République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Ecole Nationale Supérieure Polytechnique

Département d'Electronique

Laboratoire Signal et communications

Thèse de doctorat en Electronique

Option Signal et communications

Présentée par :

Mourad Abbas

Identification de thèmes pour la reconnaissance automatique de la parole

devant le jury :

Président : Mme. Mhania Guerti, Professeur à l'Ecole Nationale Polytechnique

Rapporteur : Mr. Daoud Berkani, Professeur à l'Ecole Nationale Polytechnique.

Rapporteur : Mr. Kamel Smaili, Professeur à l'université Nancy2, France.

Examinateur : Mr. Abderrezak Guessoum, Professeur à l'université de Blida.

Examinateur : Mr. Messaoud Bensebti, Professeur à l'université de Blida.

Examinateur : Mme. Latifa Hamami, Maître de Conférence à l'Ecole Nationale Polytechnique.

Année Universitaire 2007-2008

Dédicaces

Je dédie ce travail aux gens qui m'ont soutenu et aidé à préparer ce travail:
ma famille,
mes amis,
mes collègues,
mes enseignants,
et...
à toute personne qui s'intéresse à la science.

Mourad Abbas

Remerciements

J'aimerais tout spécialement remercier Monsieur Daoud Berkani, Professeur à l'Ecole Nationale Polytechnique et Monsieur Kamel Smaili Professeur à l'université de Nancy2 de m'avoir dirigé et aidé à achever ce travail aussi passionnant, et pour leur grande disponibilité durant ces années.

Je n'oublierais pas d'exprimer ma gratitude à Mr Kamel Smaili de toutes les facilités qui me les a accordées, durant mon séjour au laboratoire LORIA à Nancy.

J'exprime ma gratitude et mes vifs remerciements à Mme Mhania Guerti, Professeur à l'Ecole Nationale Polytechnique, d'avoir bien voulu présider le jury.

Je remercie vivement Mme Latifa Hamami, Maître de conférence à l'Ecole Nationale Polytechnique, Monsieur Aberrazzak Guessoum et Monsieur Messaoud Bensebti, Professeurs à l'université Saad Dahlab de Blida, d'avoir accepté de lire ce manuscrit et de m'avoir honoré d'être membres de la commission d'examen.

Je voudrais également remercier Yann Guermeur de son aide ainsi que ses conseils valeureux.

ملخص

نعالج في هذه الأطروحة فكرة غاية في الأهمية هي التعرف الموضوعي، الذي يعتبر أساسيا للحصول على كفاءات عالية في ميادين متعددة، على سبيل المثال التعرف الآلي على الكلام. في أبحاث سابقة كثيرة، تم دراسة التعرف الموضوعي وتصنيف النصوص بالنسبة للغات متعددة كالإنجليزية أو الفرنسية، إلا أن نصيب اللغة العربية منها كان ضئيلا جدا أو شبه معدوم.
بداية قمنا باستعمال طرق وخوارزميات معروفة كطريقة ت.و.ع.ت.و (تردد اللفظة / عكس تردد الوثيقة) وكذلك ما يسمى بالمكائن ذات الدعم الإتجاهي (م د إ)، و كلتاهما أدتا إلى نتائج جيدة. بعد ذلك استثمرنا طريقة أكثر تعميما من م د إ، لأول مرة في التعرف الموضوعي، هي المكائن ذات الدعم الإتجاهي متعددة الفئات "م د إ م ف"، وذلك لتجاوز "م د إ" المقصورة على الفصل بين فئتين فقط.
وقد اقترحنا طريقة جديدة أسميناها مصنف الزناد، وهي تعتمد على حساب الزنادات باستعمال المعلومة المتبادلة المتوسطة. وقد أدت هذه الطريقة إلى نتائج جيدة بالمقارنة مع "ت.و.ع.ت.و" و "م د إ م ف"، رغم استعمالنا لمجموعة مفردات ذات حجم صغير جدا.

الكلمات الجوهرية. التعرف الموضوعي، ت.ل.ع.ت.و، م د إ، التصنيف ذو الطبيعة المتعددة المواضيع باستعمال المكائن ذات الدعم الإتجاهي، اللغة العربية.

Résumé

Dans cette thèse, nous traitons un sujet d'actualité qui est l'identification de thèmes, tâche essentielle dans plusieurs domaines dont la reconnaissance automatique de la parole. Dans les travaux de l'état de l'art, l'identification de thèmes est étudiée pour différentes langues comme le Français ou l'Anglais. Toutefois, il y a eu peu d'études concernant ce sujet pour la langue Arabe. Nous avons commencé par l'application des méthodes statistiques issues de l'état de l'art, comme la TFIDF (Term Frequency/Inverse Document Frequency) et la SVM (Support Vector Machines), dans le but de les évaluer.
Nous avons utilisé une méthode plus généralisée de la SVM appelée M-SVM (Multi-Category SVM), pour la première fois dans le cadre de l'identification de thèmes afin de contourner l'insuffisance de la SVM, restreinte à la séparation de deux classes uniquement .
Une nouvelle méthode que nous avons baptisée TR-classifier, est proposée dans le but d'améliorer les résultats. L'information mutuelle moyenne est utilisée pour calculer les triggers sur lesquels est basée notre méthode. Avec des vocabulaires de tailles très petites, nous avons pu avoir des performances supérieures à celles de la TFIDF et la M-SVM.

Mots clés : *Identification de Thèmes, TFIDF, SVM, M-SVM, Langue Arabe.*

Abtsract

In this thesis, we are dealing with Topic Identification, a new interesting subject which has an important role to enhance performances of systems belonging to several domains, for example Automatic Speech Recognition. In the state of the art, Topic Identification has been studied for different languages like French and English, nevertheless, there are few works in this area for Arabic Language.
We have applied two known statistical methods, TFIDF (Term Frequency/Inverse Document Frequency) and SVM (Support Vector Machines). Both methods gave good results and seem to be efficient at least for Arabic language. Furthermore, we have applied, for the first time, the M-SVM method (Multi-category SVM) in topic identification, in order to overcome SVM which is limited only to binary categorization.
We have proposed a new method based on computing triggers, and that we have called TR-Classifier. Average mutual information is used to obtain triggers. We should notify that results which we have found using this method outperform M-SVM and TFIDF using very small corpus size.

Keywords: *Topic Identification, TFIDF, SVM, M-SVM, Arabic Language.*

Table des matières

Table des figures

Liste des tableaux

0.1 Introduction générale

A l'ère d'Internet et l'immense capacité de stockage des documents, l'information est fortement présente, néanmoins elle est éparpillée. C'est pour ce genre de problèmes que des axes de recherche sont apparus, comme la recherche d'information, le traitement automatique du langage naturel et la catégorisation de textes. Cette dernière a un lien étroit avec l'identification de thèmes : le sujet auquel nous nous intéressons à travers ce manuscrit.

L'identification de thèmes est un domaine de recherche qui a des applications dans plusieurs secteurs : reconnaissance de la parole pour adapter les modèles de langage, traduction automatique pour cibler la thématique de traduction, amélioration de la recherche dans le web, etc.

Le but ultime que nous cherchons est d'arriver à identifier le thème d'un document d'une manière efficace. Ceci aidera à améliorer le rendement des systèmes susmentionnés.

Un système de reconnaissance automatique de la parole est constitué d'un modèle acoustique responsable d'analyser l'onde acoustique qui représente une unité, qu'elle soit un mot ou une phrase, et fournir en sortie l'unité reconnue. Toutefois, bien que ces modèles acoustiques ont atteint de grands degrés de performance, ceux-ci restent toujours insuffisants à cause de la diversité et la complexité de la langue. En effet, bien que la phrase reconnue par le système pourrait coïncider à la suite de mots prononcée, elle est souvent incorrecte syntaxiquement. En considérant l'exemple de la phrase "La voiture est chargée", nous aurons une multitude de solutions correspondantes fournies par le système :

- La voiture est chargée.
- La voiture est char j'ai.
- La voiture et chargée.
- La voix tu raie char j'ai.
- La voix tu raie chargée.
- La voix tu raie chargé.
- .....

C'est pour cette raison que l'intégration d'un modèle de langage est nécessaire, puisqu'il permet de sélectionner la phrase la plus correcte possible. Les modèles de langage existants ne sont pas parfaits, il existe toujours des imperfections qui font que les systèmes de reconnaissance ne sont pas toujours très performants. Pour améliorer le rendement de ces derniers, des travaux de recherche ont été réalisés en proposant ce que l'on appelle "Adaptation des modèles de langage" qui sera appliquée au signal en cours de reconnaissance. Ces travaux ont conduit à une augmentation du gain en perplexité des modèles, ainsi qu'en taux de reconnaissance [1], [2]. Une idée qui parait prometteuse est de créer des modèles suivant le thème de la séquence en cours de reconnaissance au lieu d'avoir un modèle général. C'est à dire le système dispose de plusieurs modèles thématiques, ainsi le modèle M_i est sollicité lorsque la séquence

s'articule autour du thème T_i. Ceci aidera à améliorer le taux de reconnaissance en évitant plusieurs ambiguïtés de la langue.

Les travaux de recherche que nous avons menés ont pour objectif, tout d'abord, de tester un ensemble de méthodes de l'état de l'art, fondées sur des techniques statistiques, en utilisant un corpus en langue Arabe [3, 4, 5, 6, 7]. Certaines de ces méthodes que nous avons expérimentées, comme la TFIDF et la SVM, ont déjà fait l'objet d'étude dans le cadre de l'identification de thèmes. Cependant, nous avons exploité la M-SVM, version généralisée de la SVM, pour la première fois dans ce domaine. Nous proposons ensuite une nouvelle méthode basée sur les triggers que nous avons baptisée TR-Classifier. Nous considérons que les résultats obtenus par cette dernière sont importants vu les tailles modestes des vocabulaires de thème utilisés pour le calcul des triggers. Notre travail a nécessité la création d'un corpus en langue Arabe. Pour cela nous avons procédé à la collecte des textes arabes via le Web dans le but de construire cette matière essentielle sur laquelle notre travail va se baser : le corpus. La justification de ce choix peut être expliqué par le fait qu'il existe peu de corpus en langue Arabe, ou par le fait que leur prix est élevé (corpus de Linguistic Data Consortium).

Dans le premier chapitre, nous aborderons les méthodes connues en catégorisation de textes et en identification de thèmes. Nous présenterons également les étapes qui précédent le traitement effectué par ces méthodes, comme la représentation des documents, ainsi que la manière dont nous évaluerons ces méthodes [5].

Dans le second chapitre, nous détaillons les méthodes de l'état de l'art utilisées dans cette thèse, en l'occurrence TFIDF, SVM [5, 3]et M-SVM. Ces méthodes sont différentes selon le mode de classification. Le classifieur TFIDF se base sur le calcul de similarité entre un document et l'ensemble des thèmes, pour ensuite décider l'appartenance au thème correspondant. Les SVM, qui réalisent la discrimination binaire, reposent sur le principe de séparation par un hyperplan. Cependant, pour pouvoir traiter un nombre de thèmes supérieur à 2, la M-SVM est utilisée, où plusieurs hyperplans font la séparation.

Dans le troisième chapitre, nous exposons les expériences effectuées sur les méthodes mentionnées dans le paragraphe précédent . Avant cela, nous y présentons le corpus sur lequel nous nous sommes basés, ainsi que quelques spécificités de la langue Arabe qui nous permet de voir plus clairement les choses, particulièrement dans des situations où l'on est contraint de faire une comparaison entre les corpus des différentes langues. Nous montrons ensuite les différentes expériences appliquées aux méthodes TFIDF, SVM et M-SVM et les résultats correspondants. Notons que les tailles des documents d'apprentissage et celles des documents de test varient entre 100 et 1000 mots, voir annexe (A).Dans ce même chapitre, et dans le souci de réaliser une identification de thèmes rapide et menant à des résultats satisfaisants en même temps, nous avons testé les performances de l'une des méthodes sus-mentionnées, en l'occurrence la TFIDF, en fonction du nombre de mots qui constituent les documents [3, 4]. Certains thèmes ont atteint les performances espérées en utilisant un nombre très limité de mots, tandis que certains d'autres nécessitent un nombre plus

grand. Ce chapitre englobe aussi l'application de la M-SVM pour la première fois dans un problème d'identification de thèmes. La M-SVM permet la multi-classes classification au lieu de la séparation binaire que fait la SVM.

Nous exposons dans le quatrième chapitre, la méthode que nous avons proposé, en l'occurrence le TR-classifier. En effet cette méthode est basée sur l'exploitation du lien existant entre les mots dans le but de caractériser les thèmes faisant l'objet d'étude. Des expériences sont réalisées en variant le nombre de triggers, ainsi que la taille des vocabulaires de thèmes.

Chapitre 1

Identification de thèmes : État de l'art

1.1 Introduction

Le thème a suscité plusieurs définitions. En linguistique, et précisément en sémantique, l'élément d'un énoncé qui est réputé connu par les participants à la communication, est appelé thème. Ce dernier s'oppose au rhème qui est l'information nouvelle apportée par l'énoncé. Dans la phrase "Ton ami pratique le sport", le thème est "Ton ami", qui est supposé connu par le locuteur et son allocutaire. Tandis que le rhème est "pratique le sport". De ce fait, on peut dire que le thème est ce dont en parle, le rhème ce qu'on en dit. Par contre, en morphologie, le thème est l'ensemble constitué par un radical et ses affixes de formation sans les désinences. La définition de thème diffère donc d'un domaine à un autre. Dans les études relatives au traitement automatique des thèmes, nous trouvons différentes notions. Dans [2] les mots-clés sont considérés comme étant des thèmes, tandis que dans d'autres travaux comme ceux de [8] un thème est plus général, ainsi "politique" et "économie" sont des thèmes. Dans la plupart des travaux de l'état de l'art, ces derniers sont considérés comme étant des mots-clés. Parmi les applications qui ont considéré des thèmes généraux on peut citer celles de [9], [10].

L'identification de thèmes est l'opération qui consiste à attribuer une étiquette à un flux de données textuelles. Ceci étant la définition la plus simple mais aussi la plus précise, du moins en ce qui concerne notre étude. La figure (1.1) présente un schéma simplifié décrivant la procédure d'identification dans le cas où le nombre de thèmes est trois.

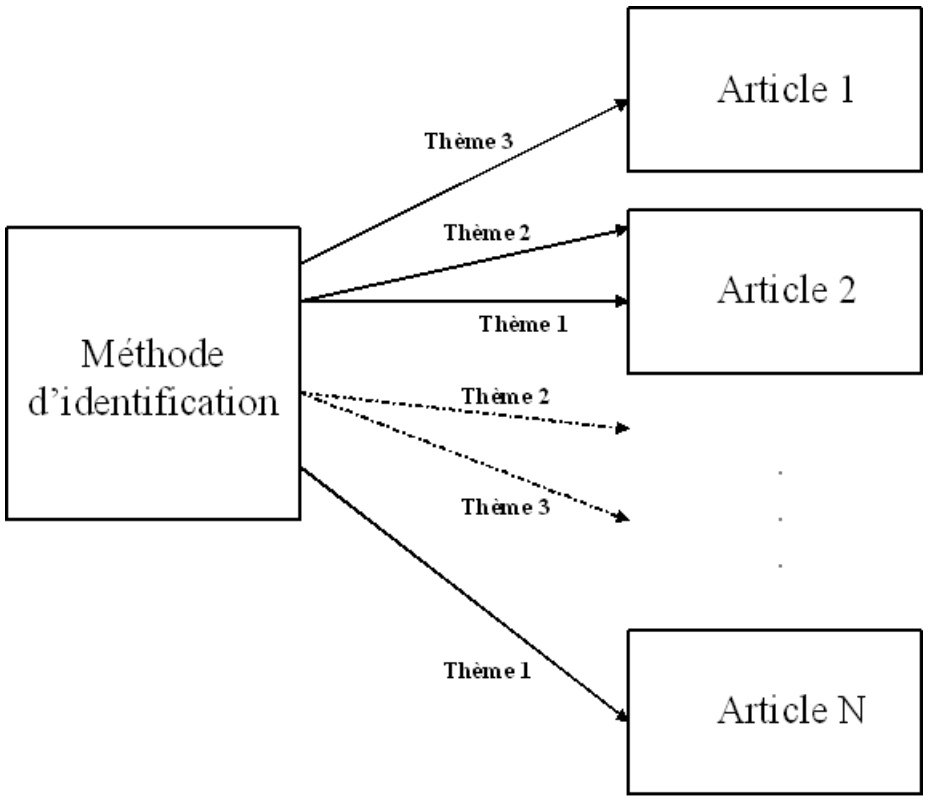


FIG. 1.1 – Schéma définissant l'identification de thèmes

1.2 État de l'art

1.2.1 Introduction

L'identification de thème a un lien très étroit avec la catégorisation de textes. L'objectif est d'assigner des catégories prédéfinies à des documents textuels. En remplaçant les catégories par des thèmes, il s'agit alors d'identification de ces derniers, qui est le but de ce manuscrit.

Plusieurs travaux ont été menés au cours de la dernière décennie sur la catégorisation de textes : les classificateurs bayesiens [11, 12, 13], arbres de décision [14, 12, 13], réseaux de neurones [15, 16], KNN [17, 18]. D'autres approches adoptées utilisaient des systèmes experts comme le système "Construe" développé par Carnegie Group.

Plusieurs recherches ont été mené pour la classification des documents écrits en langue anglaise. D'autres langues ont eu le privilège d'être l'objet d'études, comme l'allemand, l'Espagnol, l'Italien [19], ainsi que les langues asiatiques comme le Japonais et le Chinois [20].

En ce qui concerne la langue Arabe, il y a peu de travaux dans ce domaine. En effet, dans [21], l'algorithme Naïve Bayes a conduit à des performances égales à 71.96 % en terme de Rappel. Un autre système a été proposé dans [22], où ont été utilisées des méthodes statistiques de classification comme le Maximum d'entropie.

Les résultats obtenus ont conduit à un taux de Rappel égal à 84.2 % avec une valeur très basse de Précision, en l'occurrence 50 %.

D'autres méthodes basées sur des "règles d'association" pour la classification des textes arabes ont été reportées dans [23] et ont permis d'atteindre un taux de Rappel égal à 74.5 %. En outre, la société Sakhr a mis en oeuvre un système de classification permettant d'aboutir à un taux de Rappel égal à 73.78 % avec une précision faible 47.35 %. Néanmoins, aucun détail n'a été fourni sur la méthode adoptée par ce système (siraj.sakhr.com).

Un autre système appelé "ArabCat", basé sur le Maximum d'entropie a été reporté dans [24]. Il a mené à des performances égales à 80.48 % en terme de Rappel avec une Précision de 80.34 %.

Concernant la langue française, et dans le but d'obtenir une identification thématique plus fiable, plusieurs efforts ont été déployés. En effet, Brun a proposé une méthode fondée sur la similarité mot-thème, présentée en détail dans la section 1.2.2 [10], qui a conduit à un taux de rappel égal à à 82 :5 %.
D'autres travaux ont été proposés par Bigi dans [25] exploitant les relations sémantiques établies par la hiérarchie thématique. Cette voie paraît théoriquement judicieuse, vu que les thèmes sont souvent de granularité différente. Par conséquent, il a fallu exploiter cette structure hiérarchique afin de déterminer automatiquement le thème d'un texte. Le corpus utilisé dans cette méthode est extrait des forums de discussions "newsgroups" afin de tirer profit de leurs structures arborescentes, en absence de corpus traditionnellement utilisés en RAP qui ne sont pas hiérarchisés.

Nous allons présenter dans ce chapitre les différentes méthodes les plus connues de l'état de l'art concernant deux domaines très liés, en l'occurrence : l'identification de thèmes et la catégorisation de textes.

1.2.2 Quelques méthodes

Certaines des méthodes que nous allons présenter dans cette section sont des méthodes d'apprentissage supervisé. L'apprentissage automatique (Machine Learning) a comme but la création d'une fonction à partir d'un ensemble de données d'apprentissage, constitué de couples de valeurs d'entrée et de sorties désirées. La sortie de la fonction permet d'avoir la classe de l'objet d'entrée, ce qui est communément appelé classification. Les méthodes d'apprentissage sont largement utilisées dans différents domaines, on peut citer par exemple la reconnaissance de formes, la catégorisation de textes, etc. En reconnaissance de formes on peut citer l'exemple de la reconnaissance de chiffres manuscrits et la reconnaissance de visages. Cette dernière, par exemple, a comme paramètre d'entrée une image bidimensionnelle et en sortie la personne reconnue. En catégorisation de textes plusieurs applications font actuellement l'objet d'études, en l'occurrence la classification d'emails et la classification de pages web. Les méthodes d'apprentissage ne sont pas uniquement limitées à ces applications mais il y a une multitude de domaines aussi importants. Le diagnostic médical en est un, car il englobe des applications de recherche utiles comme l'évaluation des risques de cancer et la détection d'arythmie cardiaque. Dans

ce qui suit nous présenterons les méthodes d'apprentissage utilisées en catégorisation de textes et en identification de thèmes.

K plus proches voisins .

Les K plus proches voisins ou KNN (K Nearest Neighbor) a été appliquée au début des années 90 pour la catégorisation de textes [26, 27]. Pour assigner un thème T_i à un document d, cette méthode consiste à chercher les k documents les plus proches parmi ceux constituant les données d'apprentissage, d'où vient cette appellation de "plus proches voisins". Ces derniers sont obtenus en utilisant en général une distance cosine. L'appartenance du document d au thème T_i est obtenue si le score relatif à T_i est le plus élevé.

Classifieur de Naïve Bayes .

La prédiction de la catégorie à laquelle un document appartient, peut être obtenue en utilisant un classificateur basé sur le théorème de Bayes et appelé le Classificateur Naïve Bayes. Le théorème de Bayes est donné par la relation (1.1) :

$$P(Y/A) = \frac{P(A/Y)P(Y)}{P(A)} \tag{1.1}$$

$P(Y)$ est la probabilité a priori de l'hypothèse Y.

$P(A)$ est la probabilité à priori des données d'apprentissage.

$P(Y/A)$ est la probabilité de Y sachant A.

$P(A/Y)$ est la probabilité de A sachant Y.

L'hypothèse la plus probable sachant les données d'apprentissage est appelée Hypothèse maximum à posteriori et est donnée par la relation (1.2) :

$$H = argmax\, P(Y/A) \tag{1.2}$$

$$= argmax\, \frac{P(A/Y)P(Y)}{P(A)}$$

$$= argmax\, P(A/Y)P(Y)$$

Dans le cas où les hypothèses Y ont la même probabilité, alors H se calcule en utilisant le critère de la vraisemblance maximale "Maximum Likelihood", comme il est présenté dans la formule (1.3).

$$H = argmax_{Y} P(A/Y) \tag{1.3}$$

En décrivant la variable A par les attributs $a_1, a_2, \ldots, a_n$, la formule (1.2) devient comme suit (1.4).

$$H = arg\,max_{Y}\, P(Y/a_1, a_2, \ldots, a_n)$$

$$= argmax_{Y} \frac{P(a_1, a_2, \ldots, a_n/Y)P(Y)}{P(a_1, a_2, \ldots, a_n)}$$

$$= argmax_{Y}\, P(a_1, a_2, \ldots, a_n/Y)P(Y) \tag{1.4}$$

L'hypothèse Naïve Bayes est donnée par l'équation (1.5).

$$P(a_1, a_2, \ldots, a_n/Y) = \prod_i P(a_i/Y) \tag{1.5}$$

Le classifieur Naïve Bayes est donné alors par l'expression (1.6).

$$H = arg\,max_{Y}\, P(Y) \prod_i P(a_i/Y) \tag{1.6}$$

Dans le cas d'utilisation de cette méthode afin d'identifier le thème T d'un document d, il suffit alors de changer les variables A et Y , respectivement par d et T.

Réseaux de neurones .

Des travaux utilisant des réseaux de neurones en catégorisation de textes sont présentés dans [15, 16] et [28]. Wiener et al. ont utilisé des perceptrons (sans couche cachée) ainsi que des réseaux de neurones à trois couches (avec une couche cachée). Ng et al. utilisaient seulement des perceptrons. Dans chacun des deux systèmes on trouve un réseau pour chaque catégorie. Yang et al. [28] ont constaté, dans leurs expériences sur le corpus Reuters 21578, que le temps d'apprentissage des réseaux de neurones dépasse largement les autres classifieurs. Par conséquent, l'apprentissage d'un réseau pour chaque catégorie serait bien évidemment plus coûteux. Ils décidèrent alors d'exécuter un apprentissage d'un réseau pour les 90 catégories de Reuters. Toutefois, les performances de cette méthode sont dépassées largement par celles de SVM et KNN, selon [28].

Arbres de décision .

La méthode connue sous l'appellation Arbres de décision est une technique d'apprentissage supervisé très populaire. En effet les règles de décision issues de l'apprentissage sont facilement interprétables. Un arbre de décision est constitué de nœuds, de branches et de feuilles. Un nœud représente un terme, tandis que la branche représente un test sur le terme dans un document. Ceci consiste, d'une manière simple, à vérifier sa présence ou sa fréquence dans ce document. Le résultat de la classification attribué à ce dernier est celui qui correspond à la feuille obtenue par parcours de l'arbre. La classification d'un document consiste donc à le soumettre à la racine de l'arbre, et lui appliquer les tests en parcourant les branches jusqu'à atteindre la feuille qui représente le résultat de classification. Cette méthode utilise un vocabulaire général V. L'apprentissage se base sur un ensemble de documents. Pour chaque nœud appartenant au vocabulaire V, un terme est choisi. En réalisant le test sur ce dernier, l'ensemble des documents se divise en deux sous- ensembles, l'un satisfait la condition mise sur le terme, l'autre ne la vérifie pas. Ce processus est répété sur ces deux ensembles jusqu'à ce que l'on ne puisse plus trouver de terme permettant la subdivision de l'ensemble, et ainsi le résultat de la classification est obtenu par la création d'une feuille portant l'étiquette majoritairement représentée par les exemples reçus. Les tests réalisés dans les nœuds utilisent généralement le gain d'information, critère basé sur l'entropie.

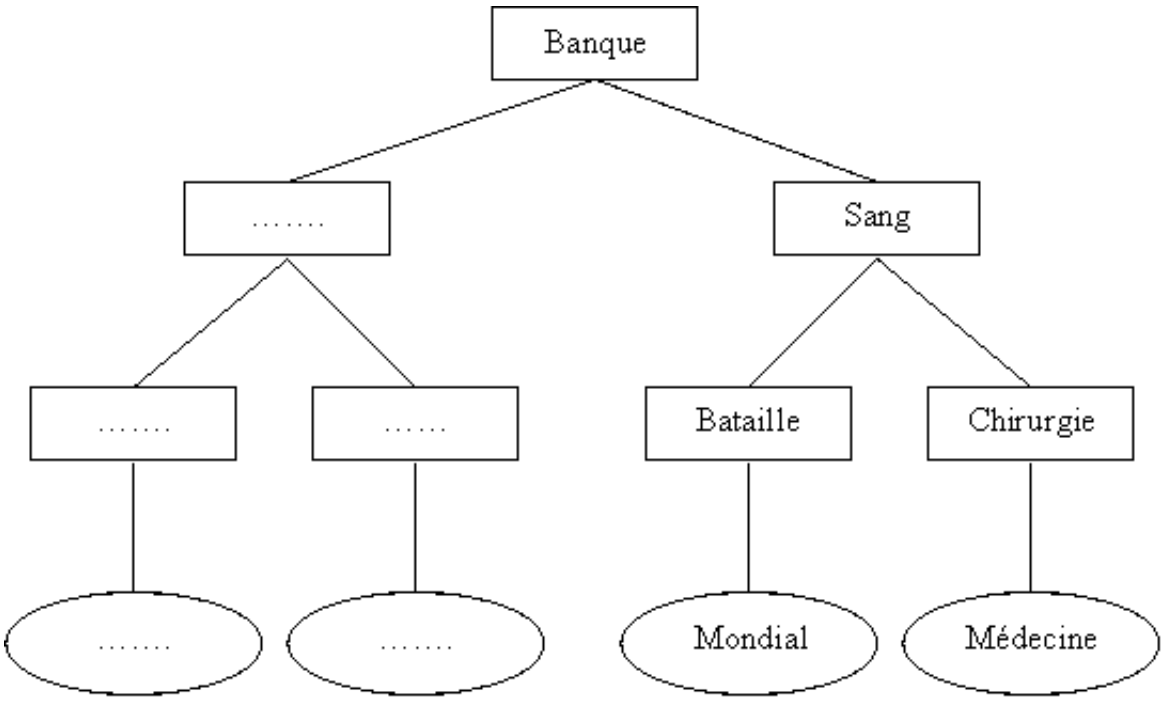


FIG. 1.2 – Arbres de décision

Le classifieur TFIDF .

L'algorithme Rocchio est l'un des algorithmes les plus populaires largement répandus dans la recherche d'informations [29]. Il a été utilisé ensuite dans la catégorisation de textes. Dans [29], Joachims a exposé l'une des plus importantes composantes heuristiques de cet algorithme qui est connue sous l'acronyme TFIDF (Term Frequency / Document Inverse Frequency). En effet c'est une méthode de pondération des mots, grâce à laquelle, cet algorithme est désormais connu sous le nom du classifieur TFIDF.

Soit les documents d_1= {La maison} et d_2={La la la maison maison}. La figure (1.3) montre une représentation vectorielle de deux documents d_1 et d_2 (vecteurs v_1 et v_2).

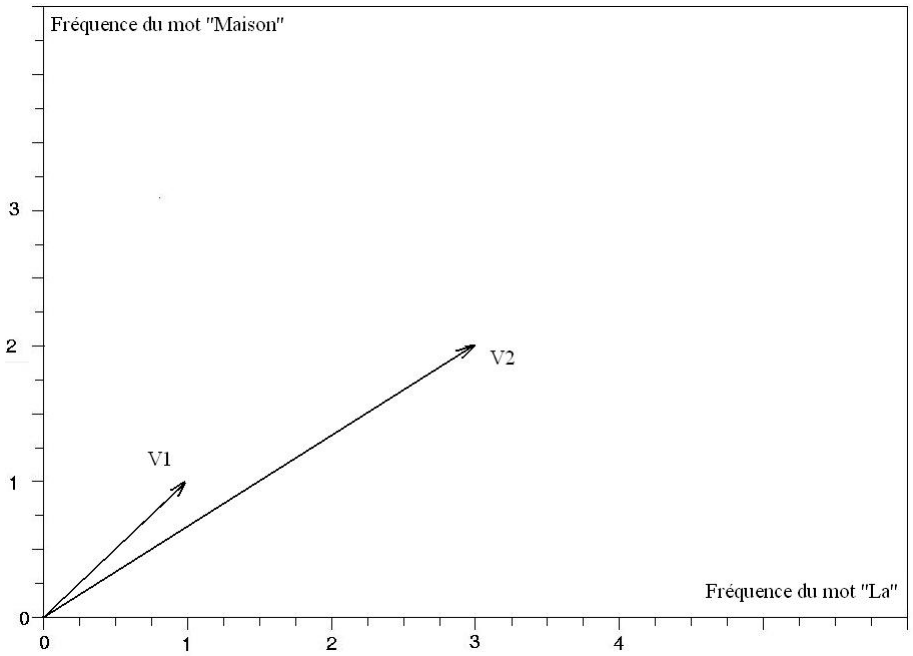


FIG. 1.3 – Un exemple de représentation vectorielle des deux documents $d_1$et d_2

La figure (1.3) montre l'idée de base de cet algorithme, en l'occurrence : la représention des documents sous forme de vecteurs. Ainsi un document doc_i constitué de n mots : $doc_i = \{w_1, w_2, \ldots, w_n\}$, est transformé en un vecteur $D_i = \{d_{i1}, d_{i2}, \ldots, d_{i|V|}\}$. où $\mid V \mid$est la taille du vocabulaire. Chaque composante d_{ik} du vecteur représente une pondération du mot w_k. Elle est obtenue en effectuant le produit des deux grandeurs (valeurs statistiques) $TF(w,d)$ et $IDF(w)$. La fréquence de mots ou Term Frequency $TF(w,d)$ exprime le nombre de fois où le terme w apparaît dans le document d. Tandis que la fréquence de documents ou Document Frequency $DF(w)$ est le nombre de documents dans lesquels apparaît le terme w une fois au minimum. La valeur d_{ik} est obtenue en utilisant l'équation (1.7).

$$d_{ik} = TF(w_{ik}, d).IDF(w_{ik}) \tag{1.7}$$

L'inverse de la fréquence de documents (Inverse Documents Frequency) $IDF(w)$ est donné par la relation (1.8).

$$IDF(w) = log\left(\frac{\mid D \mid}{DF(w)}\right) \tag{1.8}$$

$\mid D \mid$ est le nombre total des documents. Le classifieur TFIDF représente aussi chacun des thèmes ou "classes" par un vecteur en se basant sur le corpus d'apprentissage concernant ce thème. Ainsi le thème T_j est représenté par le vecteur $D_j = \{d_{j1}, d_{j2}, \ldots, d_{j|V|}\}$. La similarité $sim(D_j, D_i)$ entre le thème T_j (représenté par le vecteur D_j) et le document doc_i (représenté par le vecteur D_i) est calculée en utilisant l'équation (1.9).

$$sim(D_j, D_i) = \frac{\sum_{k=1}^{|V|} d_{jk}.d_{ik}}{\sqrt{\sum_{k=1}^{|V|}(d_{jk})^2.\sum_{k=1}^{|V|}(d_{ik})^2}} \tag{1.9}$$

Méthode SVM .

La méthode SVM (Support Vector Machines) est une méthode d'apprentissage qui a été introduite au début des années 90, dans le but de réaliser une catégorisation bi-classes [30, 31]. Elle est définie à partir des modèles linéaires dont l'apprentissage se fait en minimisant l'erreur empirique et dans le même temps, en essayant de maximiser une marge géométrique. Cette dernière est la plus petite distance entre un point de l'ensemble d'apprentissage et le plan séparateur réalisé par le modèle. La figure (1.4) présente un exemple de séparation de deux classes, ainsi que la notion de marge. Nous y remarquons également les "support vectors" situés sur les deux plans parallèles au plan séparateur. Plus de détails sur la SVM sont présentés dans le chapitre 2.

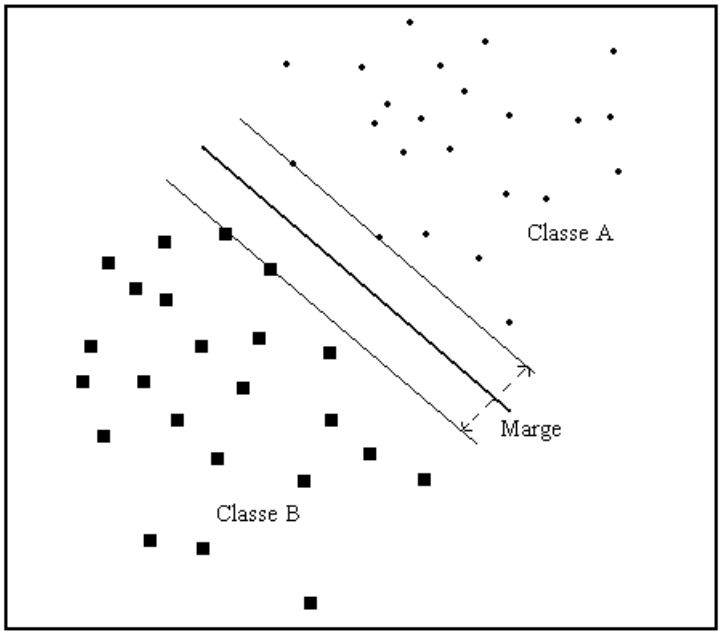


FIG. 1.4 – Séparation de deux classes par un hyperplan optimal

Le modèle Cache .

Le modèle Cache exploite les mots présents dans l'historique d'un texte (oral ou écrit) afin d'en identifier le thème [32].

Son principe est de renforcer la probabilité d'un mot si celui-ci a déjà été rencontré dans son contexte gauche. Ceci est réalisé en utilisant une fenêtre glissante (Cache) de taille fixe dans laquelle la fréquence relative des mots est examinée [33]. Cette taille peut être de plusieurs dizaines, centaines ou milliers de mots [34, 32].

La formule de base du modèle Cache est présentée par l'équation (1.10) :

$$P\left(w/h\right) = \frac{N_h\left(w\right)}{|h|} \tag{1.10}$$

$N_h\left(w\right)$est le nombre d'occurrence du mot w dans l'historique h.
$|h|$ est la taille de l'historique.

L'exploitation du modèle Cache en identification du thème consiste à comparer la distribution unigramme des mots du vocabulaire du thème et celle du Cache [1] d'un document d à un instant t.

Dans les travaux présentés dans [10] le modèle Cache a été la méthode la plus performante après le modèle unigramme thématique.

Le modèle unigramme thématique .

Le modèle unigramme thématique a pour objectif l'évaluation de la probabilité d'un thème sachant le document d en cours de traitement. Ceci peut être obtenu en utilisant la formule (1.11).

$$P\left(T_i \mid d\right) = \frac{P(T_i)P(d \mid T_i)}{\sum_{k=1}^{I} P(T_k)P(d \mid T_k)} \tag{1.11}$$

Sachant que $P\left(T_i\right)$est la probabilité à priori du thème T_i. $P(d \mid T_i)$ représente la probabilité du document d sachant le thème T_i et elle est évaluée de la manière suivante :

$$P(d \mid T_i) = \prod_{j=1}^{|V|} (P(w_j \mid T_i))^{d_j} \tag{1.12}$$

sachant que $\mid V \mid$est la taille du vocabulaire général, et d_j représente le nombre d'occurrences du mot w_j.

Méthode fondée sur la perplexité .

La perplexité est utilisée pour l'évaluation des modèles de langage. Elle est définie comme étant une mesure dérivée de l'entropie croisée [33]. Cette dernière mesure la qualité de prédiction d'un événement e par le modèle M sur un corpus textuel C . Elle est définie par la formule (1.13).

$$H(P_C, P_M) = \sum_e P_C(e).log_2(P_M(e)) \tag{1.13}$$

L'entropie croisée désigne l'entropie du corpus C perçue par un modèle M [35]. Lorsque le corpus de test est suffisamment grand, elle devient comme présentée dans l'expression (1.14). N désigne le nombre de mots du corpus C.

$$H_M = -\frac{1}{N} \sum_e log_2(P_M(e)) \tag{1.14}$$

La valeur H_M est vue comme étant le nombre moyen de bits nécessaires au codage des mots du corpus en utilisant le modèle$M$ [33]. Par conséquent, ce dernier est considéré performant si le nombre de bits alloués est minimal. D'où la formule (1.15) [36] exprimant la perplexité d'un modèle M, définie comme l'inverse de la moyenne des probabilités affectées aux mots du corpus C.

$$PP_M(C) = 2^{H_M} \tag{1.15}$$

L'utilisation de la perplexité pour identifier les thèmes, consiste à créer un modèle de langage pour chacun des thèmes traités. Ensuite, la perplexité de chacun des

modèles est alors évaluée sur le document d. L'attribution de l'étiquette du thème $T_i$au document d est faite si la perplexité du modèle correspondant est minimale. Nous définissons la perplexité sur le thème T_i par la formule (1.16). N désigne la taille du document d, et w_k sont les mots y appartenant.

$$PP_i(d) = \left(P(w_1) \prod_{k=2}^{N} P(w_k \mid w_{k-1}) \right)^{-\frac{1}{N}} \tag{1.16}$$

Méthode WSIM .

La méthode WSIM "Word Similarity" a été proposée par Brun [10]. L'idée est que les valeurs représentant un thème T_i exploitent la similarité qui existe entre le mot et le thème. Deux mots m_1 et m_2 sont considérés similaires si leurs informations mutuelles avec l'ensemble des autres mots du vocabulaire sont proches [37, 10]. En effet, l'objectif est de mettre les deux mots en contexte droit et gauche pour pouvoir les évaluer, ceci est réalisable en exploitant les valeurs de leurs informations mutuelles avec les mots v_i du vocabulaire V , en l'occurrence $I(m_1, v_i)$, $I(m_2, v_i)$, $I(v_i, m_1)$, $I(v_i, m_2)$, voir figure (1.5). Le calcul de similarité entre deux mots est exprimée par la relation (1.17) :

$$Sim(m_1, m_2) = \frac{1}{2|V|} \sum_{i=1}^{|V|} \left(\frac{min(I(v_i, m_1), I(v_i, m_2))}{max(I(v_i, m_1), I(v_i, m_2))} + \frac{min(I(m_1, v_i), I(m_2, v_i))}{max(I(m_1, v_i), I(m_2, v_i))} \right) \tag{1.17}$$

$|V|$est la taille du vocabulaire de thème T_i.

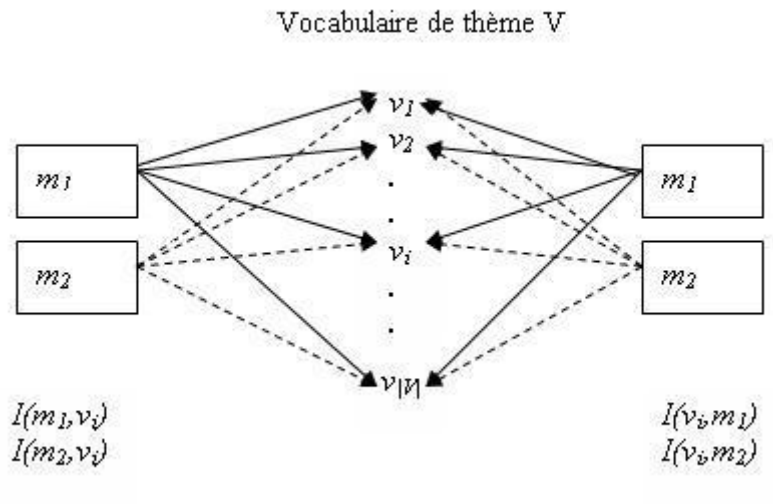


FIG. 1.5 – Similarité entre deux mots

L'équation (1.17) permet de calculer la similarité entre deux mots concernant un

thème donné. Ce qui est important dans cette méthode c'est d'obtenir la similarité entre un mot et un thème.

Un thème T_i est représenté par un vecteur R_i dont chaque composante exprime la valeur de similarité entre ce thème et chacun des mots du vocabulaire.

$$R_i = r_{im1}, r_{im2}, ..., r_{im|V|}$$

La similarité entre un mot m et un thème T_i est donnée par l'équation (1.18) :

$$r_{im} = Sim(m, T_i) = P(m \mid T_i) \frac{\sum_{k=1}^{|V|} Sim(m, t_k)}{\sum_{m=1}^{|V|} \sum_{k=1}^{|V|} Sim(m, t_k)} \tag{1.18}$$

où $P(m \mid T_i)$désigne la probabilité a priori du mot m dans le thème T_i.

Pour déterminer le thème d'un document $d = \{w_1, w_2, \ldots, w_N\}$, il faut tout d'abord calculer un score pour les thèmes traités. Ceci est défini par l'équation (1.19).

$$Score(T_i \mid d) = \frac{\sum_{j=1}^{N} r_{iw_j}}{\sum_{k=1}^{I} \sum_{j=1}^{N} r_{kw_j}} \sum_{j=1}^{N} \psi_{ji} \tag{1.19}$$

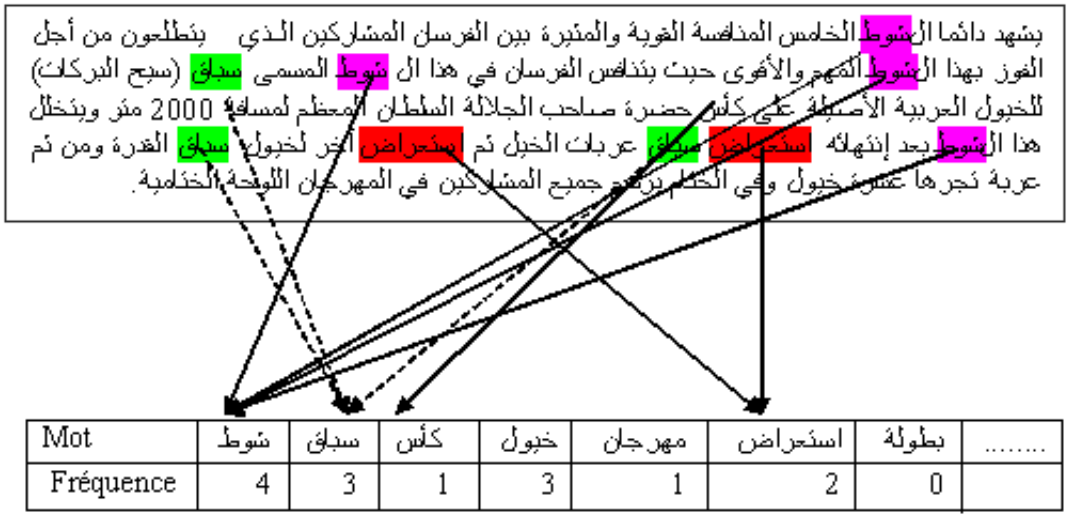
$$\text{avec } \psi_{ji} = \begin{cases} 1 & si\, w_j \in V \\ 0 & sinon \end{cases}$$

La probabilité du thème T_i sachant le document d, est calculée par la formule (1.20). Le thème assigné est celui qui maximise la probabilité présentée par cette dernière. Notons que α_i sont des coefficients de pondération estimés par validation croisée.

$$P(T_i \mid d) = \frac{\alpha_i Score(T_i \mid d)}{\sum_{k=1}^{I} \alpha_k Score(T_i \mid d)} \tag{1.20}$$

1.3 Représentation de documents

La représentation de documents est une étape primordiale dans un système d'identification de thèmes. Un document ne peut être traité par un algorithme d'identification avant qu'il ne subisse une transformation particulière permettant sa représentation sous un format spécifique. Les mots sont les paramètres représentatifs les plus adéquats pour des tâches de classification, toutefois une partie de l'information sur le document est perdue. Un sac de mots ou "Bag of words" est la méthode la plus connue dans les travaux relatant de la classification de textes et de l'identification de thèmes. A chaque mot est attribué une valeur correspondant à sa fréquence d'apparition dans le document. L'exemple de la figure (1.6) montre la façon dont un document est représenté selon cette méthode.



Mot	شوط	سباق	كأس	خيول	مهرجان	استعراض	بطولة	.....
Fréquence	4	3	1	3	1	2	0	

FIG. 1.6 – Représentation des documents par la méthode Bag of words

"Bag Of Words" n'est pas l'unique représentation utilisée. D'autres chercheurs ont tenté utilisé les n-grammes, ce qui a causé une dégradation des performances [38, 39, 40]. Ceci est dû à plusieurs raisons : la fréquence basse des unigrammes, la haute dimensionnalité (Pour n mots nous pouvons obtenir $n!$ phrases), et aussi la redondance causée par la forte présence de la synonymie des mots [41]. Toutefois des efforts ont été fournis dans le sens d'obtenir des résultats fructueux par l'addition des n-grammes à la méthode Bag Of Words. En effet, dans [42, 43], il a été montré que l'amélioration des performances est obtenue en utilisant cette méthode de combinaison de ces deux modes de représentation, néanmoins le nombre de mots constituant les phrases "les n-grammes" ne doit pas excéder la valeur 3. Une autre méthode se basant sur les bigrammes a été expérimentée dans [41]. Elle consiste à ajouter des bigrammes aux unigrammes, avec un taux de 2 % pour éviter la haute dimmensionnalité sus-mentionnée. Les bigrammes ont été soigneusement sélectionnés en utilisant leur fréquences ainsi que le gain d'information. Toutefois, ces méthodes ont amélioré les performances pour certaines catégories et en ont dégradé pour certaines d'autres.

1.4 Mesures d'évaluation

Pour évaluer les performances d'un classificateur, au moins trois mesures standard sont utilisées, en l'occurrence le Rappel R, la Précision P et la mesure F1. La performance du classificateur est le degré d'efficacité d'attribution des étiquettes aux documents. Le Rappel est obtenu en faisant un rapport du nombre des documents correctement étiquetés par le classificateur automatique et du nombre total des documents ayant cette même étiquette. Le calcul de la Précision se fait en divisant le nombre des documents correctement étiquetés par le nombre total des documents étiquetés par le classificateur. La combinaison de ces deux mesures donne une nou-

velle mesure appelée F_1, et est définie par la formule 1.21 :

$$F_1 = \frac{2RP}{R+P} \tag{1.21}$$

Cette mesure rassemble la fonction du Rappel et celle de Précision, c'est-à-dire elle permet de fournir le nombre de documents correctement étiquetés avec fiabilité. Il y a deux manières avec laquelle, le calcul de cette mesure s'effectue : la première s'appelle macro-averaging, elle consiste à considérer chaque catégorie individuellement, ensuite calculer la moyenne pour toutes ces catégories. Tandis que la deuxième micro-averaging fournit cette mesure, d'une façon globale, en considérant les n résultats ou décisions binaires. n est le produit du nombre de documents de test et le nombre de catégories [28].

Il existe d'autres mesures qui sont appliquées dans plusieurs domaines, le taux de fausses acceptations "False Acceptation Rate (FAR)" qui est défini par le fait d'accepter à tord un faux document, et le taux de faux rejets "False Rejection Rate (FRR)" qui consiste à refuser à tord un vrai document. Ainsi, pour un thème donné T_i , les deux taux peuvent être donnés par les relations suivantes :

$$FAR = \frac{Nombre\,de\,documents\,faussement\;etiquetes\,T_i}{Nombre\,de\,documents\,n'appartenant\,pas\,au\;theme\,T_i}$$

$$FRR = \frac{Nombre\,de\,documents\,faussement\,rejetes\,T_i}{Nombre\,de\,documents\,appartenant\,au\,theme\,T_i}$$

Les valeurs de ces deux grandeurs varient entre 0 et 1. Par exemple si la valeur de FAR est égale à 0, cela veut dire qu'aucun des documents n'est faussement étiqueté. La valeur 1 indique que tous les documents sont faussement identifiés. On ne peut pas dire qu'un système est performant si la valeur de l'un des taux est basse alors que pour l'autre est élevée. Un bon système doit avoir conjointement des valeurs minimales pour ces deux grandeurs.

1.5 Conclusion

Nous avons présenté dans ce chapitre les différentes méthodes issues de l'état de l'art. Ces dernières ont pour objectif d'atteindre des taux d'identification de thèmes satisfaisants. La philosophie de ces méthodes repose sur la probabilité et les statistiques, deux principes essentiels pour pouvoir contourner le problème de la compréhension de la langue et son côté sémantique. Pour cela, la plupart de ces méthodes utilise le mode de représentation Bag Of Words, que nous avons préféré l'illustrer par l'exemple de la figure (1.6), exposé en haut, montrant un texte contenant des mots et leurs fréquences correspondantes.

La fiabilité des méthodes sus-mentionnées diffère d'une méthode à une autre. La SVM, par exemple, est connue par sa supériorité sur les autres méthodes, en ce qui concerne la discrimination binaire.

Les expériences en catégorisation réalisées dans [28] ont montré que la SVM dépasse les performances de la KNN. Les réseaux de neurones et le classifieur de Naïve Bayes sont beaucoup moins performants. Quelques résultats de ces expériences sont exposés dans la table (1.1).

Méthode	SVM	KNN	Réseaux de neurones	Naïve Bayes
$F_1(\%)$	85.99	85.67	82.87	79.56

TAB. 1.1 – Performances des méthodes de catégorisation en terme de la mesure F_1 (Expériences réalisées par Yang [28])

Dans ses expériences présentées dans [29], Joachims a montré que le classifieur de Naïve Bayes dépasse la méthode TFIDF. En utilisant des méthodes différentes de sélection de vocabulaire, ainsi que des vocabulaires de tailles différentes, Brun a élaboré une comparaison des performances des méthodes étudiées [10]. Voir table (1.2).

Méthode	Unigramme	Modèle Cache	WSIM	Perplexité	SVM	TFIDF
Rappel (%)	83.4	82.5	82.5	79	78.4	74.4

TAB. 1.2 – Performances des méthodes de catégorisation en terme de Rappel (Expériences réalisées par Brun [10])

Selon la table (1.2), la SVM n'est pas considérée comme la plus performante méthode comme c'est présenté par Joachims "voir table (1.1)".

Dans nos expériences que nous présenterons dans le chapitre 3, nous avons testé les deux méthodes TFIDF et SVM, Nous avons trouvé que la SVM dépasse la TFIDF, toutefois les performances de cette dernière ne se sont pas dégradées aux valeurs obtenues dans [10].

Chapitre 2

Méthodes utilisées : Notions et Définitions

2.1 Introduction

Les travaux de recherche en identification de thèmes appliqués à la langue Arabe, sont peu ou inexistants. Par conséquent nous étions motivés de travailler dans cette voie. Dans l'état de l'art il a été démontré que les méthodes statistiques, comme la TFIDF et la SVM , ont mené à des résultats satisfaisants. Nous souhaiterions savoir la nature des résultats obtenus, en appliquant ces méthodes pour la langue Arabe. En outre, nous avons utilisé une méthode généralisée de la SVM : la M-SVM (Multicategory Support Vector Machines) pour la première fois en identification de thèmes. Dans [44], il a été montré que l'utilisation de la M-SVM dans la fusion des méthodes de prédiction de la structure secondaire des protéines améliore les résultats. Ceci est dû essentiellement à deux raisons. L'une d'elles est que ces méthodes reposent sur différents principes. L'autre est que les données sont extraites de différentes sources de connaissances. Pour notre cas d'utilisation de cette méthode, il s'agit de faire une identification directe des différents thèmes et non deux à deux comme le fait la SVM.

Dans les sections suivantes nous présenterons en détail des définitions sur la TFIDF où nous expliquerons les premières étapes primordiales de l'identification, puis nous exposerons des notions sur le classificateur binaire : la SVM ainsi que sur la M-SVM.

2.2 Méthode TFIDF

L'origine de la méthode TFIDF vient de l'algorithme proposé par Rocchio dans [45]. Elle se base, comme son nom l'indique, sur deux grandeurs : la fréquence de mots et la fréquence de documents. Son originalité est basée sur deux idées essentielles :

une représentation fiable de documents et une sélection optimale de paramètres "Feature Selection".

2.2.1 Représentation des documents

Les premiers travaux concernant la représentation de documents qui ont été élaborés par Luhn consistent en l'utilisation des mots. En effet, ce choix d'utilisation est motivé par la signification des mots, qui se traduit, selon Luhn [46] , par la fréquence de ces derniers. Ses propositions sont basées sur les points suivants :

- La répétition de certains mots est sollicitée pour insister sur un thème T.
- L'auteur utilise un sens unique d'un mot dans le texte.
- Il y a un nombre limité de mots permettant d'exprimer un concept particulier. Bien que plusieurs auteurs peuvent choisir d'autres mots pour le même concept, pour des raisons de style.

Dans le domaine de Recherche d'information, il semble que l'utilisation des mots dans la représentation a conduit à des résultats satisfaisants [40]. Il a été constaté aussi que l'ordre des mots dans un document n'est pas important [29]. C'est la raison pour laquelle, la méthode Bag of Words "Sac de Mots", exposée dans la section 1.3, est largement sollicitée dans la représentation d'un document, où chaque mot distinct est un paramètre ayant comme valeur le nombre d'apparition du mot dans le document, appelée communément "Term Frequency" ou Fréquence de Mots.

2.2.2 Sélection de paramètres "Feature Selection"

Effectuer une sélection de paramètres, c'est choisir un bon sous-ensemble de mots, qui puisse mener à une meilleure identification de thèmes ou une catégorisation de textes, selon la tâche désirée [47]. On parle ici de bon sous-ensemble car l'utilisation de la totalité des mots constituant un document, peut nuire à cette tâche par le fait d'ajouter du bruit. L'autre cas, c'est à dire l'utilisation de très peu de mots, ne peut cependant pas aider à formuler de bonnes hypothèses [29]. Par conséquent, l'élimination des mots non fréquents ainsi que ceux ayant une fréquence très élevée est nécessaire. Les mots non fréquents sont ceux qui apparaissent au plus i fois dans le corpus d'apprentissage. La valeur de i est généralement choisie égale à 3 [45]. Ceci aidera à réduire la majorité des erreurs d'orthographe, et permettra d'accélérer les étapes qui suivent. Les mots dont la fréquence est très élevée sont connus, ce sont les mots outils de la langue qui n'apportent aucune information utile dans le domaine d'identification de thème, ce qui rend leur suppression essentielle.

L'ensemble des mots qui serviront à la représentation des documents, peuvent être sélectionnés en utilisant plusieurs méthodes. Nous citons la fréquence de mots, la fréquence de documents, l'information mutuelle et le gain d'information [10]. Le classifieur TFIDF représente chacun des mots retenus en faisant la combinaison des deux grandeurs $TF(w, d)$ et $IDF(w)$ précisées dans la section 1.2.2. L'appartenance du document à un thème est obtenue en utilisant l'équation (1.9) qui permet de

calculer la distance cosine entre les deux vecteurs. Dans les paragraphes suivants, nous exposons quelques méthodes de sélection de mots-clés.

Fréquence de mots .

La fréquence de mots est l'une des méthodes de sélection des mots-clés la plus basique. Elle consiste à calculer la fréquence d'apparition de chaque mot dans le corpus d'apprentissage. Les mots qui apparaissent fréquemment sont considérés les plus représentatifs, et par conséquent ils contribuent efficacement dans la détection de thème.

Fréquence de documents .

La fréquence de documents d'un mot est une méthode répandue. Elle est définie comme étant le nombre de documents dans lesquels ce mot apparaît. Les mots conservés sont ceux présents dans le plus grand nombre de documents. En comparant cette méthode à l'utilisation de l'ensemble des mots constituant le corpus d'apprentissage, [48] a montré que les performances sont les mêmes sachant que le nombre de mots est réduit d'un facteur 10. Certains l'ont utilisée ayant comme objectif la suppression des mots apparaissant dans moins de k documents [49, 50].

Gain d'information .

Le gain d'information permet de quantifier le nombre de bits d'informations nécessaires pour la catégorisation de documents sachant la présence ou non d'un mot dans le document [10]. Selon les études de [48, 51] le gain d'information est le meilleur critère de sélection des mots-clés. Toutefois [52, 10] ont conclu que la mesure d'information mutuelle donne de meilleures performances.

Information mutuelle .

Pour mesurer la quantité d'information que présente un mot w_i pour un thème T_j, nous donnons la relation (2.1)[53] :

$$I(w_i, T_j) = P(w_i, T_j) log \frac{P(w_i, T_j)}{P(w_i)P(T_j)} \tag{2.1}$$

$P(w_i)$et $P(T_j)$ représentent respectivement la probabilité a priori du mot w_i et la probabilité a priori du thème T_j . $P(w_i, T_j)$ est la probabilité conjointe de w_i et T_j. Si $I(w_i, T_j)$ est élevée alors une forte relation existe entre le mot et le thème. Si, toutefois, elle est nulle cela veut dire qu'il y a une indépendance totale entre eux.

Brun a testé les performances de ces quatre méthodes de sélection de paramètres "features", en utilisant trois méthodes d'identification, en l'occurrence la méthode TFIDF, le modèle unigramme thématique et la méthode SVM [10]. Les résultats obtenus ont montré que la méthode de fréquence de mots est celle la plus performante

malgré sa simplicité, 78.46 % en terme de Rappel. Cependant, le gain d'information est classé le dernier, 77.4 %.

Dans la partie suivante de ce chapitre, nous présenterons des notions sur les deux méthodes que nous avons utilisées en identification de thème, en l'occurrence la SVM et la M-SVM.

2.3 SVMs et M-SVMs

Les SVMs sont un ensemble de méthodes appelées méthodes d'apprentissage supervisé "supervised learning methods". Elles sont utilisées notamment dans la classification et dans la régression. L'apprentissage supervisé est une technique "machine learning" dont le but est *de* créer une fonction à partir des données d'apprentissage. Ces dernières sont constituées de paires de vecteurs d'entrées, et de sorties désirées. La sortie de la fonction peut prédire la classe de l'objet d'entrée, on parle donc de classification, comme elle peut être continue, il s'agit alors de régression. L'apprentissage supervisé peut générer deux types de modèles : un modèle global, le plus répandu, et un ensemble de modèles locaux. Dans ce manuscrit, nous étudions la méthode SVM qui a comme but la catégorisation biclasse, nous traiterons ensuite la M-SVM "Multi-class Support Vector Machines" qui traite la catégorisation multi-classes.

2.3.1 Notion du risque empirique

Le modèle global consiste à trouver une fonction g, sachant un ensemble de points $(x, g(x))$. Ceci engendre une perte causée par la prédiction de la valeur de g pour un point donné x_i : $L(f(x_i), g(x_i))$, $f(x_i)$ étant la valeur prédite de $g(x_i)$. Le risque associé à une fonction f est défini comme suit :

$$R(f) = \sum L(f(x_i), g(x_i)).p(x_i) \tag{2.2}$$

le but est de trouver une fonction f^* qui rend minimal $R(f^*)$. Comme la variable $p(x_i)$ n'est pas connue, le risque empirique est représenté par l'équation (2.3) :

$$\tilde{R_n}(f) = \frac{1}{n} \sum L(f(x_i), y_i) \tag{2.3}$$

La sélection d'une fonction qui rend minimal est connue sous le terme "minimisation du risque empirique". Le risque réel est borné par deux quantités : le risque empirique et l'intervalle de confiance. Ce dernier est fonction de la valeur m/h. m étant la taille de l'échantillon, et h est la VC-dimension du modèle [54]. Si la quantité de m/h est suffisamment grande, la minimisation du risque empirique garantit une faible valeur du risque réel. Par contre dans le cas ou m/h est petit, l'intervalle de confiance aura une valeur importante, et la seule minimisation du risque ne suffit pas à réaliser cette tâche. C'est pour cela que Vapnik propose le principe de minimisation du risque structurel, basé sur la minimisation conjointe du risque empirique et l'intervalle de confiance [54].

Dans la section suivante, nous présentons les Support Vector Machines, et comment elles font la classification linéaire et celle non linéaire..

2.3.2 Support Vector Machines (SVM)

L'idée de la méthode SVM est de créer un hyperplan optimal qui sépare les deux classes avec une marge maximale. Autrement dit, cet hyperplan est situé à la distance maximale des vecteurs les plus proches parmi l'ensemble des exemples. L'extraction des paramètres d'un tel hyperplan se fait en résolvant le problème d'optimisation de la programmation quadratique (Quadratic Programming). Nous exposons ci-dessous les deux types de classification par la SVM, en l'occurrence la classification linéaire et celle non linéaire.

La classification linéaire .

Soit un ensemble $S = \{(x_1, y_1), ..., (x_m, y_m)\}$linéairement séparable. La solution du problème d'optimisation mathématique suivant fournit l'hyperplan optimal [54]. Cela revient à minimiser la quantité $\|\omega\|_2^2/2$ tel que :

$$\begin{cases} \omega.x + b \geq 1 & si\, y_i = +1 \\ \omega.x + b \leq -1 & si\, y_i = -1 \end{cases} \tag{2.4}$$

y_i étant la classe du vecteur x_i.

Ces deux inéquations peuvent être présentées par l'inéquation 2.5 :

$$y_i((\omega.x_i) + b) \geq 1 \,\forall\, i \in \{1, ..., m\} \tag{2.5}$$

Il s'agit d'un problème quadratique dont la fonction objective est à minimiser. Cette fonction objective est le carré de l'inverse de la double marge. L'unique contrainte stipule que les exemples doivent être bien classées et qu'ils ne dépassent pas les hyperplans canoniques.

L'idéal est de trouver un hyperplan optimal, c'est-à-dire à marge maximale. La valeur de cette dernière étant égale à $2/\|\omega\|_2$, ceci revient donc à minimiser $\|\omega\|_2^2$.

Un exemple d'un ensemble de données séparables est montré dans la figure (2.1).

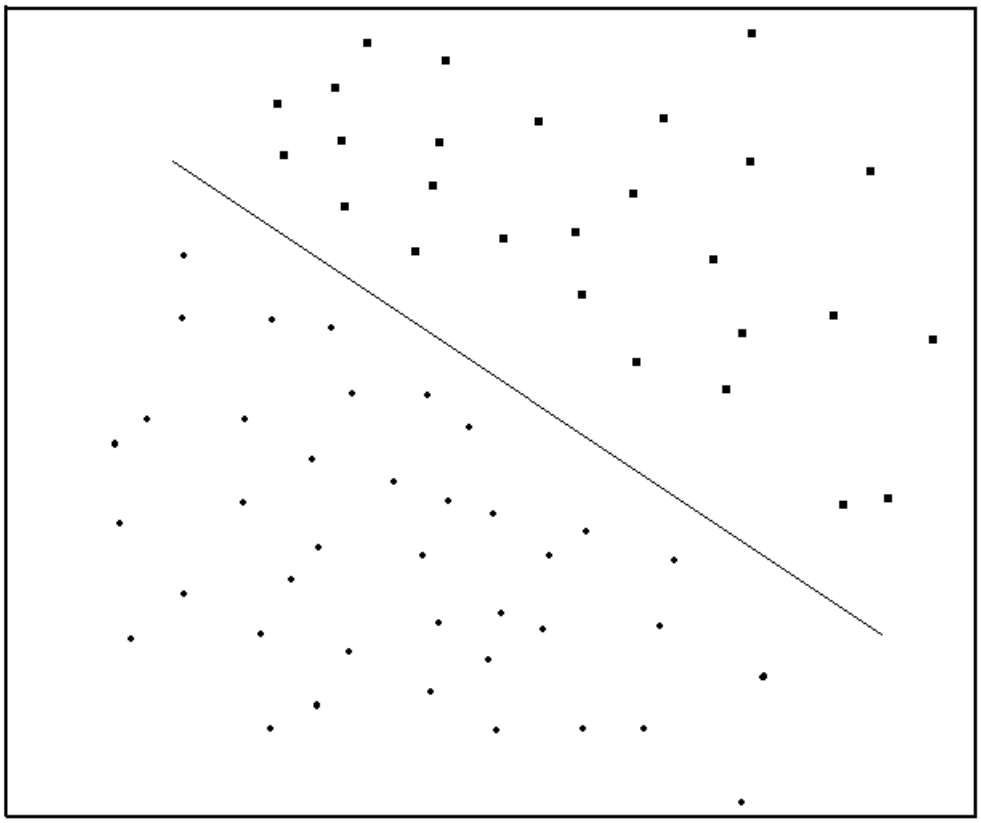


FIG. 2.1 – Données linéairement séparables

Dans cette formulation, les variables à fixer sont les composantes ω_i et b. D'un point de vue "Machine Learning", ces variables correspondent aux λ_i de la machine d'apprentissage. Le vecteur ω possède un nombre de composantes égal à la dimension de l'espace d'entrée. En gardant cette formulation telle quelle, des problèmes similaires à ceux des méthodes classiques du "Machine Learning" (overfitting, curse of dimensionality) apparaissent. Pour éviter cela, il est nécessaire d'introduire ce qu'on appelle dual du problème. Les variables du problème primal sont appelées variables primales, et les variables du problème dual sont appelées les variables duales qui n'interviennent pas dans le primal.

Pour dualiser le problème initial "le primal", le Lagrangien doit être formé. Il s'agit de faire renter les contraintes dans la fonction objective et de pondérer chacune d'entre elles par une variable duale :

$$L_p(\omega, b, \lambda) = \frac{1}{2}(\omega.\omega) - \sum_{i=1}^{m} \lambda_i \left[y_i \left((\omega.x_i) + b \right) - 1 \right] \tag{2.6}$$

En appliquant le principe de Kuhn-Tucker, le point-selle ω_0, b_0, λ_0 du lagrangien présenté par la formule (2.6) vérifie l'équation (2.7).

$$\lambda_i^0 \left[y_i \left(\left(\omega^0.x_i \right) + b^0 \right) - 1 \right] = 0, i \in \{1, ..., m\} \tag{2.7}$$

Les variables duales λ_i sont appelées *multipicateurs de Lagrange*. Les x_i qui vérifient l'équation 2.7 sont dits vecteurs supports. En partant du principe de Fermat, concernant l'annulation des dérivées partielles du Lagrangien, nous obtenons les relations (2.8) et (2.9) qui permettent d'obtenir l'hyperplan optimal :

$$\omega^0 = \sum_{i=1}^{m} \lambda_i^0 y_i x_i \tag{2.8}$$

et

$$\sum_{i=1}^{m} \lambda_i^0 y_i = 0 \tag{2.9}$$

La forme duale du Lagrangien s'écrit donc comme présentée dans l'équation (2.10) :

$$W(\lambda) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i.x_j) \tag{2.10}$$

L'équation de l'hyperplan optimal représenté par (2.11) s'obtient en maximisant $W(\lambda)$ avec $\lambda_i \geq 0$:

$$\sum_{i=1}^{m} \lambda_i^0 y_i (x.x_i) + b_0 = 0 \tag{2.11}$$

Le signe du membre gauche de l'équation (2.11) permet d'attribuer la classe d'un vecteur test x.

Ce que nous avons présenté précédemment c'est le cas d'une séparation linéaire. Dans le cas d'une séparation non linéaire, il suffit de remplir la condition (2.12) :

$$\begin{cases} \omega.x + b \geq 1 - \xi_i & si\, y_i = +1 \\ \omega.x + b \leq -1 + \xi_i & si\, y_i = -1 \end{cases} \tag{2.12}$$

$$\text{avec } \xi_i \geq 0 \, \forall \, i \in \{1, ..., m\}$$

La figure (2.2) présente le cas d'une séparation non linéaire :

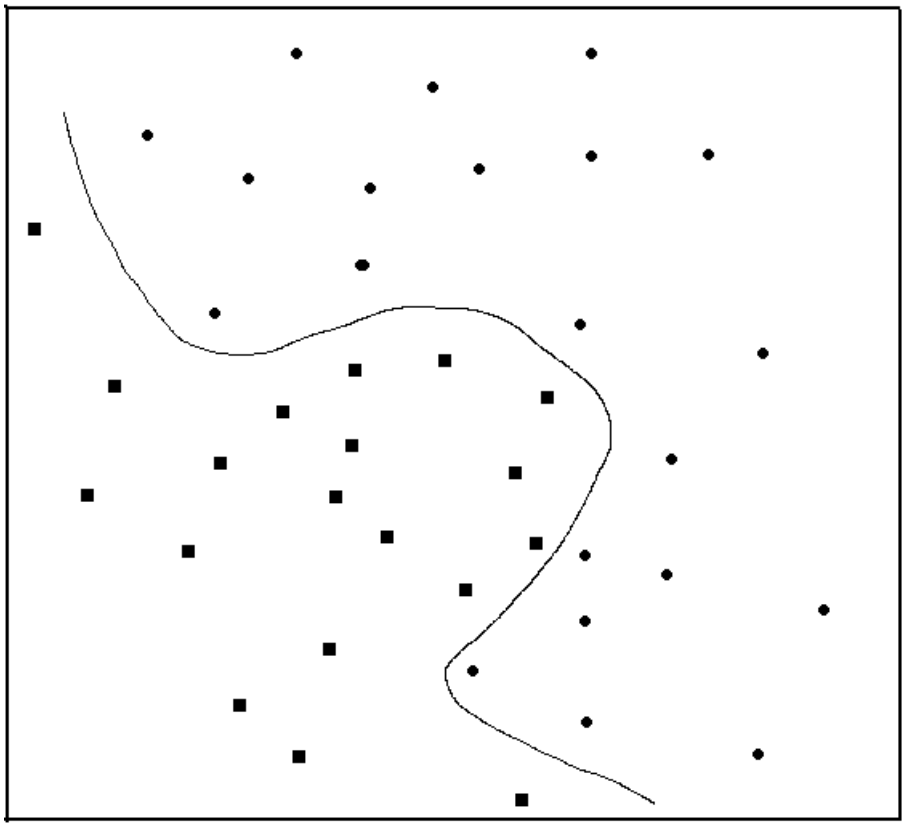


FIG. 2.2 – Données non séparables

Nous cherchons maintenant à minimiser la quantité présentée par l'équation (2.13) sous la contrainte (2.12) :

$$\phi(\omega, \xi) = \frac{1}{2}(\omega.\omega) + C\sum_{i=1}^{m}\xi_i \tag{2.13}$$

En suivant les mêmes étapes que dans le cas séparable, c'est-à-dire en maximisant le lagrangien dual, avec cette fois-ci $0 \leq \lambda_i \leq C$ pour $i \in \{1, ..., m\}$, nous obtenons un hyperplan optimal, toutefois il permet une appartenance non désirée de quelques échantillons à une classe donnée.

La classification non-linéaire .

Nous avons vu dans la section précédente, la difficulté de réaliser une bonne classification dans le cas non séparable. La classification non linéaire surmonte cette difficulté en représentant les vecteurs d'apprentissage dans un espace de dimension suffisamment grande. La procédure de la construction d'un hyperplan optimal est presque similaire à celle déjà vue dans la section précédente, en effet on remplace

le produit scalaire des vecteurs d'entrée par celui des vecteurs de l'espace de représentation $(e_i.e_j)$ qui est égal à $K(x_i, x_j)$. La quantité est appelée noyau, c'est une fonction symétrique qui vérifie les conditions de Mercer.

Le lagrangien dual aura la forme présentée par l'équation (2.14) :

$$W(\lambda) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j K(x_i.x_j) \tag{2.14}$$

En le maximisant sous les contraintes $0 \leq \lambda_i \leq C$ pour $i \in \{1, ..., m\}$, on obtiendra l'équation de l'hyperplan optimal présentée par la formule (2.15) :

$$\sum_{i=1}^{m} \lambda_i^0 y_i K(x.x_i) + b_0 = 0 \tag{2.15}$$

La fonction de décision non-linéaire s'exprime comme suit par l'équation (2.16) :

$$f(x) = sgn \left[\sum_{i=1}^{N} \lambda_i y_i K(x.\omega_i) + b \right] = 0 \tag{2.16}$$

N est le nombre des vecteurs supports.

2.3.3 La M-SVM

La M-SVM (Multi-category Support Vector Machines) a pour objectif la discrimination multi-classes. Les méthodes employées sont diverses. La première catégorie la plus basique est appelée un contre les autres. Une autre approche comme celle présentée dans [55] et [56] traite l'apprentissage des M-SVMs par l'intermédiaire d'un problème de minimisation quadratique conformément à ce qui est fait lorsqu'il y a deux classes. Une méthode récente a été présentée par Guermeur [57], elle est basée sur le résultat de la convergence uniforme. Elle permet de trouver un compromis satisfaisant entre la complexité et la performance de l'apprentissage. Des notions détaillées sur les principes de cette méthode peuvent être trouvées dans [57, 58].

La M-SVM linéaire (cas séparable) .

Considérons l'ensemble de variables :

$$S_N = \{(x_1, C(x_1)), ..., (x_i, C(x_i)), ..., (x_N, C(x_N))\}$$

S_N est un ensemble de points x_i attribués chacun à sa classe $C(x_i)$. Avec $C(x_i)$ appartenant à l'ensemble de classes $\{1, ..., Q\}$. La fonction qui permet le calcul des hyperplans optimaux pour la catégorisation multi-classes est donnée par la relation (2.17) :

$$h_k(x) = \omega_k^T x + b_k \tag{2.17}$$

avec $[\omega_k] \in R^{Qd}$ et $[b_k] \in R^Q$. $h_k(x)$ est la valeur de la fonction calculée par la classe C_k pour un point donné x. Pour ceci, on est amené à minimiser la fonction objective (2.18) [57] :

$$J(\omega) = \frac{1}{2} \sum_{k=1}^{Q-1} \sum_{l=k+1}^{Q} \|\omega_k - \omega_l\|_2^2 \tag{2.18}$$

sous les contraintes :

$$\forall i \in \{1, ..., N\}, \forall k \in \{1, ..., Q\} / C_k \neq C(x_i)$$

$$\omega_{C(x_i)} - \omega_k)^T x_i + b_{C(x_i)} - b_k \geq 1 \tag{2.19}$$

Ceci revient à trouver le point-selle du lagrangien présenté par la formule (2.20) :

$$L(\omega, b, \alpha) = \tfrac{1}{2} \textstyle\sum_{k=1}^{Q-1} \sum_{l=k+1}^{Q} \|\omega_k - \omega_l\|_2^2 -$$

$$\sum_{i=1}^{N} \sum_{k=1, k \neq C(x_i)}^{Q} \alpha_{ik} \left[(\omega_{C(x_i)} - \omega_k)^T x_i + b_{C(x_i)} - b_k - 1 \right] \tag{2.20}$$

Les coefficients α_{ik} sont des multiplicateurs de Lagrange non-négatifs.

En mettant $\dfrac{\partial L(\omega, b, \alpha)}{\partial b} = 0_Q$ on obtient le lagrangien simplifié présenté par l'équation (2.21) :

$$L(\omega, \alpha) = \frac{1}{2} \textstyle\sum_{k=1}^{Q-1} \sum_{l=k+1}^{Q} \|\omega_k - \omega_l\|_2^2 - \sum_{i=1}^{N} \sum_{k=1}^{Q} \alpha_{ik}$$

$$\left[(\omega_{C(x_i)} - \omega_k)^T x_i \right] + \sum_{i=1}^{N} \sum_{k=1}^{Q} \alpha_{ik} \tag{2.21}$$

Pour des raisons théoriques et techniques, le problème de la programmation quadratique est résolu dans l'espace dual. Ceci est réalisé souvent en utilisant une matrice régulière de Hessian qui est obtenue en changeant les variables Primal. Cela permet d'aboutir à l'équation de l'hyperplan qui sépare les catégories C_k et C_l.

$$\frac{1}{Q} \left\{ \sum_{x_i \in C_k} \sum_{m=1}^{Q} \alpha_{im} x_i^T - \sum_{i=1}^{N} (\alpha_{ik} - \alpha_{il}) x_i^T - \sum_{x_i \in C_l} \sum_{m=1}^{Q} \alpha_{im} x_i^T \right\} x + b_k - b_l = 0 \tag{2.22}$$

La M-SVM non-linéaire (cas non-séparable) .

Dans le cas non séparable, on introduit des variables positives ξ_{ik}, $(1 \leq i \leq N)$, $(1 \leq k \leq Q)$, On doit minimiser la fonction objective :

$$J(\omega) = \frac{1}{2} \sum_{k=1}^{Q-1} \sum_{l=k+1}^{Q} \|\omega_k - \omega_l\|_2^2 + C \sum_{i=1}^{N} \sum_{k=1}^{Q} \xi_{ik} \tag{2.23}$$

sous les contraintes :

$$\forall i \in \{1, ..., N\}, \forall k \in \{1, ..., Q\} / C_k \neq C(x_i)$$

$$\omega_{C(x_i)} - \omega_k)^T x_i + b_{C(x_i)} - b_k \geq 1 - \xi_{ik}$$

La valeur C permet à l'utilisateur de faire un compromis entre la complexité et l'efficacité de l'apprentissage. Nous sommes donc face à un problème de programmation quadratique, qui sera résolu de la même manière effectuée dans le cas linéaire. Il est à noter que les M-SVMs non-linéaires peuvent être obtenues en remplaçant, dans l'équation (2.22) les produits scalaires $x_i^T x$ par des noyaux $K(x_i, x)$, en respectant le théorème de Mercer.

2.4 Conclusion

Dans ce chapitre, nous avons abordé les méthodes de l'état de l'art utilisées dans ce manuscrit, en l'occurrence la TFIDF, la SVM et la M-SVM. Au début, nous avons exposé les deux tâches essentielles : la représentation des documents et la sélection des paramètres. Dans le cas de l'identification de thèmes, la méthode Bag of words paraît la plus efficace pour la représentation des documents. La sélection de paramètres a suscité un nombre de travaux ayant le but de fournir la méthode la plus adéquate qui puisse aboutir à des performances maximales. Selon [10] c'est la méthode basée sur l'information mutuelle qui permet au classifieur TFIDF d'atteindre ses meilleures performances. Toutefois ces dernières ne sont que légèrement supérieures à celles fournies par les autres méthodes de sélection pour le même classifieur.

La méthode SVM est conçue pour effectuer la séparation entre deux classes. Ceci est réalisé en créant un hyperplan optimal séparateur avec une marge maximale. Lorsque les données sont linéairement séparables, il suffit de réaliser une classification linéaire. Cependant, dans le cas où elles ne le sont pas, la classification non-linéaire est sollicitée, en représentant ces données dans un espace de dimension suffisamment grande. Ceci est réalisé en utilisant des fonctions noyaux.

Le fait que la SVM est limitée à la séparation bi-classes présente un inconvénient que nous avons voulu surmonter en utilisant la M-SVM. Pour cette dernière, le

problème c'est de chercher l'équation des hyperplans optimaux qui permet de faire la catégorisation de Q classes (avec $Q \geq 3$). L'idée est de construire un modèle multivariable $h_k(x) = \omega_k^T x + b_k$.

Chapitre 3

Expériences et résultats

3.1 Introduction

Les méthodes exposées dans ce chapitre sont des méthodes statistiques issues de l'état de l'art. L'application des deux premières méthodes TFIDF et SVM a donné de résultats satisfaisants [3, 5, 4]. Nous avons ensuite testé la M-SVM, une version plus développée de la SVM, pour la première fois dans le domaine de l'identification de thèmes. Toutefois, le corpus dédié à cette expérience est moins important que celui utilisé dans celles de la TFIDF et la SVM en raison du temps de calcul énorme que l'outil M-SVM mette pour accomplir la tâche d'identification.

Les documents dont nous nous sommes servis pour réaliser nos expériences ont subi une transformation les rendant sous un format adéquat pour l'utilisation des méthodes d'identification. Des recherches ayant été déjà effectuées sur les méthodes de représentation de documents [10], notre objectif n'étant pas de réaliser une étude sur la manière avec laquelle nous représentons nos données, nous avons opté pour une méthode très populaire en catégorisation de textes, celle utilisée par le classifieur TFIDF.

Nous avons donc abordé quelques expériences ayant pour objectif une identification rapide de thèmes, et ce en utilisant seulement les premiers mots du document de test. Cette accélération du processus d'identification est nécessaire dans le cas d'adaptation des modèles de langage dans un système de reconnaissance automatique de la parole.

Dans ce qui suit, nous allons commencer par citer quelques spécificités de l'Arabe, langue de notre corpus, ensuite exposer les étapes de représentation de documents constituant le corpus d'apprentissage, à partir duquel nous procédons à la construction du vocabulaire, en se basant sur la fréquence des mots.

3.2 Corpus

3.2.1 Quelques spécificités de la langue Arabe

La langue Arabe s'identifie par rapport aux autres langues latines par différents aspects. L'un de ces derniers est qu'elle s'écrit de droite à gauche, contrairement à la langue anglaise ou française. Une autre caractéristique de cette langue est qu'elle est économique dans le cas de la représentation textuelle d'un énoncé, une particularité qui paraît intéressante et présente un inconvénient en même temps. En effet son aptitude à regrouper plusieurs mots en un seul, constitue d'un coté un gain de temps et d'espace, mais d'un autre coté elle nécessite un travail de plus dans le domaine de traitement automatique des textes.

Énoncé en Français	Équivalent en Arabe
Et ils y entrent.	fasayadkhulunaha

TAB. 3.1 – Exemple présentant l'équivalence en Français d'un mot arabe

Une autre caractéristique est l'absence des voyelles dans les textes, ce qui est non habitué pour les autres langues. Si on prend l'exemple présenté dans la table (3.1), et on en enlève les voyelles on obtiendra "fsydkhlunha". On mentionne que "u" et "a" sont des prolongements appelés voyelles longues. La conséquence de l'omis intentionnel de ces voyelles appelées mouvements chez les grammairiens arabes, rend le mot plus comprimé.

Lorsqu'on parle de la taille d'un corpus, on doit prendre en considération la langue. La forme compacte qui caractérise la langue Arabe fait qu'un corpus dans cette langue constitué de n millions mots pourrait être équivalent à un corpus indo-européen composé de $k \cdot n$ millions mots. En effet la taille du corpus issu du journal Le monde pour la période de quatre années est 80 millions mots [10]. Tandis que celle du corpus extrait de AFP Arabic Newswire, en 2001 par LDC (Linguistic Data Consortium) pour une période avoisinant 7 années est de 76 millions mots [59, 60].

En outre en comparant notre corpus extrait du journal Omanais Alwatan avec celui issu du journal Le monde pour la même durée, on remarque que la taille de ce dernier est double de celle du corpus arabe [3], table (3.2).

Journal	Période	Taille du corpus
Alwatan	1 année	9.8 Millions
Le monde	1 année	20 Millions

TAB. 3.2 – Différence de taille de corpus des deux journaux Alwatan et Le monde

3.2.2 Collecte de corpus

Nous avons utilisé un outil de téléchargement, en l'occurrence WinHttrack (Annexe C.1). Ce dernier permet de stocker toutes les pages html d'une manière orga-

nisée, dans un répertoire qui peut être défini par l'utilisateur. L'attribution d'étiquettes thématiques aux articles est effectuée manuellement au préalable. Seulement, les données sont stockées d'une manière déterminée dans des pages sous le format html. A partir de ces dernières on peut accéder aux articles relatifs à un thème parmi ceux existants dans le journal. L'extraction des textes est réalisée en utilisant un outil que nous avons implémenté pour cet effet.

Le corpus dont nous disposons est donc réparti sur six thèmes, en l'occurrence : Culture, Religion, Économie, Local, International et Sports.Toutefois cette attribution des thèmes aux documents n'est pas toujours exacte, ceci est dû parfois aux articles qui traitent un thème parlant d'un autre thème comme le texte présenté dans la figure (3.1). Et parfois d'autres thèmes, comme le thème culture, ont besoin d'être divisés en sous-thèmes pour pouvoir aboutir à des résultats satisfaisants.

الاحمر لون الفوز في الرياضة

عالما اجناس بريطانيان يربطان الفوز باللون الاحمر الذي يعبر عن سيطرة الذكر في عالم الحيوانات.
ميدل ايست اونلاين

لندن - كشفت دراسة أجراها فريق من الباحثين في جامعة دورهام في إنجلترا أن فرص الرياضيين أفرادا أو فرقا في الفوز تكون أعلى إذا ارتدوا اللون الاحمر.
وجاء في مقال الباحثين الذي نشر في صحيفة جورنال العلمية البريطانية "بين مختلف الرياضات وجدنا أن ارتداء أزياء باللون الاحمر يرتبط بشكل مستمر بزيادة فرص الفوز".
وربط الفريق بقيادة عالمي الاجناس راسل هيل وروبرت بارتون الفوز باللون الاحمر الذي يعبر عن سيطرة الذكر في عالم الحيوانات.
وقال العالمان "على الرغم من أن الالوان الاخرى موجودة أيضا في عالم الحيوان إلا أن حضور وشدة اللون الاحمر يرتبط بسيطرة الذكر وبمستويات هرمون التستوستيرون (هرمون الذكورة)".
وخلص الباحثون إلى أن "في السلوك البشري يرتبط الغضب باحمرار لون البشرة بسبب تزايد اندفاع الدم في الجسم في حين أن الخوف يترتبط بزيادة شحوب الوجه في المواقف الخطرة أيضا". وطبق العلماء نظريتهم على أربع رياضات تنافسية خلال دورة الالعاب الاولمبية التي أقيمت في أثينا العام الماضي وهي الملاكمة والتايكوندو والمصارعة الرومانية والمصارعة الحرة حيث يرتدي اللاعبون إما أزياء حمراء أو زرقاء. واكتشف العلماء أن الذين ارتدوا اللون الاحمر كانوا فائزين في أغلب الاحيان.
وأظهرت فئات الوزن في كل رياضة النتيجة نفسها ففي نحو ١٩ من بين ٢٩ فئة كان الفائزون من أصحاب اللون الاحمر أكثر عددا مقارنة بستة فائزين فقط من اللون الازرق.
وانتهت الدراسة إلى أن الالوان الصناعية قد تؤثر على نتيجة المسابقات التي تتطلب مجهودا جسديا لدى البشر.

FIG. 3.1 – Exemple d'un texte traitant deux thèmes

L'idée du texte présenté dans la figure (3.1) est de montrer scientifiquement que la couleur rouge est la couleur qui aide à gagner dans les compétitions sportives. En utilisant les méthodes statistiques ce texte peut être assigné à la classe sport si les termes relatifs à ce thème sont prédominants. Comme il peut être considéré appartenant au thème science si c'est l'inverse. Mais le plus juste c'est de dire que ce texte est un texte scientifique traitant un autre thème qui est celui de sport.

Les deux principales sources de notre corpus sont "Alwatan" qui est un journal Omanais, et le journal "Akhbar Alkhaleej". Nous présentons dans la table (3.3) le nombre de mots N constituant chacun des corpus thématiques extraits du journal "Alwatan", avant et après enlèvement des mots outils M.O. Des informations sur ces derniers sont présentées dans la section 3.4.

Thèmes	N. de mots avant	N. de mots après
Culture	1.359.210	1.013.703
Religion	3.122.565	2.133.577
International	855.945	630.700
Économie	1.460.462	1.111.246
Local	1.555.635	1.182.299
Sports	1.423.549	1.067.281
Total	9.813.366	7.139.486

TAB. 3.3 – Taille du corpus "journal Alwatan" avant et après enlèvement des mots outils

3.3 Représentation de documents

Comme c'est mentionné ci-dessus, l'élaboration du corpus utilisé dans nos expériences est basée sur deux principales sources, en l'occurrence deux journaux arabophones, "Akhbar Alkhaleej" du sultanat d'Oman, et le journal koweitien "Alwatan".

Avant de procéder à l'identification de thèmes, nous avons commencé à faire la représentation des documents, dans le but d'appliquer, en premier lieu, la méthode TFIDF, et ensuite les autres méthodes de l'état de l'art comme la SVM. Comme le corpus utilisé dans nos expériences est obtenu du réseau Web, nous étions contraints d'extraire les documents qui sont sous le format html, en implémentant des outils appropriés pour cette opération. Un schéma simplifié montrant les étapes essentielles menant à la représentation des documents est présenté dans la figure (3.2).

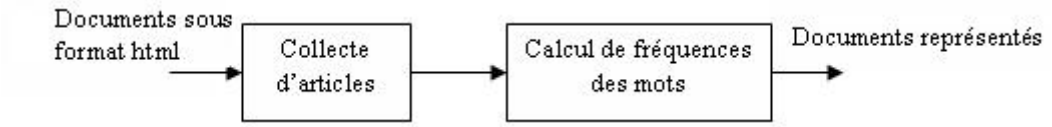


FIG. 3.2 – Transformation d'un document sous format html à un vecteur contenant des paramètres

Chaque document est bien évidemment constitué d'un nombre de mots. Pour le rendre adéquat au traitement, une opération le transformant à un format spécifique est nécessaire. Elle consiste à représenter chaque mot par une valeur bien définie. Ainsi un document constitué d'un ensemble de mots devient un vecteur contenant des éléments appelés "features" en anglais, caractérisant le document en question.

La figure (3.3) montre les mots appartenant à un texte extrait de notre corpus, et à la droite de ceux-ci un ensemble de valeurs représentant les "Term Frequency/Inverse Document Frequency), caractérisant les mots du texte.

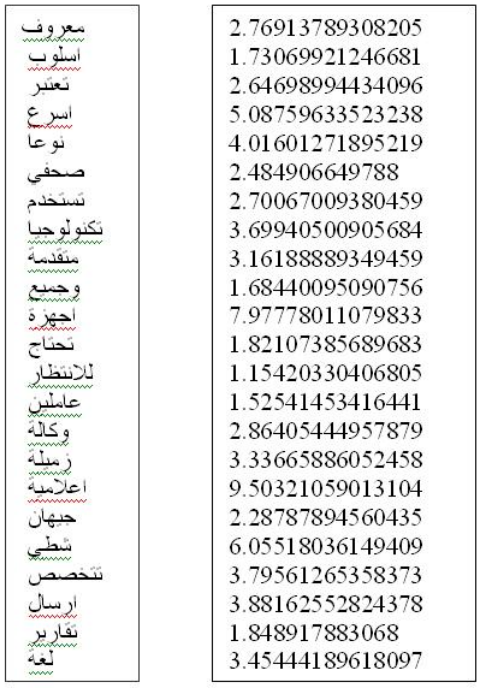


FIG. 3.3 – Exemple de représentation d'un texte par la méthode Bag of Words

3.4 Les mots outils

Les mots outils sont écartés en général lorsqu'on est devant un problème d'identification de thèmes, car ils ne sont pas caractérisants de ces derniers. Néanmoins des travaux dans [61] ont cité que de bonnes performances ont été obtenues en présence de ces mots. En effet, nous avons testé les performances de l'une des méthodes utilisées, en l'occurrence le classifieur TFIDF, dans les deux cas : en présence et en absence des mots outils. Les résultats que nous avons obtenus, cités dans la section 3.6.1 confirment que leur absence est une condition nécessaire pour avoir de bonnes performances.

La suppression des mots outils dans la plupart des textes arabes, y compris notre corpus, n'est pas aussi simple que celle effectuée dans les textes écrits en langues indo-européennes. Dans ces derniers, les mots outils "le", "la", "et", etc. sont séparés des autres mots, ce qui facilite la tâche de les enlever automatiquement. A l'inverse, pour l'Arabe, les articles définis, ainsi que plusieurs d'autres prépositions se trouvent toujours collés aux autres mots du texte. Outre cela, on rencontre souvent le "waw" (équivalent de "et") collé au mot qui le suit. Ce qui pose de difficultés en plus de les enlever.

3.5 Construction du vocabulaire

Pour construire notre vocabulaire, nous nous sommes basés sur une méthode connue : la fréquence de mots, bien qu'il existe d'autres méthodes comme la fréquence de documents, la mesure d'information mutuelle ou encore le gain d'information [10]. En effet dans [10] on trouve que pour chaque méthode d'identification est choisie une méthode de sélection de vocabulaire. A titre d'exemple pour le modèle unigramme, c'est la méthode de fréquence de documents qui est utilisée, par contre pour le classifieur TFIDF c'est la fréquence de mots qui est optimale. En outre, l'information mutuelle est sollicitée pour la SVM. Néanmoins, il n'y a pas de grandes différences dans les performances que fournissent ces méthodes de sélection. En effet, dans [48] et [51] il a été conclu que le gain d'information est la meilleure méthode de sélection de mots-clés du vocabulaire. Ce qui n'a pas été le cas dans [52] et [10] qui ont montré que la mesure d'information mutuelle mène à des performances qui dépassent celles obtenues en utilisant le gain d'information. Dans nos expériences, nous ne cherchons pas à trouver quelle est la méthode adéquate de sélection de vocabulaire autant que nous voulions étudier les performances des différentes méthodes d'identification de thèmes que nous appliquerons en utilisant corpus en langue Arabe. Le vocabulaire ne doit pas contenir tous les mots différents du corpus d'apprentissage, car ceci pourra facilement fausser les résultats. En effet, d'après [37] et [29], les mots dont la fréquence ne dépasse pas un certain seuil n'apportent aucune information. La valeur de ce seuil reste empirique, toutefois des études optent pour la valeur 3, d'autres ont choisi la valeur 5 [62].

3.6 Expérimentation des méthodes d'identification

Dans cette expérience, l'objectif est de tester les performances du classifieur TFIDF et la méthode SVM. Les thèmes à identifier sont : International, Local, Sports, et Économie. Le corpus d'apprentissage est extrait du journal "Akhbar Al Khaleej" et il est composé d'environ 2500 articles. Dans la table (3.8) nous présentons les thèmes par le nombre de mots distincts les constituant, avant et après l'écartement de ceux dont la fréquence n'excédant pas la valeur 3.

Mots distincts	Sports	Local	International	Économie
Avant	22822	31292	25540	18994
Après	7357	10883	8687	6502

TAB. 3.4 – Nombre de mots distincts caractérisant chacun des thèmes avant et après l'enlèvement des mots dont la fréquence < 3 (Taille du corpus 2500 articles).

3.6.1 Le classifieur TFIDF

Comme il est décrit dans le deuxième chapitre, le classifieur TFIDF est basé sur le calcul de la distance cosine entre deux vecteurs, dans le but d'en mesurer la

similarité. Un thème T_i est attribué à un document d si la valeur de la similarité obtenue entre ces deux entités est la plus grande.

Présence de mots outils .

Les articles constituant le corpus d'apprentissage sont utilisés pour construire le vocabulaire général dont la taille initiale obtenue est de 74000 mots distincts. Nous avons ensuite écarté les mots dont la fréquence est inférieure à 3, c'est à dire les mots n'apportant aucune information, le vocabulaire ainsi obtenu est constitué de 20000 mots.

Performances	Rappel (%)	Précision (%)	F_1(%)
Sports	84.37	75	79.41
Économie	70.31	58.44	63.83
Local	39.06	83.33	53.19
International	82.81	68.33	74.87

TAB. 3.5 – Performances du classifieur TFIDF en présence de mots outils

Absence de mots outils .

L'absence de mots outils a contribué dans l'amélioration du taux d'identification. Ce qui a été bien évidemment prévu, car leur présence nuit à l'identification en ne fournissant aucune information pertinente, en plus elle alourdit le calcul. Ceci concerne les travaux qui se basent sur les mots comme c'est notre cas : l'identification de thèmes. En effet, en se référant à d'autres types spécifiques de classification comme la classification de phrases 'sentence classification', il a été montré que leur présence, au contraire améliore les résultats [63].

Performances	Rappel (%)	Précision (%)	F_1(%)
Sports	88.52	96.43	92.30
Économie	88.10	82.22	85.06
Local	87.23	86.31	86.77
International	97.77	95.65	96.70

TAB. 3.6 – Performances du classifieur TFIDF en absence des mots outils

Amélioration de la représentation des thèmes et les résultats correspondants .

Pour une meilleure représentation des thèmes, nous avons élevé la taille du corpus à 5120 articles. A partir de ce corpus nous avons construit notre vocabulaire dont la taille préliminaire a été de 100000. Cette valeur est réduite à 43000 après avoir supprimé les mots dont la fréquence d'apparition est inférieure à 3. Le nombre de

mots constituant notre nouveau corpus total avoisine trois (03) millions de mots. Le corpus d'apprentissage est formé de 90 % du corpus total, ce qui correspond à 4608 articles. Les 10 % restants sont réservés pour le test. Dans la table (3.7) nous présentons le corpus total et celui d'apprentissage après enlèvement de mots outils, en fournissant en détail le nombre de mots attribués à chacun des quatre thèmes [3, 5].

Thème	sports	Local	International	Économie
Corpus total	628.000	893.000	754.000	578.000
Corpus d'apprentissage	485.000	680.000	567.000	440.000

TAB. 3.7 – corpus après écartement des mots outils

Le nombre de mots distincts utilisés pour chacun des thèmes est dressé dans la table (3.8) .

Mots distincts	sports	Local	International	Économie
Avant	41980	55390	45145	36088
Après	15078	21108	17213	13632

TAB. 3.8 – Nombre de mots distincts caractérisant chacun des thèmes avant et après l'enlèvement des mots dont la fréquence n'excédant pas la valeur 3 (Taille du corpus 5120 articles)

Les performances de la méthode TFIDF se sont améliorées en augmentant respectivement la taille du corpus et celle du vocabulaire général qui sont respectivement de trois millions mots et 43000. Nous résumons les valeurs du rappel, précision et la mesure F_1 dans la table (3.9).

Performances	Rappel (%)	Précision (%)	Mesure F_1(%)
Sports	94.53	100	97.19
Économie	85.15	85.82	85.48
Local	85.94	79.71	82.71
International	97.65	99.20	98.42

TAB. 3.9 – performances en terme de rappel , de précision et de mesure F_1, avec une taille de vocabulaire égale à 43000 .

3.6.2 Expérimentation de la SVM

Pour pouvoir utiliser l'outil SVM^{light} de Joachims et faire la discrimination bi-classes nous avons organisé les documents sous un format approprié. La première étape consiste à rassembler N documents (positifs) du premier thème (N=1152 dans notre cas) et N documents (négatifs) du deuxième thème, en vu de réaliser l'étape

d'apprentissage. Dans la deuxième étape qui est celle de test, nous avons prévu 128 documents pour le premier thème et pareil pour le deuxième. Ceci est effectué $n(n-1)/2$ fois pour chaque paire de thèmes. Cette méthode est appelée one-against-one. Dans la table (3.10) nous présentons respectivement les performances en terme de rappel, précision et de la mesure F_1 [5, 3].

Nous remarquons que les résultats pour les thèmes Mondial et Sport, sont très satisfaisants. Pour les deux autres thèmes : Local et Économie, les performances sont inférieures. L'explication est que le corpus du thème Local contient parfois des articles traitant de l'économie, et vice versa. La méthode SVM se montre supérieure, toutefois les résultats concernant ces deux thèmes sont, relativement moins importants. Les résultats montrés dans la table (3.10) concernent la classification binaire, dite la méthode one-against-one, qui veut dire la différenciation d'un élément d'un autre. Et comme la méthode SVM ne peut réaliser une classification directe de l'ensemble des thèmes dont le nombre est supérieur à 2, nous avons utilisé une autre méthode appelée one-versus-rest, qui consiste à classifier un thème par rapport à une mixture des autres thèmes. Les résultats sont légèrement dégradés vis-à-vis de ceux présentés dans la table (3.10), néanmoins ils sont toujours satisfaisants, en l'occurrence les taux de rappel et de précision sont respectivement de 95.44 % et 96.26 %. Une autre alternative de la méthode one-versus-rest est la M-SVM que nous allons expérimenter dans les prochaines sections.

La table (3.11) présente une comparaison des résultats issus des deux méthodes SVM et TFIDF en fonction du rappel, précision et la mesure F1.

Performances	Rappel (%)	Précision (%)	Mesure F_1(%)
TFIDF	90.82	91.18	90.95
SVM	97.26	98.52	97.88

Tab. 3.11 – Tableau récapitulatif des performances de la SVM et du classifieur TFIDF

3.7 Vers une identification dynamique

L'un des objectifs de l'amélioration des performances des méthodes utilisées en identification de thèmes, est l'adaptation des modèles de langage d'un système de reconnaissance automatique de la parole. Par conséquent, il est judicieux d'accélérer cette tâche d'identification de thèmes, pour que la reconnaissance s'effectue en temps réel. Pour cela nous réalisons l'expérience exposée ci-dessous, qui consiste à prendre en compte seulement les n premiers mots contenus dans l'article réservé pour le test. Nous allons commencer par prendre de petites valeurs de n, ensuite nous l'augmenterons au fur et à mesure, dans le but de déterminer au-delà de quelle valeur les performances restent inchangées, et de ce fait il serait inutile de prendre en considération les mots qui viennent après, ce qui nous évitera un temps de calcul important.

Dans cette expérience [3, 4], la méthode utilisée est la TFIDF, et les données sont

Thèmes	International			Local			Économie			Sports		
Performances	Rap	Pré	F_1	Rap	Pré	F_1	Rap	Pré	F_1	Rap	Pré	F_1
International	-	-	-	99.22	100	99.61	100	99.22	99.61	100	100	100
Local	99.22	100	99.61	-	-	-	89.06	92.68	90.83	97.66	99.21	98.43
Economie	100	99.22	99.61	89.06	92.68	90.83	-	-	-	97.66	100	98.81
Sports	100	100	100	97.66	99.21	98.43	97.66	100	98.81	-	-	-

TAB. 3.10 – Performances de la méthode SVM en terme de rappel, précision et de la mesure F1

celles utilisées dans la section (3.6.1). Les trois figures (3.4), (3.5) et (3.6) présentent les performances de l'identification concernant les quatre thèmes en question, et cela en fonction des n premiers mots. Pour plus de détails nous avons dressé des tables représentatives des données dans l'annexe B. Pour $n = 40$ nous remarquons que les taux de rappel pour les thèmes (International et Local) ont atteint des valeurs élevées égales à celles obtenues en utilisant tous les mots constituant l'article de test. Il y a des thèmes qui atteignent des taux de rappel similaires à ceux obtenus de l'expérience, dans laquelle, nous avons considéré tous les mots constituant l'article de test. Voir la table (3.9), (et ce en utilisant seulement $n = 10$ mots pour le thème Local et $n = 40$ mots pour le thème International). Au-delà de ces valeurs de n, les taux ne changent pas trop. Par contre pour les deux autres thèmes, nous avons pour $n = 10$ mots, les taux de rappel qui sont égaux à 18.75 % et 57.03 % respectivement pour Sport et Économie. Au-delà de ces deux valeurs pour n, nous avons remarqué une amélioration presque linéaire des résultats. A partir de la valeur $n = 100$ mots, on a trouvé un taux de rappel égal à 82.81 % (pour le thème Économie) presque similaire à celui mentionné dans la table (3.9). Pour que les performances arrivent exactement au même niveau exposées dans cette dernière, en l'occurrence, 85.16 %, il a fallu choisir la valeur n=200 mots.

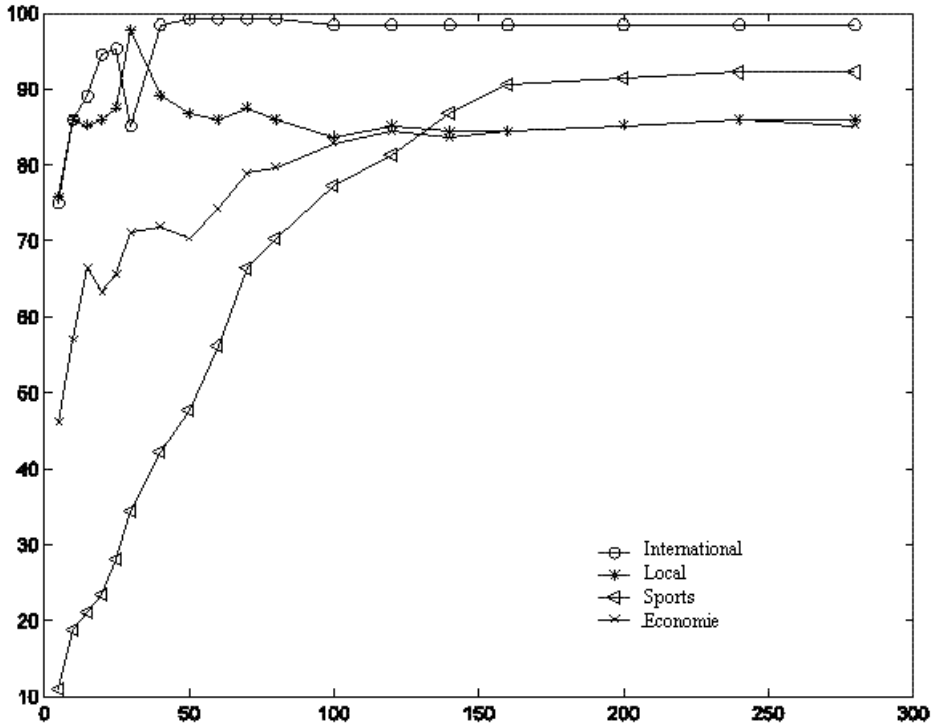


FIG. 3.4 – Performances en terme de rappel, en fonction des n premiers mots

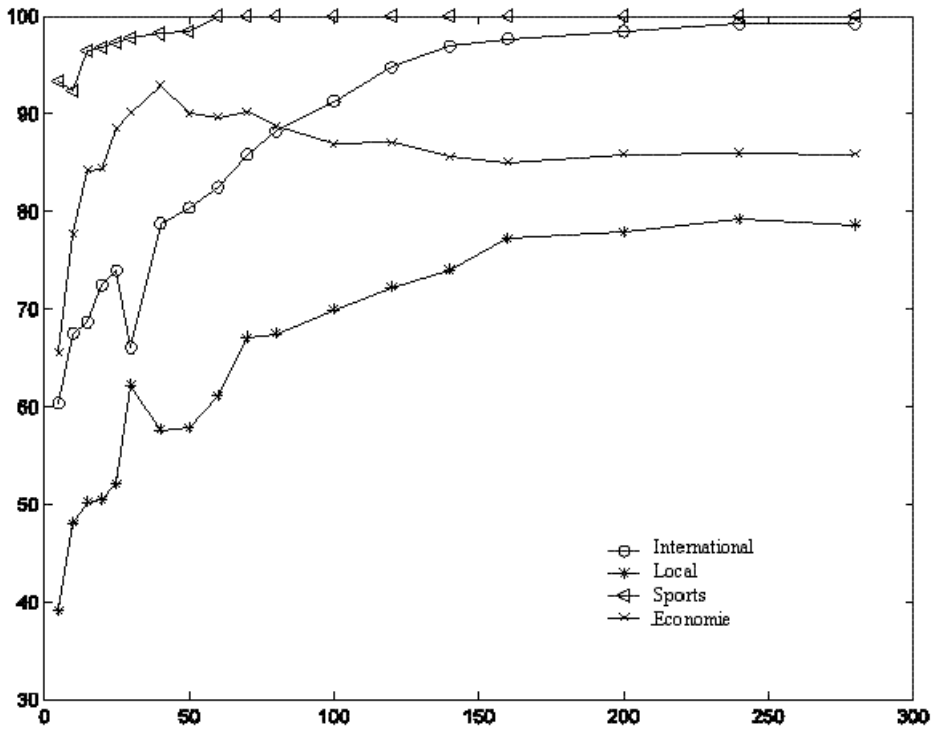


FIG. 3.5 – Performances en terme de précision, en fonction des n premiers mots

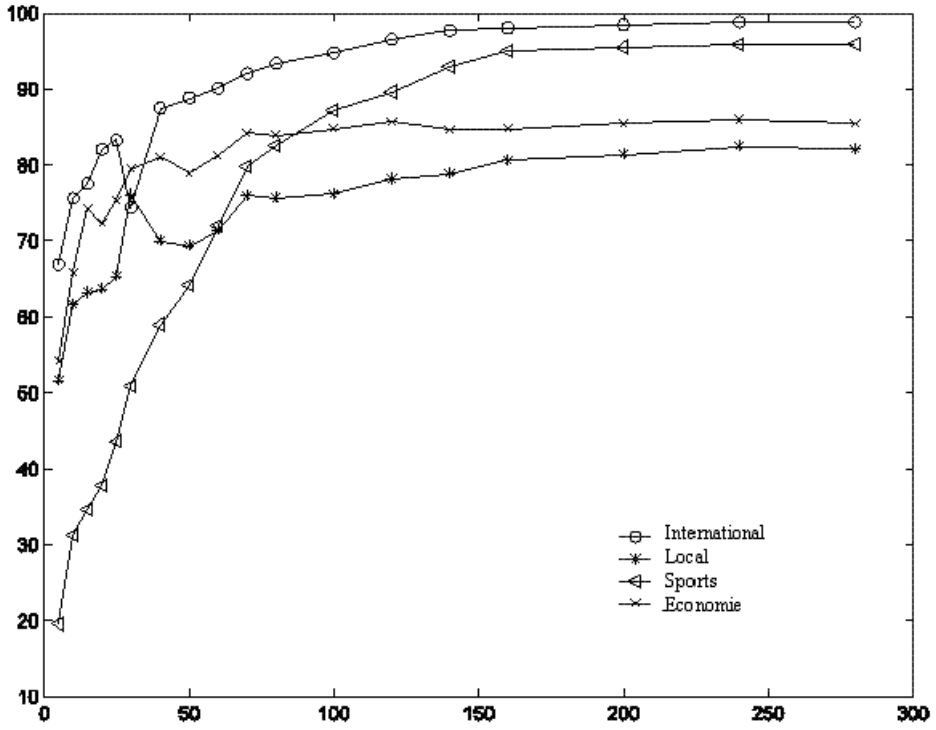


FIG. 3.6 – Performances en terme de mesure F_1, en fonction des n premiers mots

Pour le thème Sports, le taux de rappel augmente linéairement avec n. Pour$n = 240$ mots, le taux de rappel est égal à 92.19 %. Les deux thèmes Local et International ont atteint des taux élevés pour$n$ petit (n'ont pas besoin de beaucoup de mots, le taux de rappel atteint rapidement un bon résultat), tandis que les thèmes Sports et Économie ne les ont atteint que pour n plus grand (A chaque fois qu'on ajoute des mots le taux augmente légèrement). Ceci peut s'expliquer par le fait que les mots des thèmes Local et International sont plus représentatifs que ceux des autres thèmes. Nous avons remarqué que les articles étiquetés Sports et ceux étiquetés Économie sont souvent identifiés International ou Local pour des valeurs faibles et moyennes de n. Ce qui veut dire que les thèmes Économie et Sports sont représentés par des mots dont la fréquence est plus importante pour les thèmes International et Local. L'augmentation du corpus d'apprentissage est une solution judicieuse pour avoir de bons résultats, car cela permettra de mieux représenter les thèmes. En terme de précision nous remarquons que les thèmes Sports et Économie atteignent des valeurs importantes et ce pour $n = 5$ mots. Par contre, pour les deux autres thèmes, les valeurs de précision sont moyennes pour $n = 5$ mots, mais elles continuent à augmenter avec n.

3.8 Vérification de la fiabilité des méthodes utilisées dans le cas d'augmentation du nombre de thèmes

Dans les expériences précédentes, nous avions étudié les performances de deux méthodes de classification sachant que le nombre de thèmes étant égal à quatre. Néanmoins, dans cette présente expérience, notre objectif est de savoir si les performances ne se seraient pas dégradées si le nombre de thèmes à étudier augmente. Pour cela, nous présentons dans cette section, les performances des méthodes utilisées précédemment, mais cette fois-ci pour l'identification de six thèmes qui sont : culture, religion, économie, local, International et sports. Le corpus que nous avons utilisé est extrait d'un autre journal arabophone "Alwatan", et comprend 9000 articles. Nous avons réservé 8100 articles pour l'apprentissage et le reste pour le test. Les mots distincts obtenus lors de la suppression de ceux dont la fréquence d'apparition n'excédant pas la valeur 3, sont présentés dans la table (3.12).

Thème	Nombre de mots différents
Culture	77.154
Religion	37.862
Politique	48.723
Économie	40.162
Local	43.359
Sport	36.221
Total	283.481

TAB. 3.12 – Mots distincts après suppression des mots dont la fréquence <3

La taille du vocabulaire obtenu est 40000. La table (3.13) montre les performances de la méthode TFIDF dans l'identification des six thèmes susmentionnés.

Performances	Rappel (%)	Précision (%)	F_1(%)
Culture	83.50	74.50	78.74
Religion	94.33	95.50	94.91
Économie	85.60	92.75	89.03
Local	90.66	83.00	86.66
Politique	94.75	89.00	91.78
Sports	97.66	99.50	98.57
Moyenne	91.08	89.04	89.94

TAB. 3.13 – Performances de la méthode TFIDF dans l'identification de six thèmes

Les résultats diffèrent d'un thème à l'autre, ceci est dû à la différence entre les tailles des corpus thématiques, c'est-à-dire les articles constituant chacun des thèmes. En moyenne, nous avons obtenu 89.94 % en terme de la mesure F_1. Cela pourra être considéré comme étant un résultat significatif, particulièrement en le comparant à d'autres expériences menées pour la langue Française en utilisant la même méthode [10]. Pour ce qui concerne la méthode SVM, nous avons constaté une supériorité remarquable. En effet, sa performance en terme de mesure F_1 vaut 96.63 %. Une comparaison des résultats émanant des deux méthodes TFIDF et SVM est résumée dans la table (3.16).Le corpus d'apprentissage est composé de 8100 articles, en d'autres termes chaque thème est représenté en se basant sur 1350 articles.

Catégorisation bi-classes	Culture	Religion	Politique	Économie	Local	Sports
Culture	-	80	96	97.33	98	99.33
Religion	100	-	98.67	100	98.67	100
Politique	98	88.67	-	96	98.67	100
Économie	98.67	90.67	96.67	-	93.33	93.33
Local	98	90	98.67	92	-	96.67
Sports	100	93.33	99.33	100	100	-
Moyenne	98.93	88.53	97.86	97.06	97.73	99.06

TAB. 3.14 – Performances de la SVM en terme de rappel

Catégorisation bi-classes	Culture	Religion	Politique	Économie	Local	Sports
Culture	-	100	97.96	98.65	98	100
Religion	83.33	-	89.70	91.46	90.80	93.75
Politique	96.08	98.52	-	96.64	98.67	99.34
Économie	97.37	100	96.03	-	92.11	100
Local	98	98.54	98.67	93.24	-	100
Sports	99.34	100	100	99.34	96.77	-
Moyenne	94.82	99.41	96.47	95.86	95.27	98.62

TAB. 3.15 – Performances de la SVM en terme de précision

Performances	Rappel (%)	Précision (%)	Mesure F_1(%)
TFIDF	91.08	89.04	89.94
SVM	96.53	96.74	96.63

TAB. 3.16 – Comparaison des performances des deux méthodes TFIDF et SVM

3.9 Comparaison entre les corpus de test et uniformité de l'Arabe standard

Dans cette expérience, le corpus extrait du journal Alwatan est utilisé dans l'étape d'apprentissage. Pour l'évaluation nous avons pris trois corpus de trois journaux arabophones originaires de différents pays. Il s'agit du journal Algérien "Alkhabar", "Akhbar Alkhaleej" de Bahrain, et "Alarabonline". Les résultats exposés dans la table (3.17) sont presque similaires. De ce fait on peut tirer la conclusion que l'Arabe standard est une langue uniforme, du moins pour les pays où nous avons puisé les données de test textuelles. Les performances du classifieur TFIDF ne sont pas affectées sachant que nous avons utilisé un corpus d'apprentissage dont la source est le journal 'Alwatan', et d'autres corpus de test émanant d'autres sources. En effet, l'évaluation effectuée par l'utilisation de ces corpus de test varie de 87.64 % à

88.85 % en terme de la mesure F_1. Ce qui est très proche du résultat obtenu en utilisant les données de test issues du journal 'Alwatan' qui est de 89.94 %.

Thème	Journal	Rappel (%)	Précision (%)	F_1(%)
Culture	Akhbar Alkhaleej	83.00	78.50	80.68
	AlKhabar	81.00	75.66	78.23
	Alarabonline	81.50	75.00	78.11
Religion	Akhbar Alkhaleej	92.75	90.00	91.35
	AlKhabar	90.00	92.50	91.23
	Alarabonline	91.00	91.00	91.00
Économie	Akhbar Alkhaleej	86.75	91.50	89.06
	AlKhabar	84.50	88.50	86.45
	Alarabonline	85.33	90.33	87.75
Local	Akhbar Alkhaleej	88.50	83.00	85.66
	AlKhabar	88.00	81.50	84.62
	Alarabonline	90.50	79.66	84.73
Politique	Akhbar Alkhaleej	92.66	87.50	90.00
	AlKhabar	91.75	86.33	88.95
	Alarabonline	93..55	88.50	90.95
Sports	Akhbar Alkhaleej	96.66	95.75	96.20
	AlKhabar	95.75	96.33	96.03
	Alarabonline	97.00	97.33	97.16
Moyenne	Akhbar Alkhaleej	90.05	87.70	88.85
	AlKhabar	88.50	86.80	87.64
	Alarabonline	89.81	86.97	88.36

TAB. 3.17 – Évaluation par des corpus de test issus de trois journaux arabophones

Par cette expérience, il a été montré qu'il suffit de construire un corpus à partir d'une seule source, sans qu'il y ait de dégradation de performances en utilisant des documents de test de sources différentes.

3.10 La méthode M-SVM

La SVM a été initialement conçue pour réaliser la catégorisation bi-classes. Si un problème de classification de plus de deux catégories s'impose, la méthode one-versus-rest peut être appliquée. Il s'agit de séparer une classe de la mixture des autres classes. Toutefois cette approche s'avère inadéquate dans d'autres circonstances. C'est pour cela qu'une nouvelle stratégie apparaît qui considère toutes les classes en même temps. Dans [44] il a été montré que l'utilisation de la M-SVM dans la fusion des méthodes de prédiction de la structure secondaire des protéines améliore les résultats. Ceci est dû essentiellement à deux raisons. L'une d'elles est que ces méthodes reposent sur différents principes. L'autre est que les données sont extraites de différentes sources de connaissances. Dans le domaine de la classification de textes,

la M-SVM est plus performante que la SVM. Nous présentons dans la table (3.18) les performances de la méthode traditionnelle one-versus-rest, en prenant un vocabulaire de taille 8000. Les documents d'apprentissage ainsi que ceux de test sont constitués de 300 mots. Le corpus d'apprentissage dédié à cette expérience comprend 4200 articles. Dans notre cas où le nombre de thèmes à identifier est égal à six, nous avons entraîné six classifieurs de type one-versus-rest. Un document d est assigné au thème T_i qui donnera la valeur maximale de la fonction $g_i(d)$(avec i variant de 1 à 6). Sachant que $g_i(d)$ est la solution SVM issue de l'apprentissage du thème T_i vis-à-vis des autres thèmes. Cette méthode a donné des performances de 81.11 % en terme de rappel. Voir table (3.18). Nous mentionnons que des valeurs beaucoup plus importantes sont atteintes avec cette méthode, en utilisant des vocabulaires de taille plus grande. Voir les tables (3.14), (3.15) et (3.16).

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	80	92.31	85.71
Religion	53.33	100	69.56
Économie	80	92.31	85.71
Local	93.33	70	79.99
International	86.67	92.86	89.66
Sports	93.33	87.5	90.32
Moyenne	81.11	89.16	84.94

TAB. 3.18 – Performances de la méthode one-versus-rest en utilisant un vocabulaire de taille 8000 et des documents ne dépassant pas 300 mots.

La catégorisation multi-classes qui repose sur la méthode qui considère les classes en même temps, est réalisée en utilisant l'outil implémenté par Guermeur. En effet la première tâche que nous avons planifiée était d'utiliser cette méthode pour l'identification de trois thèmes qui sont : culture, religion, économie. Chaque document contient 80 mots, la taille du vocabulaire étant égale à 1000.

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	90	86	87.95
Religion	95	95	95.00
Économie	90	90	90.00
Moyenne	91.66	90.33	90.99

TAB. 3.19 – Performance de la méthode M-SVM en utilisant un vocabulaire de taille 1000 et un nombre de thèmes égal à trois

Comme on voit dans la table (3.19) les performances de la méthode M-SVM atteignent la valeur 91.66 % en terme de rappel, ce qui est remarquable en la comparant avec la méthode one-versus-rest achevée par la SVM, en utilisant les mêmes données. Voir table (3.20).

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	85	88	86.47
Religion	89.63	84	86.72
Économie	86.33	92	89.07
Moyenne	86.99	88	87.49

TAB. 3.20 – Performance de la méthode SVM (one-versus-rest) en utilisant un vocabulaire de taille 1000 et un nombre de thèmes égal à trois

3.11 Évaluation des performances de la M-SVM dans le cas d'augmentation du nombre de thèmes

En appliquant la M-SVM pour la première fois en identification de thèmes, notre premier souci était de vérifier à quel point cette méthode pourrait être efficace dans ce domaine. Nous nous sommes donc engagés à réaliser un test sur un nombre de thèmes égal à trois qui a abouti aux résultats exposés dans la table (3.19). Un taux de Rappel de 91.66 % et une Précision de 90.33 % est un bon signe que nous soyons sur le bon chemin. Cependant, pour pouvoir juger de la crédibilité de cette méthode, nous avons trouvé qu'il est plus judicieux de tester ces performances, en augmentant le nombre de thèmes traités. Il s'agit de ceux suscités dans les sections précédentes. La table (3.21) montre les performances de la méthode M-SVM en utilisant un corpus d'apprentissage constitué de 4800 articles. La taille de ces derniers ne dépasse pas 100 mots, tandis que celle du vocabulaire est de 1000.

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	66	60	62.85
Religion	92	96	93.95
Économie	80	74	76.88
Local	60	64	61.93
International	82	84	82.98
Sports	86	91	88.43
Moyenne	77.66	78.16	77.91

TAB. 3.21 – performances de la méthode M-SVM en utilisant un corpus d'apprentissage constitué de 4800 articles, un vocabulaire est de 1000 et des documents de taille ne dépassant pas 100 mots.

Nous avons continué cette expérience en variant les paramètres comme la taille du corpus d'apprentissage, des documents de test et celle du vocabulaire, ainsi que le paramètre noyau. Les performances se sont nettement améliorées pour arriver à un taux de rappel égal à : 90.45 %. Nous avons fait plusieurs essais de l'outil M-SVM qui nécessite un bon choix des différents paramètres comme la Capacité et le noyau. En diminuant la capacité la marge augmente et vice versa. Les résultats obtenus différent selon les paramètres choisis. En effet Voir table (3.22).

Expérience	TV	Noyau	TD	CA	Rappel (%)
1	600	1	100	1200	81.66
2	600	3	100	1200	82.5
3	600	1	100	4800	80.41
4	600	3	100	4800	84.79
5	2500	1	150	900	85.55
6	2500	3	150	900	84.44
7	2500	2	150	900	82.22
8	2500	1	250	900	84.44
9	2500	3	250	900	83.33
10	3000	1	100	3600	85.55
10	3000	1	300	4200	86.25
11	8000	1	300	3600	89.88
12	8000	1	300	4200	90.45

TAB. 3.22 – Performances de la M-SVM en variant différents paramètres

TV : Taille du vocabulaire.
TD : Taille des documents.
Noyau 1 : Noyau linéaire.
Noyau 2 : Noyau gaussien.
Noyau 3 : Noyau polynomial
CA : nombre de documents constituant le corpus d'apprentissage.

3.12 conclusion

A travers ce chapitre nous avons présenté un nombre d'expériences que nous avons menées afin d'étudier les performances des méthodes utilisées. En effet nous nous sommes intéressés en premier lieu au classifieur TFIDF, qui est largement répandu en catégorisation de textes. Puis nous avons expérimenté la SVM, outil efficace dans la séparation biclasses, qui a abouti à des résultats très satisfaisants. Néanmoins, le choix de faire une discrimination binaire n'est pas judicieux dans le cas de plus de deux thèmes "catégories". Nous avons ensuite sollicité la M-SVM qui permet de traiter l'ensemble de thèmes en même temps, à l'inverse de la SVM. La taille du vocabulaire joue un rôle primordial dans l'amélioration des performances. En effet, pour l'identification des six thèmes mentionnés précédemment, le classifieur TFIDF a mené à un taux de rappel égal à 91.08 %, et ce en utilisant un vocabulaire de taille 40000. La SVM a conduit à un taux de rappel égal à 96.50 % en utilisant la même taille. En comparant les performances de la SVM et ceux de la M-SVM, nous remarquons une supériorité de cette dernière. En effet, pour une taille de 8000 mots, beaucoup inférieure à celle utilisée précédemment, les taux de Rappel des deux méthodes SVM et M-SVM sont respectivement de 81.11 % et 90.45 %.

Chapitre 4

Proposition d'une nouvelle méthode en identification de thèmes : le TR-classifier

4.1 Introduction

Le langage naturel peut être vu comme un processus stochastique, du fait qu'on a une connaissance imparfaite des suites de mots qui peuvent être modélisés par un modèle de langage. Il peut aussi être considéré comme une source d'information. Pour pouvoir le traiter, une multitude de solutions a été proposée, parmi lesquelles se trouvent les modèles à historiques. Ce choix basé sur l'historique d'un document n'est pas arbitraire, car de potentielles sources d'informations y existent. Ces dernières peuvent être présentées par plusieurs modèles : nous citons les modèles à historiques courts qui englobent les n-grammes et les n-classes, et les modèles à historiques longs dont appartiennent les modèles triggers et le modèle Cache.

Nous proposons dans ce chapitre une nouvelle méthode d'identification de thèmes que nous avons baptisée TR-classifier (TRiggers-based classifier). L'idée est de quantifier le lien existant entre les mots dans le but de caractériser chacun des thèmes. Le mot "trigger" veut dire déclencheur ; en d'autres termes l'apparition d'un mot dans un texte donné pourrait déclencher un autre mot. En le prenant comme exemple, le mot "match" peut déclencher une multitude de mots "football" , "basketball", "arbitre", etc. L'information mutuelle moyenne nous permet de mesurer le degré de corrélation entre ces mots, ce qui nous permettra par la fin de représenter chaque thème par l'ensemble des triggers les caractérisant.

Nous présentons dans ce qui suit, quelques notions sur l'information mutuelle et l'information mutuelle moyenne, ainsi que quelques brèves définitions des modèles à historiques dont les triggers, avant d'aborder le TR-classifier et les expériences menées pour son évaluation.

4.2 Modèles à historiques

Le langage naturel a suscité de nombreuses recherches ayant comme but sa modélisation avec succès. Les modèles à historiques courts, comme leurs noms l'indiquent, ne prennent en considération que les derniers mots de l'historique. Un exemple de ces modèles est les modèles n-grammes. Un mot w_i peut être prédit par les $n-1$ mots les plus proches, comme c'est décrit par l'équation (4.1).

$$P(w_i \mid h) = P(w_i \mid w_{i-n+1}, \cdots, w_{i-1}) \tag{4.1}$$

En pratique la valeur de n peut aller jusqu'à cinq, toutefois les valeurs les plus utilisées sont $n = 2$ (modèles bigrammes) et $n = 3$ (modèles trigammes) [10]. Outre leur simple mise en oeuvre, ces modèles ne nécessitent aucun effort de calcul supplémentaire en cours de reconnaissance [32]. L'utilisation de grandes valeurs de n causent toutefois un problème de la rareté des historiques correspondants du corpus d'apprentissage, ce qui conduit à une faible validité statistique des fréquences calculées [32].

Les modèles à historiques longs utilisent un historique qui peut arriver jusqu'à plusieurs milliers de mots [32]. Néanmoins, au contraire des modèles à historiques courts, seulement une information extraite de l'historique qui est utilisée. Le modèle Cache (défini dans la section 1.2.2) et les modèles Triggers en sont des exemples.

4.3 Information mutuelle moyenne

L'information mutuelle de deux variables aléatoires X et Y permet de mesurer leur dépendance statistique. Si nous considérons la densité de probabilité jointe $P(X,Y)$ aux deux variables X et Y, et les distributions marginales $P(X)$et $P(Y)$ alors la quantité d'information apprise sur X en observant les valeurs de Y est donnée par la relation (4.2).

$$I(X,Y) = P(X,Y)\, log_2 \frac{P(X,Y)}{P(X).P(Y)} \tag{4.2}$$

L'information mutuelle est nulle s'il y a une indépendance entre les deux variables, et elle croit lorsque la dépendance entre elles augmente. Dans le cas du traitement automatique de la langue, ce sont généralement les mots qui sont considérés comme étant des événements. La cooccurrence de deux mots n'est pas due au hasard, l'information mutuelle permet donc de recenser les couples de mots ayant une forte relation entre eux

L'information mutuelle moyenne est donnée par la relation (4.3) :

$$I(X,Y) = \sum_{moy} P(X,Y)\, log_2 \frac{P(X,Y)}{P(X).P(Y)} \tag{4.3}$$

4.4 Le TR-classifier

Soit w_k un mot appartenant à un document, V_i un vocabulaire d'un thème T_i, $Trig(w_k)$ c'est l'ensemble de mots déclenchés par le mot w_k.

L'utilisation des triggers dans l'identification de thèmes consiste à :

1. Donner pour chaque mot w_k du vocabulaire V_i les mots déclenchés (triggers).

2. Considérer les M meilleurs mots déclenchés, qui caractériseront le thème T_i.

3. En phase de test, nous calculons pour chaque mot w_k du document de test

$d = \{w_1, w_2, ..., w_k\}$, les triggers qui s'y trouvent.

4. Calculer les valeurs Q_i en utilisant la distance gauche-droite (4.4) :

$$Q_i = \frac{\sum_{i,k} IMM(w_k, w_k^i)}{\sum_{l=0}^{n-1}(n-l)} \tag{4.4}$$

i variant de 1 à N, où N est le nombre de thèmes. avec w_k^i les triggers contenus dans le document d et caractérisant le thème T_i

5. L'appartenance du document d dans un thème T_i est obtenue en choisissant la valeur max des Q_i .

L'idée de base est de calculer pour chaque couple de mots appartenant au vocabulaire de thème V_i l'information mutuelle moyenne. Les triggers sont obtenus en sélectionnant les couples de mots qui ont des valeurs d'information mutuelle moyenne importantes. Pour cela nous considérons les documents constituant le thème T_i (corpus de T_i). Chaque thème sera muni d'un ensemble de triggers dont nous choisirons les M meilleurs.

L'information mutuelle moyenne correspondant à un couple de mots a et b est donnée par la relation (4.5) :

$$\begin{aligned} IMM(a,b) = p(a,b).log\frac{p(a,b)}{p(a).p(b)} + p(\bar{a},b).log\frac{p(\bar{a},b)}{p(\bar{a}).p(b)} + \\ p(a,\bar{b}).log\frac{p(a,\bar{b})}{p(a).p(\bar{b})} + p(\bar{a},\bar{b}).log\frac{p(\bar{a},\bar{b})}{p(\bar{a}).p(\bar{b})} \end{aligned} \tag{4.5}$$

$p(a,b)$: Nombre de documents dans lesquels se trouvent a et b en même temps.

$p(a,\bar{b})$: Nombre de documents dans lesquels se trouve a en absence de b .

$p(\bar{a},\bar{b})$: Nombre de documents ou a et b sont tous les deux absents.

$p(\bar{a})$: Nombre de documents qui ne contiennent pas le mot a .

$p(a)$: Nombre de documents qui contiennent le mot a .

Identifier un thème parmi d'autres, en utilisant le TR-clasifier, revient donc à suivre les étapes suivantes :

- Prendre un mot w_k appartenant au document de test d.

- Chercher dans le vocabulaire V_i si le mot w_k y appartient.

- Si c'est le cas, comptabiliser le nombre de triggers du mot w_k (obtenus dans le document d) qui se trouvent parmi les triggers du thème T_i concernant le même mot w_k . Si ce n'est pas le cas alors attribuer au mot w_k la valeur nulle en ce qui concerne le thème T_i.

- Les étapes précédentes seront appliquées pour tous les mots du document d.

- Dans notre expérience le nombre de thèmes est 6. Par conséquent, la valeur maximale de Q_i est prise en considération, avec i variant de 1 à 6. Si la valeur maximale est celle de Q_3 alors le document d traite le thème T_3.

4.5 Expériences et résultats

4.5.1 Description

Cette méthode consiste à utiliser les triggers pour l'identification de thèmes. La première étape que nous avons suivie dans cette voie est de construire pour chacun des six thèmes un ensemble de mots représentatifs (vocabulaires de thèmes). L'idée est de choisir les mots les plus fréquents. Nous avons ensuite attribué à chacun de ces mots les triggers correspondants. Nous avons établi des expériences pour étudier les performances de cette méthode. Nous avons remarqué que l'augmentation du nombre des triggers améliore le taux d'identification. Dans la première expérience, nous avons pris des tailles différentes des vocabulaires de thème (voir table (4.1)). Nous avons utilisé un nombre de triggers égal à 10 puis 20. Les taux de rappel correspondants étaient respectivement de 59.22 % et 67.44 %. Puis nous avons augmenté légèrement les tailles des vocabulaires, en espérant avoir des résultats meilleurs, ce qui n'a pas été le cas, en l'occurrence le taux de rappel a été de 48.88%. Nous avons mené d'autres expériences en utilisant des tailles identiques des vocabulaires de thèmes, tout en augmentant le nombre de triggers. L'utilisation d'une taille égale à 100 a donné un taux de rappel égal à 83.11%. Tandis que les tailles 200 et 300 ont mené respectivement à des taux de rappel égaux à 84% et 89.67%. Par ailleurs, nous avons remarqué qu'en attribuant des tailles différentes de vocabulaire à chaque thème, nous n'avons pas atteint les performances espérées. Par contre le choix des tailles identiques de vocabulaires de thèmes améliore les performances.

4.5.2 Vocabulaires thématiques de tailles différentes

Nous avons attribué à chacun des thèmes un vocabulaire de taille différente des autres. Les figures (4.2) et (4.3) montrent les Performances du TR-classifier en utilisant respectivement un nombre de triggers 10 et 20.

Thème	Taille du vocabulaire
Culture	192
Religion	309
Économie	250
Local	213
International	181
Sports	136
Total	1281

TAB. 4.1 – Vocabulaires de thèmes avec différentes tailles

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	54	52.94	53.46
Religion	64.66	81.51	72.11
Économie	37.33	49.56	42.58
Local	56.66	34.98	43.25
International	60	37.38	46.06
Sports	82.66	37.37	51.47
Moyenne	59.22	54.95	57.00

TAB. 4.2 – Performances du TR-classifier en utilisant un nombre de triggers 10

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	60.66	57.96	59.27
Religion	72.66	83.84	77.85
Économie	52.66	63.71	57.66
Local	57.33	43.87	49.70
International	74	85.38	79.28
Sports	87.33	80.37	83.70
Moyenne	67.44	69.18	67.91

TAB. 4.3 – Performances du TR-classifier en utilisant un nombre de triggers 20

Dans cette première expérience, nous avons alloué des tailles de vocabulaires différentes. Le nombre de triggers N utilisé est un paramètre important pour avoir de bonnes performances. En effet, pour $N = 10$ le taux de Rappel obtenu est égal à 59.22%. Cette valeur est améliorée à 67.44% pour $N = 20$. Toutefois, nous allons voir dans les expériences suivantes, que l'utilisation des tailles identiques des vocabulaires de thème, pour cette même valeur du nombre de triggers donne un taux de Rappel égal à 71.55%.

4.5.3 Vocabulaires thématiques de tailles identiques

Dans cette expérience nous allons tester la méthode en faisant varier la taille des vocabulaires de thèmes ainsi que le nombre de triggers. Nous avons présenté les performances de notre méthode pour un vocabulaire de taille 100 dans les tables (4.4), (4.5), (4.6) et (4.7). Puis d'autres tailles de vocabulaires ont été testées, en l'occurrence 200 et 300.

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	54	67.5	60
Religion	87.33	82.39	84.78
Économie	60.66	68.94	64.53
Local	52.66	65.29	58.30
International	81.33	76.73	78.96
Sports	93.33	66.98	77.99
Moyenne	71.55	71.30	71.42

TAB. 4.4 – Performances pour un nombre de triggers égal à 20 et une taille de vocabulaire de thème égale à 100

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	59.33	85.57	70.07
Religion	96	82.76	88.89
Économie	76	75.49	75.74
Local	66	71.74	68.75
International	84.66	77.44	80.89
Sports	94.66	84.02	89.02
Moyenne	79.44	79.50	79.47

TAB. 4.5 – Performances pour un nombre de triggers égal à 40 et une taille de vocabulaire de thème égale à 100

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	66.66	82.64	73.79
Religion	96.66	82.85	89.22
Économie	74.66	78.87	76.71
Local	72	72	72
International	91.33	84.56	87.81
Sports	92.66	92.66	92.66
Moyenne	82.33	82.26	82.29

TAB. 4.6 – Performances pour un nombre de triggers égal à 60 et une taille de vocabulaire de thème égale à 100

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	66	88.39	75.57
Religion	97.33	81.11	88.48
Économie	76.66	78.76	77.69
Local	75.33	70.62	72.90
International	91.33	85.09	88.10
Sports	92	97.98	94.89
Moyenne	83.11	83.64	83.37

TAB. 4.7 – Performances pour un nombre de triggers égal à 80 et une taille de vocabulaire de thème égale à 100

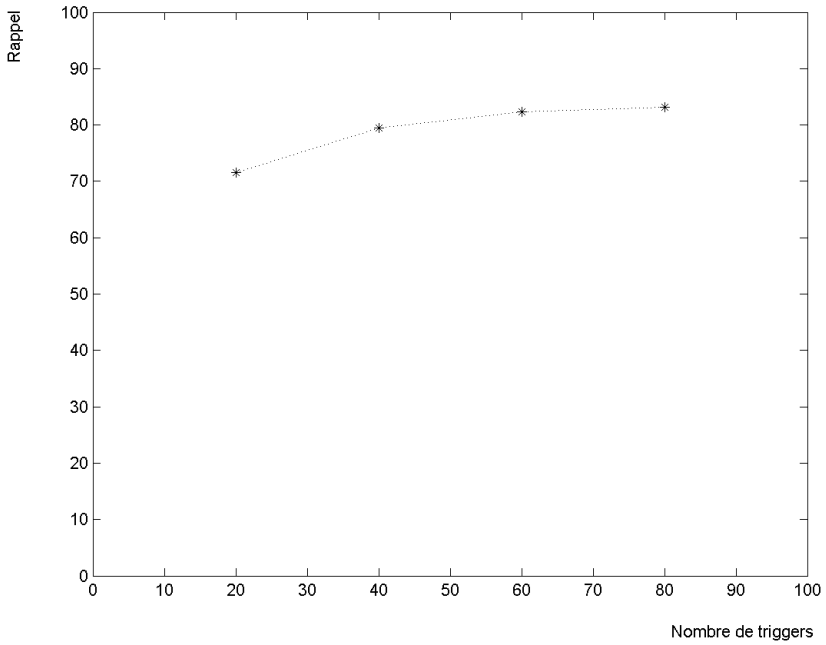


FIG. 4.1 – Taux de Rappel en fonction du nombre de triggers pour une taille de vocabulaire égale à 100

Le choix d'un nombre de triggers égal à 20, avec une taille des vocabulaires de thème de valeur 100, a permis d'obtenir un taux de Rappel global égal à 71.55 %. Le fait que les thèmes Local, Culture et Économie aient des valeurs du Rappel non satisfaisantes est dû à la variété qui caractérise ces thèmes. En effet, leur division en sous-thèmes s'avère utile, voire nécessaire, si l'on veut aboutir à de bonnes performances. D'autre part les trois autres thèmes sont relativement facilement identifiés, particulièrement le thème Sports qui a abouti à un taux de 93.33 %.

Notre souci étant de pousser les performances de cette méthode au maximum, nous avons effectué d'autres essais en augmentant le nombre de triggers.

En prenant la valeur 40, le taux de Rappel global passe à 79.44%, soit une amélioration d'environ 8 %. Les nombres de triggers 60 et 80 ont permis d'obtenir les performances perspectives 82.33 % et 83.11 %. Notons que la différence entre ces deux dernières est moins de 1 %.

Ces résultats sont en dessous de nos aspirations, nous avons décidé donc d'utiliser d'autres tailles de vocabulaires de thème, en l'occurrence 200 et 300.

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	56	64.62	60.00
Religion	60.66	74.6	66.91
Économie	58	60.42	59.18
Local	54	44.75	48.94
International	74	76.55	75.25
Sports	86.66	73.03	79.26
Moyenne	64.88	65.66	65.26

TAB. 4.8 – Performances pour un nombre de triggers égal à 20 et une taille de vocabulaire de thème égale à 200

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	61.33	66.66	63.88
Religion	71.33	79.85	75.35
Économie	66	72.26	68.99
Local	66.66	55.55	60.60
International	80.66	81.75	81.20
Sports	94	86.50	90.10
Moyenne	73.33	73.76	73.54

TAB. 4.9 – Performances pour un nombre de triggers égal à 40 et une taille de vocabulaire de thème égale à 200

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	66	69.72	67.81
Religion	75.33	83.70	79.29
Économie	70.66	73.61	72.10
Local	69.33	61.90	65.40
International	84	82.90	83.44
Sports	93.33	88.05	90.61
Moyenne	76.44	76.64	76.54

TAB. 4.10 – Performances pour un nombre de triggers égal à 60 et une taille de vocabulaire de thème égale à 200

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	68	69.86	68.92
Religion	78.66	83.69	81.09
Économie	71.33	75.35	73.28
Local	68.66	63.97	66.23
International	87.33	85.62	86.46
Sports	93.33	89.17	91.20
Moyenne	77.88	77.94	77.91

TAB. 4.11 – Performances pour un nombre de triggers égal à 80 et une taille de vocabulaire de thème égale à 200

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	65.33	71.53	68.29
Religion	80.66	82.87	81.75
Économie	73.33	74.83	74.07
Local	70	65.62	67.73
International	88.66	85.26	86.93
Sports	94.66	92.21	93.42
Moyenne	78.77	78.72	78.74

TAB. 4.12 – Performances pour un nombre de triggers égal à 100 et une taille de vocabulaire de thème égale à 200

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	64	75	69.06
Religion	84	83.44	83.72
Économie	76	76.51	76.25
Local	72	68.79	70.36
International	91.33	83.54	87.26
Sports	94	93.38	93.69
Moyenne	80.22	80.11	80.16

TAB. 4.13 – Performances pour un nombre de triggers égal à 120 et une taille de vocabulaire de thème égale à 200

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	68	75	71.33
Religion	85.33	83.12	84.21
Économie	76	81.43	78.62
Local	76.66	69.69	73.01
International	90	85.99	87.95
Sports	94.66	95.95	95.30
Moyenne	81.77	81.86	81.81

TAB. 4.14 – Performances pour un nombre de triggers égal à 140 et une taille de vocabulaire de thème égale à 200

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	70.66	84.12	76.80
Religion	92	82.14	86.79
Économie	76	83.21	79.44
Local	80.66	71.18	75.62
International	92.66	87.97	90.25
Sports	92	97.87	94.84
Moyenne	84	84.41	84.20

TAB. 4.15 – Performances pour un nombre de triggers égal à 160 et une taille de vocabulaire de thème égale à 200

L'augmentation de la taille du vocabulaire en prenant la valeur 200 a conduit à une amélioration légère du taux de Rappel d'environ 1%. Néanmoins ceci a nécessité l'utilisation d'un nombre de triggers égal à 160 pour arriver à un taux de Rappel de 84%. Sachant que dans l'expérience où nous avons utilisé une taille de vocabulaire égale à 100, le taux de Rappel était de 83.11% avec un nombre de triggers égal à 80. La figure (4.2) montre une évolution lente des valeurs du taux de Rappel en fonction du nombre de triggers à partir de la valeur 40. Le choix d'une taille de vocabulaire égale à 200 n'est pas efficace en comparant les résultats à ceux de l'expérience précédente -voir figure (4.1)-. Pour cela nous avons continué à augmenter la taille du vocabulaire, en faisant varier le nombre de triggers, -figure (4.3)-.

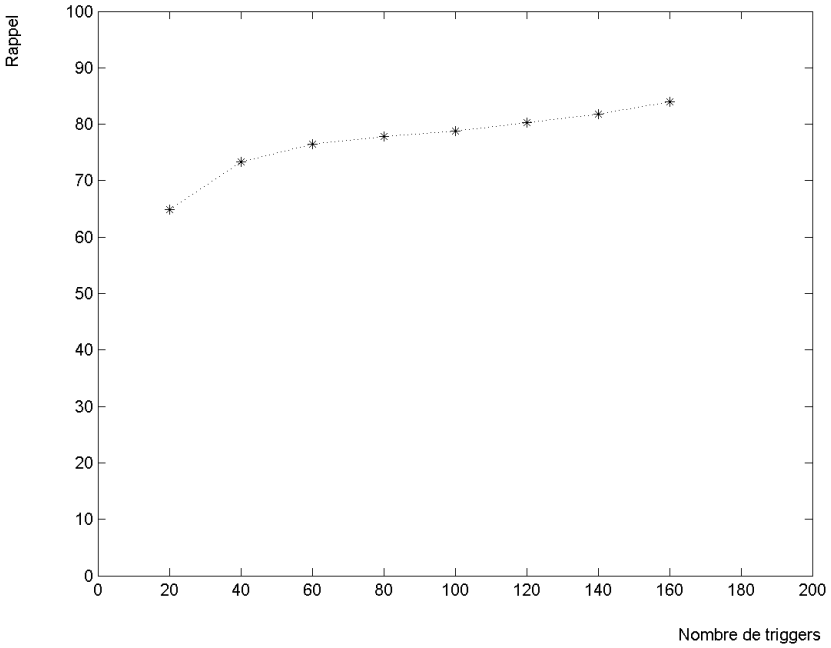


FIG. 4.2 – Taux de Rappel en fonction du nombre de triggers pour une taille de vocabulaire égale à 200

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	64	51.89	57.31
Religion	58	85.29	69.04
Économie	54	54	54
Local	46	37.91	41.56
International	49.33	52.11	50.68
Sports	58	62.59	60.21
Moyenne	54.88	57.30	56.06

TAB. 4.16 – Performances pour un nombre de triggers égal à 20 et une taille de vocabulaire de thème égale à 300

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	69.33	73.24	71.23
Religion	83.33	83.89	83.61
Économie	73.33	72.37	72.84
Local	72	68.35	70.13
International	81.33	82.99	82.15
Sports	90	88.81	89.40
Moyenne	78.22	78.27	78.24

TAB. 4.17 – Performances pour un nombre de triggers égal à 100 et une taille de vocabulaire de thème égale à 300

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	73.33	73.82	73.57
Religion	83.33	83.89	83.61
Économie	76	77.55	76.76
Local	75.33	71.97	73.61
International	88	88	88
Sports	91.33	92.57	91.94
Moyenne	81.22	81.30	81.26

TAB. 4.18 – Performances pour un nombre de triggers égal à 160 et une taille de vocabulaire de thème égale à 300

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	77	77.08	77.034
Religion	87.50	84.21	85.82
Économie	82	81.25	81.62
Local	84	72.44	77.79
International	92	89.03	90.49
Sports	93.33	93.96	93.64
Moyenne	85.97	82.99	84.45

TAB. 4.19 – Performances pour un nombre de triggers égal à 200 et une taille de vocabulaire de thème égale à 300

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	82.66	80.55	81.59
Religion	96.33	83.56	89.49
Économie	83.50	84.05	83.77
Local	86.25	82.53	84.35
International	93.33	90.66	91.97
Sports	96	97.33	96.66
Moyenne	89.67	86.44	88.02

TAB. 4.20 – Performances pour un nombre de triggers égal à 250 et une taille de vocabulaire de thème égale à 300

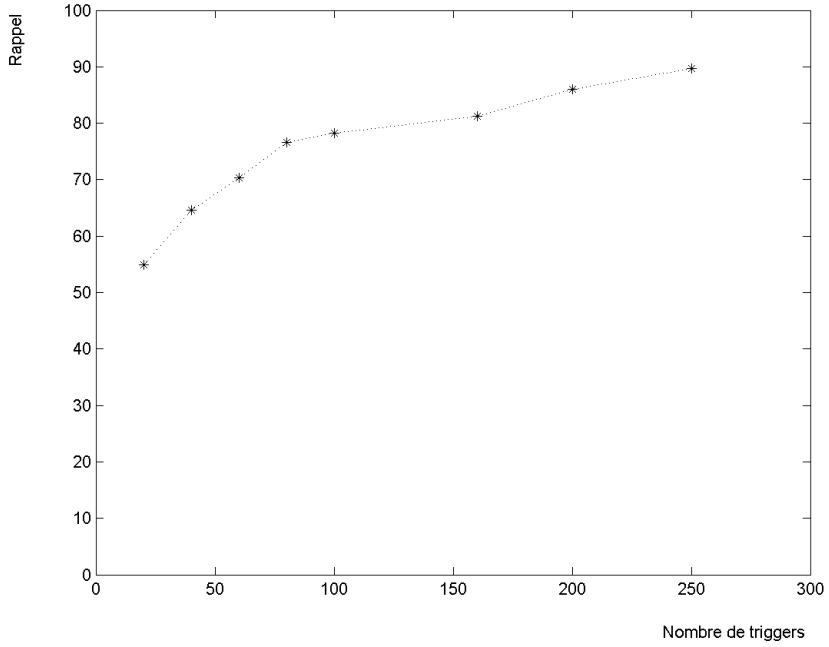


FIG. 4.3 – Taux de Rappel en fonction du nombre de triggers pour une taille de vocabulaire égale à 300

Nous présentons dans la figure (4.3) les performances du TR-classifier pour une taille de vocabulaire égale à 300. Contrairement à la courbe présentée dans la figure (4.2), les valeurs du taux de Rappel évoluent mieux. En effet, pour un nombre de triggers égal à 250 les performances du TR-classifier en terme de Rappel sont exprimées par la valeur 89.69%.

L'augmentation de la taille du vocabulaire doit être accompagnée par l'utilisation d'un nombre de triggers supérieur pour assurer l'amélioration des performances. En effet,nous remarquons qu'à chaque fois nous augmentons la taille du vocabulaire,

les performances diminuent dans le cas où nous nous contentons d'utiliser le même nombre de triggers.

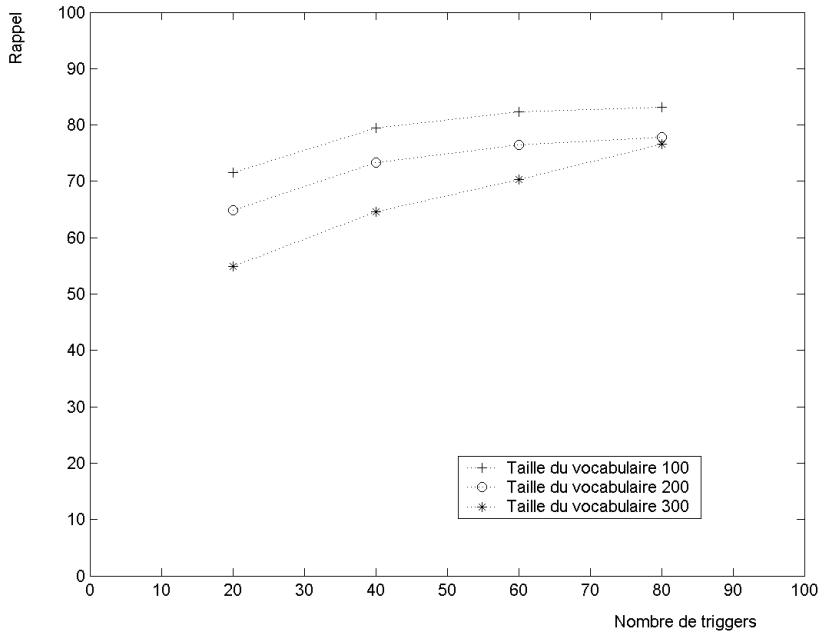


FIG. 4.4 – Les performances du TR-classifier pour les trois tailles du vocabulaire

Dans la figure (4.4), nous avons tracé les trois courbes représentatives des performances du TR-classifier en fonction du nombre de triggers variant de 20 à 80, en utilisant trois tailles différentes du vocabulaire. Les meilleures performances obtenues sont celles réalisées par l'utilisation d'un vocabulaire d'une taille égale à 100, comme le montre la figure (4.4).

La taille 200 a donné des taux de Rappel inférieurs. Nous remarquons que la différence entre les performances de ces deux tailles est constante pour l'ensemble des valeurs du nombre de triggers allant de 20 à 80. En augmentant la taille du vocabulaire à 300, les performances continuent à se dégrader de la même manière. Toutefois les valeurs du taux de Rappel commencent à se rapprocher de celles obtenues en utilisant la taille 200 à partir du nombre de triggers 60. Les performances maximales ont été atteintes en utilisant un nombre de triggers maximal pour une taille de vocabulaire 300.

Les résultats obtenus lors de l'utilisation de trois tailles différentes du vocabulaire indiquent les points suivants :

– L'augmentation de la taille du vocabulaire améliore la représentativité, mais ceci n'est pas suffisant pour avoir de bons résultats, il est donc nécessaire de choisir le nombre de triggers approprié pour que les performances augmentent. En effet, en prenant un nombre de triggers égal à 20, nous remarquons une

dégradation des performances à chaque fois que nous augmentons la taille du vocabulaire. Autrement dit, dans le cas d'une taille de vocabulaire réduite, cela veut dire que le nombre de triggers utilisé est plus significatif.

– En augmentant la taille du vocabulaire, un nombre de triggers supérieur est nécessaire pour dépasser les performances issues d'un vocabulaire de taille moins.

– L'augmentation conjointe de la taille du vocabulaire et le nombre de triggers permettent d'obtenir les meilleures performances.

En tenant compte des tailles très petites du vocabulaire que nous avons utilisées, nous pouvons dire que le TR-classifier a donné des résultats satisfaisants.

4.6 Comparaison entre TR, TFIDF, SVM et M-SVM

Le but de cette expérience est de comparer le TR-classifier aux autres méthodes testées jusqu'à maintenant. Cependant, le TR-classifier utilise un vocabulaire par thème V_i, $(i = 1, 2, \ldots, 6)$, ce qui n'est pas le cas pour les autres méthodes. Pour cela nous avons construit un vocabulaire général par la concaténation des V_i. Nous avons gardé les mêmes données d'apprentissage et de test. Les vocabulaires thématiques sont construits à partir des mots ayant des fréquences maximales. Ainsi la taille du vocabulaire général obtenu est de 800 mots.

Les résultats trouvés par la SVM ne se sont pas dégradés en utilisant un vocabulaire général de taille 800 mots. Ceci revient à la capacité de cet outil de fournir de bons résultats de classification et à la façon dont nous avons construit le vocabulaire (prise en considération du rangement des mots selon leurs fréquences du maximum au minimum). Ainsi, le vocabulaire est formé des mots ayant des fréquences maximales, ce qui veut dire qu'il offre plus de chance pour une meilleure représentation, et ceci en permettant d'éviter au maximum le chevauchement entre les mots appartenant aux différents vocabulaires.

La SVM, n'est pas adéquate dans le cas où le nombre de classes est supérieur à deux. Nous avons vu, certainement, que les résultats obtenus en utilisant cette méthode sont satisfaisants, néanmoins son point faible est qu'elle réalise uniquement la classification binaire. Une version récente généralisée de la SVM qui est en mesure de traiter plusieurs classes à la fois a été implémentée au cours de ces proches dernières années, en l'occurrence la M-SVM. Pour le TR-classifier, un taux de rappel égal à 89.67 % est encourageant, en sachant que la taille de chaque vocabulaire de thème dans les expériences que nous avons réalisées ne dépasse pas 300 mots. Les résultats émanant des quatre méthodes sont exposées dans les tables(4.21, 4.22, 4.23, 4.25 et 4.24).

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	71.33	88.43	78.96
Religion	93.33	86.95	90.03
Économie	83.33	80.64	81.96
Local	80	76.92	78.43
International	93.33	84.33	88.60
Sports	94	100	96.91
Moyenne	85.88	86.21	86.04

TAB. 4.21 – Performances du classifieur TFIDF

Catégorisation bi-classes	Culture	Religion	Économie	Local	Politique	Sports
Culture	-	86	96.67	97.33	96.67	99.33
Religion	99.33	-	100	98	99.33	100
Économie	98.67	99.33	-	90	96.67	98.67
Local	96	100	89.33	-	99.33	98
Politique	93.33	99.33	95.33	96	-	100
Sports	99.33	100	100	99.33	99.33	-
Moyenne (/5)	97.33	96.93	96.26	96.13	98.26	99.20

TAB. 4.22 – Performances de la méthode SVM en terme de rappel

Catégorisation bi-classes	Culture	Religion	Économie	Local	Politique	Sports
Culture	-	99.23	98.64	96.05	93.55	99.33
Religion	87.65	-	99.5999.34	100	99.33	100
Économie	96.73	100	-	89.40	95.39	100
Local	97.30	98.04	89.93	-	96.13	99.32
Politique	96.55	99.33	96.62	99.31	-	99.34
Sports	99.33	100	98.68	98.03	100	-
Moyenne (/5)	95.51	99.32	96.57	96.55	96.88	99.59

TAB. 4.23 – Performances de la méthode SVM en terme de précision

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	82.66	80.55	81.60
Religion	96.33	83.56	89.50
Économie	83.50	84.05	83.77
Local	86.25	82.53	84.35
International	93.33	90.66	91.97
Sports	96	97.33	96.66
Moyenne	89.67	86.44	88.02

TAB. 4.24 – Performances du TR-classifier

Thème	Rappel (%)	Précision (%)	F_1(%)
Culture	75	78	76.47
Religion	95	96	95.50
Économie	83.5	75	79.02
Local	74	64	68.64
International	86.75	83	84.83
Sports	90	89.5	89.75
Moyenne	84.04	80.91	82.44

TAB. 4.25 – Performances de la méthode M-SVM

4.7 Supériorité du TR-classifier sur TFIDF classifier

Les documents extraits du corpus utilisé sont déjà classés, toutefois des erreurs humaines commises conduisent à attribuer de faux thèmes à un nombre de documents. c'est ce qui a été constaté dans [10] où il a été montré que la fiabilité humaine dans l'étiquetage (classification manuelle), joue un rôle important dans l'amélioration des performances des méthodes d'identification.

L'expérience suivante a pour but de montrer l'effet de la fiabilité/non fiabilité humaine sur les performances obtenues, et de tester la capacité des deux méthodes TR et TFIDF à s'adapter à de telles situations. Cette expérience consiste à revoir les articles de test mal identifiés par les deux méthodes TR et TFIDF, et s'assurer si ces étiquettes attribuées par ces dernières est correcte. Le taux de Rappel est recalculé selon les nouveaux résultats. La tâche de relecture et d'identification manuelle a été assignée à un nombre d'experts ayant la compétence et la capacité de trancher sur l'appartenance d'un texte donné à un thème ou un autre.

Pour mieux comprendre, nous présentons l'exemple de la figure (4.5) qui représente un texte dont le thème attribué au préalable est "culture", tandis que le vrai thème est "Nouvelles internationales". Les deux méthodes l'ont identifié comme étant un article traitant des "nouvelles internationales", ce qui est correct.


مقدمة يحاول هذا البحث دور حلف شمال الأطلسي اثبات فرضيتين اساسيتين أولاهما حتمية الدور الطليعي للناتو في بناء الأمن والاستقرار عبر جانبي الاطلسي وعدم وجود اي مؤسسة غربية تستطيع منافسته في ذلك وثانيتهما ان نجاح الحلف في مهمته الأمنية الأطلسية سيكون مقدمة لتحوله الى قوة سياسية وعسكرية كبرى متحكمة في مجريات السياسة الدولية وفارضة شروطها وآلياتها عليها الفصل الاول اطار نظري في الدور الجديد لحلف شمال الاطلسي هذا الاطار يشكل قاعدة انطلاق فكرية لما يسعى الباحث د نزار اسماعيل الحيالي في الفصول اللاحقة وقد تمثل هذا الاطار في التأكيد النظري على وحدانية النظام الدولي الراهن وأزليته وان الناتو سيكون له دور متميز في جميع شروطه وضوابطه وآليات عمله المعروفة كالهيمنة وتوازن القوى والمصالح والتنافس والتكامل الاقتصادي يقول الباحث وعليه مادام النظام الدولي يمثل ظاهرة تاريخية ممتدة ومستمرة ومادام محتفظا بمؤسساته التقليدية كالدولة والاحلاف والمنظمات الدولية والاقليمية فان دراسة دور حلف شمال الاطلسي بعد انتهاء الحرب الباردة ينبغي ان تجرى وفق السياقات او الاليات التي عمل بموجبها هذا النظام مثل الهيمنة والقطبية وتوازن القوى وتوازن المصالح والتنافس والتوافق الاقتصادي الفصل الثاني حلف شمال الأطلسي وأوروبا حاول المؤلف في هذا الفصل اثبات الفرضية الاولى وهي حتمية الدور الطليعي للناتو في بناء الأمن والأستقرار لذلك تضمن هذا الفصل ثلاثة مباحث تبين مدى الحاجة الملحة الى وجود الناتو كضمان الامن عبر جانبي الاطلسي حيث تناول في المبحث الاول النشأة والتطور التاريخي للحلف وفي المبحث الثاني علاقة الحلف بالبيئة الاوروبية الجديدة المتولدة من انتهاء الحرب الباردة وفي المبحث الثالث نظرة القوة الكبرى ومواقفها منه وقد وضعت المعاهدة المنشئة للحلف ثلاث وظائف أساسية ينبغي على الحلف القيام بها وهي أولا الوظيفة العسكرية وثانيا الوظيفة السياسية وثالثا الوظيفة الاقتصادية وخلاصة القول حسب رأي الكاتب ان مواقف الدول الغربية من عملية توسيع الحلف تعكس مدى تأثرها بمصالحها القومية والاستعمارية وان هذه التأثرات انعكست بدورها على صياغة الاستراتيجية الجديدة التي جاءت جامعة .

FIG. 4.5 – Texte traitant le thème "Nouvelles internationales"

Nous avons réalisé notre expérience en prenant un seul thème : "culture". Les taux de Rappel obtenus en utilisant les classifieurs TR et TFIDF sont respectivement 92.66% et 78.66% . Nous remarquons une amélioration importante du taux de Rappel pour le TR-classifier de 10%, en la comparant avec la table (4.24). Tandis que l'amélioration introduite par la méthode TFIDF a été moins importante, en l'occurrence 7.33% en terme de Rappel, voir table (4.21).

Cette expérience a permis non seulement d'améliorer les performances mais de montrer l'aptitude des méthodes testées à attribuer des vraies étiquettes, malgré le faux étiquetage manuel fait au préalable. Notons que les performances du TR-classifier ont été plus importantes.

4.8 conclusion

Nous avons présenté dans ce chapitre une nouvelle méthode de détection de thème. Son originalité réside dans l'exploitation des triggers pour représenter fidèlement chacun des thèmes. Contrairement aux méthodes que nous avons étudiées, nous avons cette fois-ci attribué un vocabulaire par thème.

L'avantage de cette méthode est le fait d'utiliser des vocabulaires de tailles très réduites. En effet, les expériences que nous avons réalisées ont conduit à des performances satisfaisantes. Le nombre de triggers lui aussi est un paramètre important dans l'amélioration des résultats. Pour une taille donnée du vocabulaire de thème,

nous avons remarqué une amélioration des performances de la méthode en augmentant le nombre de triggers.

En comparant le TR-classifier avec les autres méthodes que nous avons étudiées, nous remarquons qu'il est le plus performant après la SVM. Outre cette comparaison, nous avons montré que le TR-classifier est plus efficace que la TFIDF en ce qui concerne la vérification des documents mal étiquetés par des individus. En effet, l'amélioration introduite par le TR-classifier et la TFIDF sont exprimées respectivement par les valeurs de Rappel 10 % et 7.33 %.

Toutefois, nous pouvons dire que nous n'avons pas exploité au maximum les performances de ce classifieur, du moment où nous n'avons pas utilisé des tailles de vocabulaires supérieures à celles déjà exploitées.

Conclusion générale et perspectives

Comme nous l'avons mentionné dans l'introduction de ce manuscrit, le but de notre travail est d'effectuer une étude sur l'identification de thème appliquée à la langue Arabe. Ceci est réalisé dans le but d'améliorer les performances des systèmes de reconnaissance automatique de la parole, en adaptant les modèles de langage au texte.

Au début nous avons présenté les différentes méthodes de détection de thème existantes comme la TFIDF qui est une référence dans ce domaine, la SVM, le modèle Cache, la méthode WSIM, le modèle unigramme thématique, les réseaux de neurones, etc. La plupart de ces méthodes a fait l'objet de plusieurs études comparatives qui ont montré les capacités des unes et les inconvénients des autres.

Pour la langue Arabe, les travaux non nombreux concernant la catégorisation de textes ont pu fournir des résultats presque comparables à ceux obtenus dans [28]. En effet, dans [21], l'algorithme Naïve Bayes a conduit à des performances égales à 71.96% en terme de Rappel, alors que la même méthode que Yang a utilisée a fourni un taux de 79.56% . Le système "ArabCat", basé sur le Maximum d'entropie, reporté dans [24] a mené à une valeur de la mesure F_1 égale à 80.41%.

Le mode de représentation que nous avons adopté est le mode Bag of Words, qui est sans doute très approprié à la détection de thème. Toutefois, l'introduction des bigrammes ou même des trigrammes dans la liste des unigrammes, aurait amélioré les résultats davantage. Citons par exemple le cas du bigramme "Club sportif" et le trigramme "Chef de l'état" qui serait intéressant de les traiter tel qu'ils sont, au lieu de considérer indépendamment les termes.

Les méthodes de sélection des mots-clés ou "mots du vocabulaire", ayant déjà fait l'objet d'études notamment dans [10], ne mènent pas à de différences palpables en terme de performance, ce qui nous a conduit à choisir une méthode qui est simple mais aussi performante, en l'occurrence la fréquence de mots.

Dans le chapitre 2 nous avons présenté en détail les méthodes de l'état de l'art que nous avons exploitées pour la réalisation de notre tâche. En commençant par la méthode TFIDF -présentée aussi dans le chapitre 1-, nous avons exposé les propositions de Luhn [46] sur lesquelles il s'est basé pour montrer que le mode de représentation adéquat est Bag of Words. En effet, il considérait que la répétition de certains mots est un indice pour insister sur un thème quelconque. Il considérait aussi que l'auteur utilise un sens unique d'un mot dans un texte.

Néanmoins ce n'est pas uniquement les propos de Luhn qui nous a bien évidemment poussé à utiliser cette méthode de représentation, en effet presque toutes les études réalisées dans ce domaine ont adopté cette voie, en atteignant de bons résultats. Dans cette première étude, nous nous sommes contentés de faire l'usage de cette méthode qui a toujours montré ses performances, toutefois nous mettons en perspective l'utilisation des autres modes cités ci-dessus.

Dans ce même chapitre, la méthode SVM et la méthode M-SVM sont exposées en détail. La justification intuitive de la SVM repose sur le fait que si les données d'apprentissage sont linéairement séparables, il semble naturel de les séparer d'une

façon qu'elles soient le plus loin possible du plan séparateur choisi. C'est l'idée principale de la SVM à partir de laquelle émanent les différentes notions que nous avons exposées dans ce chapitre. En effet, nous y avons vu comment obtenir l'hyperplan de séparation en maximisant la marge, et ceci se fait en résolvant un problème de programmation quadratique.

Outre la classification linéaire, nous avons passé en revue la classification non linéaire, car en pratique on rencontre souvent les données qui ne sont pas linéairement séparables.

La M-SVM est une version plus généralisée de la SVM. En effet, au lieu de traiter uniquement deux classes comme le fait la SVM, cette méthode adopte une approche directe en considérant l'ensemble des classes "ou des thèmes" en même temps. Dans ce manuscrit nous avons expérimenté celle proposée par Guermeur, qui a pour objectif de trouver un compromis satisfaisant entre la complexité et la performance de l'apprentissage.

Nous avons réservé le troisième chapitre aux différentes expériences réalisées dans cette thèse. Les méthodes exploitées sont des méthodes statistiques issues de l'état de l'art, en l'occurrence : la TFIDF, la SVM et la M-SVM. Le but est d'évaluer ces méthodes en utilisant un corpus en langue Arabe. Nous avons procédé à la collecte des textes arabes dans les journaux via Internet, et ceci en utilisant un logiciel conçu à cet effet (C.1). Toutefois, il a fallu réaliser plusieurs opérations pour obtenir des données prêtes à subir les deux tâches respectives d'apprentissage et de test.

La suppression des mots outils de la langue, le calcul de la fréquence des mots et la construction du vocabulaire constituent des exemple de ces opérations.

En réalisant nos premières expériences nous nous sommes basés, tout d'abord, sur un petit corpus constitué de 2500 documents, un nombre de thèmes égal à 4 et un vocabulaire de taille 74000 mots distincts. Nous avons testé les performances de la méthode TFIDF en préservant les mots outils. Les résultats ont montré que leur présence dégrade les performances de plus de 21% en terme de Rappel, nous avons ensuite augmenté la taille du corpus pour atteindre plus de 5000 documents. Le vocabulaire utilisé est d'une taille égale à 43000. Les performances en terme de Rappel se sont améliorées de 0.4% (de 90.45% à 90.81%).

La méthode SVM, outil efficace dans la séparation biclasses, a abouti à des résultats très satisfaisants. En effet, elle a dépassé la méthode TFIDF en terme de Rappel d'environ 7%.

En parallèle, nous avons réalisé une expérience dans le but d'accélérer le processus d'identification de thèmes. Cette tâche est primordiale dans les applications temps réel, comme c'est le cas des systèmes de Reconnaissance Automatique de la Parole. Cette expérience a montré que la prise en compte des n premiers mots du document de test, avec des valeurs très basses de n garde presque les mêmes performances.

Nous avons par la suite augmenté le nombre de thèmes à six pour voir si les performances restent stables. Par conséquent nous avons utilisé cette fois-ci un corpus relativement important dans le but d'avoir une meilleure représentativité des thèmes, environ 10 millions de mots. Les valeurs résultantes du Rappel concernant la TFIDF et la SVM sont respectivement de 91.08% et 96.53%, pour un vocabulaire

de taille 40000. Nous remarquons qu'elles ne se sont pas dégradées bien que nous avions élevé le nombre de thèmes. Nous pouvons dire que ces méthodes ont montré leur efficacité dans l'identification de thèmes en langue Arabe.

La SVM est conçue uniquement pour effectuer la séparation entre deux classes. Néanmoins, en pratique on se trouve souvent face à des problèmes traitant plusieurs catégories (> 2). Nous nous sommes donc intéressés par l'exploitation de la M-SVM, en utilisant la méthode proposée par Guermeur. C'est la première fois que celle-ci est utilisée en identification de thème.

Pour des raisons de complexité d'utilisation de la M-SVM, les tailles de vocabulaire utilisées sont limitées. Par conséquent pour une taille égale à 8000, nous avons obtenu un taux de Rappel égal à 90.45%. Tandis qu'en utilisant la méthode SVM, avec la même taille, nous avons obtenu une valeur de 81.11%.

Dans le quatrième chapitre, nous avons présenté une nouvelle méthode de détection de thème, le TR-classifier. L'objectif est de mesurer le degré de corrélation entre les mots qui constituent chacun des corpus thématiques, afin de mieux caractériser les thèmes. Dans cette expérience, chaque thème lui a été attribué un vocabulaire dont la taille ne dépasse pas quelques centaines de mots. Ces derniers sont rangés selon leurs fréquences.

Dans ces conditions, l'évaluation de cette méthode a été déroulée en variant les deux paramètres : la taille du vocabulaire et le nombre de triggers. Le nombre de triggers est important dans l'amélioration des résultats. Pour une taille donnée du vocabulaire de thème, nous avons remarqué une amélioration des performances de la méthode en augmentant le nombre de triggers. Ce qui est remarquable, c'est que les performances de cette méthode ont dépassé celles des autres méthodes sauf la SVM.

Pour voir clair, nous avons présenté dans l'annexe (E) les performances perçues de chaque méthode vis-à-vis de chacun des thèmes.

En perspective nous envisageons exploiter au maximum les performances des méthodes utilisées. Pour la M-SVM et le TR-classifier, il serait plus intéressant d'augmenter les tailles respectives du corpus d'apprentissage et des vocabulaires. Dans le côté du traitement de la langue elle même, notons que quelques aspects du traitement du corpus, comme l'extraction des racines, n'ont pas été effectués dans ce travail, vu la difficulté qui caractérise la langue Arabe. L'application de tels traitements serait très efficace dans la détection de thème.

Nous projetons l'exploitation des résultats, que nous avons obtenus, dans l'amélioration du rendement des systèmes de reconnaissance automatique de la parole (RAP). Pour ce faire, la solution qui sera envisageable se présente en l'adaptation d'un système de RAP existant comme celui de HTK ou CMU.

Bibliographie

[1] R. Kuhn, R. De Mori. A cache-based natural language model for speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(6), pp. 570-582, 1990.

[2] K. Seymore, R. Rosenfeld. Using Story Topics for Language Model Adaptation. In Proceeding of the European Conference on Speech Communication and Technology.

[3] M. Abbas, D. Berkani. Topic Identification by Statistical Methods for Arabic Language. Wseas Transactions on Computers, Issue 9. Volume 5. pp. 1908-1913. septembre 2006, Athens, Greece, ISSN 1109-2750.

[4] M. Abbas, D. Berkani, A Topic Identification Task for Modern Standard Arabic. 10th WSEAS International Conference on Computers, July 10-15, 2006, Vouliagmeni Beach, Athens, Greece ISSN : 1790-5117, ISBN : 960-8457-47-5, pp1092-1096.

[5] M. Abbas, K. Smaili. Comparison of Topic Identification Methods for Arabic Language. International conference RANLP05 : Recent Advances in Natural Language Processing, pp. 14-17, 21-23 septembre 2005, Borovets, Bulgary.

[6] M. Abbas, K. Smaili, D. Berkani. Topic Identification : Application to Arabic texts. JeTIC2007 : Journées d'étude sur les TICs, du 18 au 19 avril 2007, Centre Universitaire de Béchar, Algérie.

[7] M. Abbas, K. Smaili, D. Berkani. Identification de thèmes des textes arabes : Un outil pour la contribution au développement du projet Trésor de la Langue Arabe. Colloque International sur les travaux scientifiques du Professeur Hadj-Salah et les sciences modernes du langage, 3 et 4 juin 2008, CRSTDLA, Alger.

[8] Y. Yamashita, T. Tsunekawa, R. Mizoguchi. Topic Recognition for News Speech based on Keyword Spotting. Dans IEEE International Conference on Spoken Language Processing, Page 23, Sydney, Australia, 1998.

[9] M. Aery, N. Ramamurthy, Y. Alp. Aslandogan, Topic Identification of Textual Data. Technical Report CSE-2003-25, July 2003.

[10] A. Brun. Détection de thème et adaptation des modèles de langage pour la reconnaissance automatique de la parole. Doctorat de l'université Henri Poincaré, Nancy1, 2003.

[11] K. Tzeras, S. Hartman. Automatic Indexing Based on Bayesian Inference Networks. In : Proc. 16th Ann. Int.ACM SIGIR Conference on Research and development in Information Retrieval (SIGIR'93), 1993, pp. 22-34.

[12] DD. Lewis, M. Ringuette. Comparison of two Learning Algorithms for Text Categorization. In : Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94), 1994.

[13] I. Moulinier. Is Learning Bias an issue on Text Categorization Problem ?. Technical Report, LAFORIA-LIP6, Université Paris VI, 1997.

[14] N. Fuhr, S. Hartman, G. Lustig, M. Schwantner, K. Tzeras. A rule-based Multistage Indexing Systems for Large Subject fields. In : Proceedings of RIAO'91, 1991, pp.606-623.

[15] E. Wiener, J. O. Pedersen, A.S. Weigend. A neural network approach to topic spotting. In Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), pages 317-332, Nevada, Las Vegas, 1995. University of Nevada, Las Vegas.

[16] H. T. Ng, W.B. Goh, K.L. Low. Feature selection perceptron learning, and a usability case study for text categorization. In 20th Ann Int ACM SIGIR Conference on Research and De velopment in Information Retrieval (SIGIR'97), pages 67-73, 1997.

[17] R. H. Creecy, B. M. Masand, S. J. Smith, D. L. Waltz. Trading Mips and Memory for Knowledge Engineering : Calssifying Census Returns on the Connection Machine. Comm. ACM, 1992, 35 :48-63.

[18] Y. Yang. Expert Network : Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval. In : 17th Ann. Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94), 1994, pp. 13-22.

[19] F. Ciravegna, L. Gilardoni, A. Lavelli, M. Ferraro, N. Mana, S. Mazza, J. Matiasek, W. Black, F. Rinaldi. Flexible Text Classification for Financial Applications : the FACILE System. In Proceedings of PAIS-2000, Prestigious Applications of Intelligent Systems sub-conference of ECAI2000. (2000).

[20] F. Peng, X. Huang, D. Schuurmans, S. Wang. Text Classification in Asian Languages without Word Segmentation. In Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages (IRAL 2003), Association for Computational Linguistics, July 7, Sapporo, Japan. (2003).

[21] M. El-Kourdi, A. Bensaid, T. Rachidi. Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm. 20th International Conference on Computational Linguistics . August 28th. Geneva (2004).

[22] H. Sawaf, J. Zaplo, H. Ney. Statistical Classification Methods for Arabic News Articles. Arabic Natural Language Processing, Workshop on the ACL'2001. Toulouse, France, July (2001).

[23] A. El-Halees. Mining Arabic Association Rules for Text Classification. In the proceedings of the first international conference on Mathematical Sciences. Al-Azhar University of Gaza, Palestine, 15 -17 (2006). To be appear.

[24] A. El-Halees. Arabic Text Classification Using Maximum Entropy. The Islamic University Journal (Series of Natural Studies and Engineering) Vol. 15, No.1, pp 157-167, 2007, ISSN 1726-6807, http//www.iugzaza.edu.ps/ara/research/.

[25] B. Bigi, K. Smaïli. Identification thématique hiérarchique : Application aux forums de discussions. TALN 2002, Nancy, France, 24–27 juin 2002.

[26] B. Masand, G. Linoff, D. Waltz. Classifying news stories using memory based reasoning. In 15th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92), pages 59-64, 1992.

[27] M. Iwayama, T. Tokunaga. Cluster-based text categorization : a comparison of category search stratefies. In 18th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95), pages 273-281, 1995.

[28] Y. Yang, X. Liu. A re-examination of text categorization methods. Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99, pp 42–49), 1999.

[29] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Rapport techique, School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213, 1996.

[30] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 144-152, Pittsburgh, 1992.

[31] C. Cortes and V. Vapnik. Support-vector network. Machine Learning. vol. 20, pp. 273-297, 1995.

[32] D. Langlois. Notions d'événements distants et d'événements impossibles en modélisation stochastique du langage : appalication aux modèles n-grammes de mots et de séquences. Doctorat de l'université Henri Poincaré, Nancy1, Juillet 2002.

[33] J.P. Haton, C. Cerisara, D. Fohr, Y. Laprie, K. Smaili. Reconnaissance automatique de la parole Du signal son interprtation , DUNOD (370 pages), mai 2006.

[34] P. R. Clarkson, A. J. Robinson. Language model adaptation using mixtures and an exponentially decaying cache. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pages 799-802, Munich, Allemagne, Avril 1997.

[35] R. Rosenfeld. Adaptive Statistical Language Modeling : A Maximum Entropy Approach. PhD thesis, Computer Science Department, Carnegie Mellon University,TR CMU-CS-94-138, April 1994.

[36] F. Jelinek et al. Perplexity a mesure of Difficulty of Speech Recognition Tasks. 9th Meeting of the Acoustical Society of America, Miami, 1977.

[37] I. Dagan , Y. Karov, D. Toth. Mistake-driven learning in text categorization. In 2nd conference on Empirical Methods in Natural Language Processing, EMNLP-97, pages 55-63, Providence, US, 1997.

[38] D. Lewis. Representation and Learning in Information Retrieval. Technical Report UMCS- 1991-093. Department of Computer Science, University of Massachusetts, Amherst, MA.

[39] D. Lewis. Feature selection and feature extraction for text categorization. In Proceedings of a Workshop on Speech and Natural Language, (pp. 212-217). San Mateo, CA : Morgan Kaufmann.

[40] D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In Croft et. al. (Ed.), Proceedings of SIGIR-95, 15th ACM International Conference on Research and Development in Information Retrieval (pp. 37-50). New York : ACM Press.

[41] Chade-Meng Tan et all. The use of bigrmas to Enhance Text Categorization.

[42] D. Mladenic, M. Grobelnik. Word sequences as features in text learning. In Proceedings of the 17th Electrotechnical and Computer Science Conference (ERK-98) (pp. 145-148), 1998, Ljubljana, Slovenia.

[43] Fürnkranz. A Study Using n-gram features for Text Categorization. Technical Report OEFAI-TR-98-30, Austrian Research Institute for Artificial Intelligence, Vienna, Austria, (1998).

[44] Y. Guermeur, G. Pollastri, A. Elisseff, D. Zelus, H. Paugam-Moisy, P. Baldi. Combining protein secondary structure prediction models with ensemble methods of optimal complexity. Neurocomputing 56 (2004) 305-327.

[45] J. Rocchio. Relevance Feedback in Information Retrieval. in The SMART Retrieval System : Experiments in Automatic Document Processing, Chapter 14, pages 313- 323, Prentice-Hall, 1971.

[46] H. P. Luhn. A Statistical Approach to Mechanized Encoding and Searching of Library Information. IBM Journal, pp. 309-17, 1957.

[47] R. Caruana, D. Freitag. Gready Attribute Selection. International Conference on Machine Learning. 1994.

[48] Y. Yang et J.O. Pedersen. A comparative study on feature selection in text categorization. In 14th International Conference on Machine Learning, ICML-97, redacteur Douglas H. Fisher, pages 412-420, San Francisco, US, 1997. Morgan Kaufmann.

[49] M. Maron. Autoamtic indexing : an experimental inquiry. Journal of the Association for Computing Machinery, 1961.

[50] D.J. Ittner, D.D. Lewis, D.D. Ahn. Text Categorization of Low Quality Images. In 4th Annual Symposium on Document Analysis and Information Retrieval, SDAIR-95, pages 301-315, 1995.

[51] F. Sebastiani. Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1) :l-47, 2002.

[52] D. Mladenic. Machine Learning on non-homogeneous distributed text data. Phd thesis, Computer and Information Science, 1998.

[53] K. Seymore, S. Chen, R. Rosenfeld. Nonlinear Interpolation of Topic Models for Language Model Adaptation. In Proceedings of the International Conference on Spoken Language Processing, 1998.

[54] Y. Guermeur and H. Paugam-Moisy. Théorie de l'apprentissage de Vapnik et SVM, Support Vector Machines. READ, Vol. 3, N. 1, 17-38, 1999.

[55] J. Weston, C. Watkins. Multi-class support vector machines. Technical report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science, 1998.

[56] V. N. Vapnik. Statistical Learning Theory. John Wiley & Sons, Inc., N.Y., 1998.

[57] Y. Guermeur, A. Elisseeff, H. Paugam-Moisy. A new multi-class SVM based on a uniform convergence result. IJCNN'00, Come, Vol. IV, 183-188, 2000.

[58] A. Elisseeff, Y. Guermeur, H. Paugam-Moisy. Margin error and generalization capabilities of multi-class discriminant systems. Technical Report 1999-051, NeuroCOLT2, 1999.

[59] A. Abdelali, J. Cowie. Regional corpus of modern standard Arabic. In second international conference on Arabic Language Engineering, 2005, Vol. 1, No. 1, 2005, pp. 1-12.

[60] A. Abdelali, J. Cowie, H. Soliman. Building a modern standard corpus. Workshop on Computational Modeling of Lexical Acquisition. The Split Meeting, Split, 2005.

[61] M. Mahajan, D. Beeferman, X.D. Huang. Improved Topic-Dependent Language Modeling Using Information Retrieval Techniques. In IEEE Transactions on Acoustics, Speech, and Signal Processing, 1999.

[62] C. Apté, F. Damerau, S.M. Weiss. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, 12(3) :233-251, 1994.

[63] A. Khoo, Y. Marom, D. Albrecht. Experiments with Sentence Classification, ALTW06.

Annexe A

Corpus de test thématiques

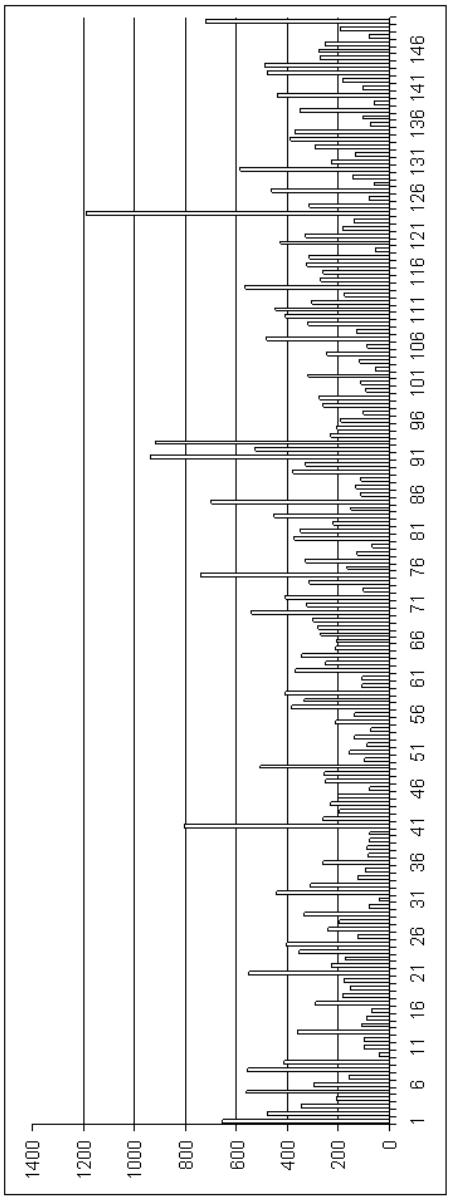


FIG. A.1 – Nombre de mots constituant les documents de test du thème Culture

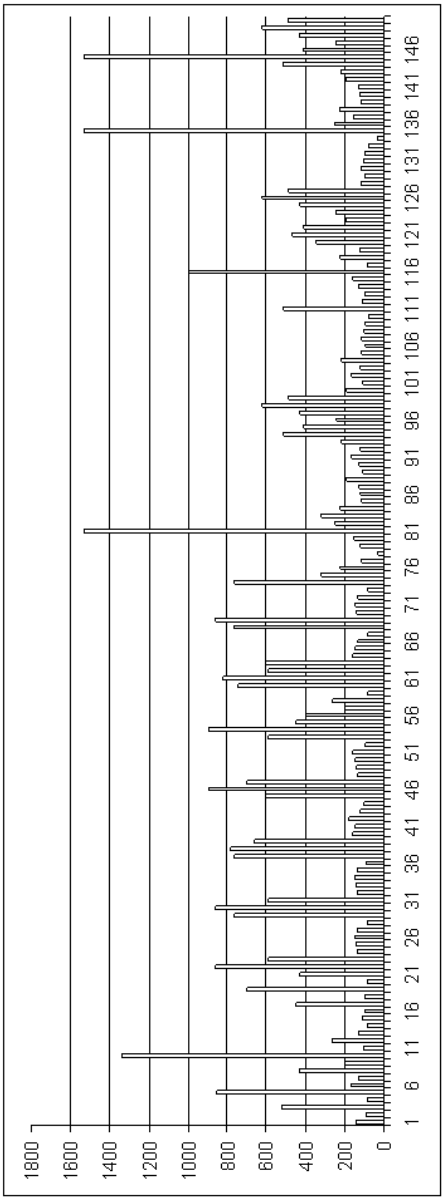


FIG. A.2 – Nombre de mots constituant les documents de test du thème Religion

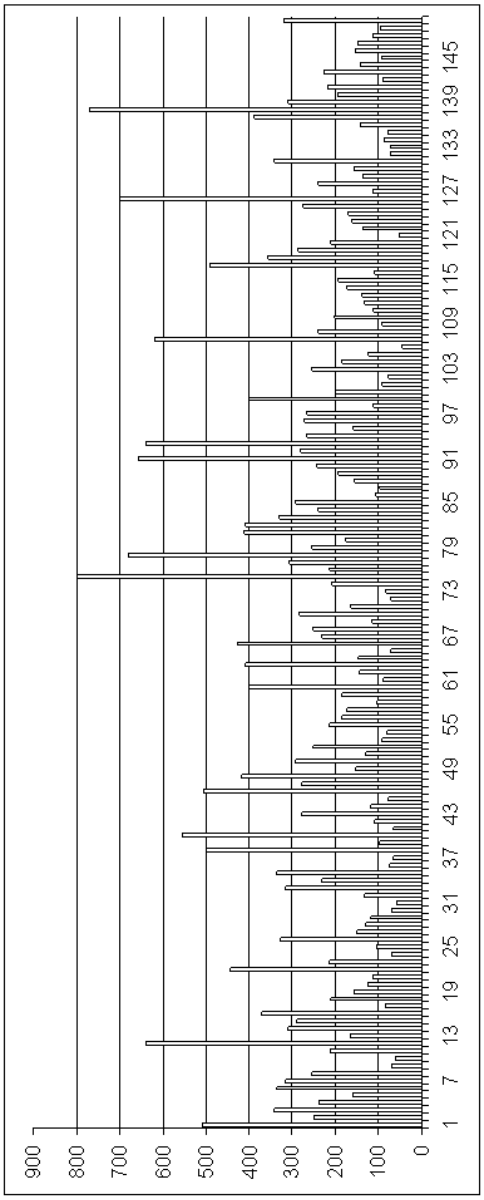


FIG. A.3 – Nombre de mots constituant les documents de test du thème Économie

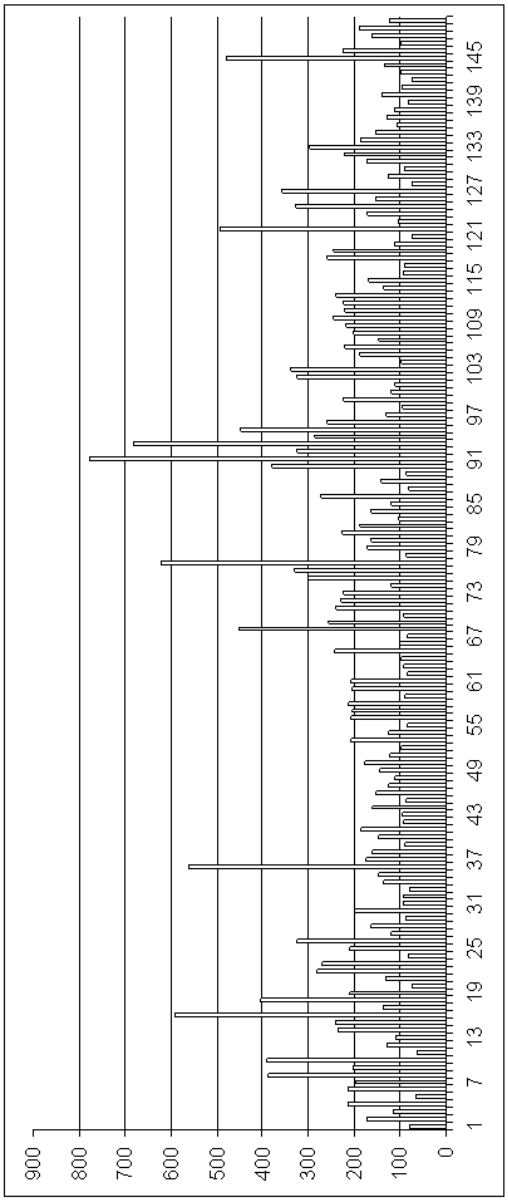


FIG. A.4 – Nombre de mots constituant les documents de test du thème Local

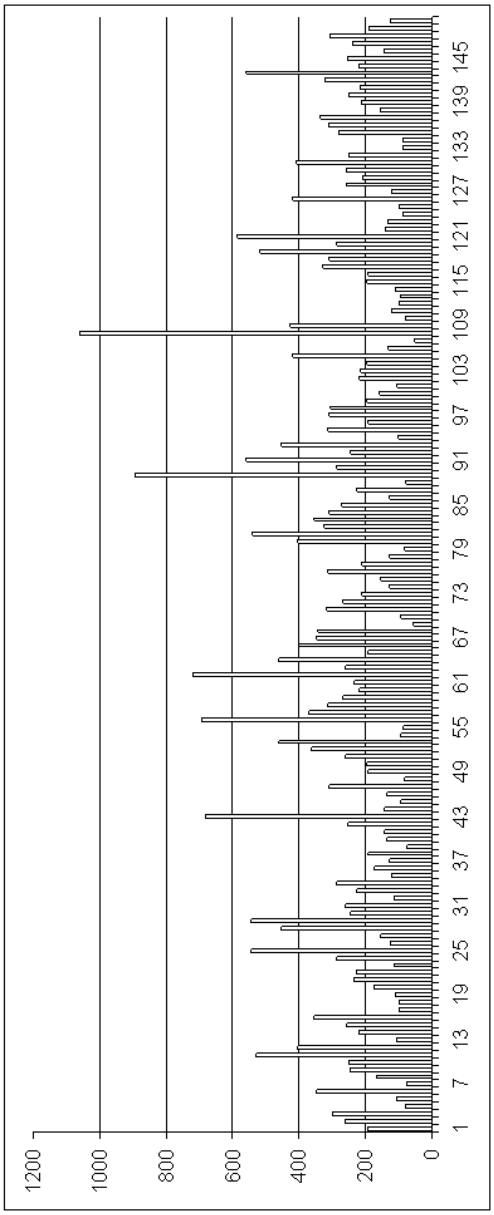


FIG. A.5 – Nombre de mots constituant les documents de test du thème International

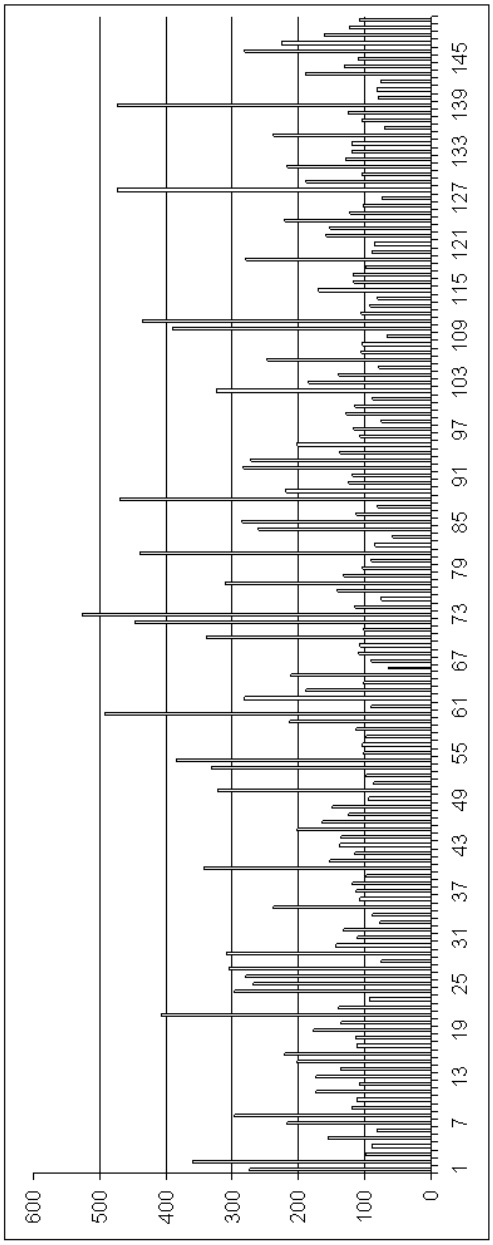


FIG. A.6 – Nombre de mots constituant les documents de test du thème Sports

Annexe B

Identification dynamique

n	Rappel (%	Précision (%	F_1(%
5	10.93	93.33	19.56
10	18.75	92.31	31.16
15	21.09	96.43	34.61
20	23.44	96.77	37.73
25	28.12	97.3	43.63
30	34.37	97.77	50.86
40	42.19	98.18	59.02
50	47.65	98.38	64.20
60	56.25	100	72
70	66.4	100	79.81
80	70.31	100	82.56
100	77.34	100	87.22
120	81.25	100	89.65
140	86.72	100	92.88
160	90.62	100	95.07
200	91.41	100	95.51
240	92.19	100	95.93
280	92.19	100	95.93

TAB. B.1 – Performances obtenues en fonction du nombre de mots -thème Sport-

n	Rappel (%	Précision (%	F_1(%
5	46.1	65.55	54.13
10	57.03	77.66	65.76
15	66.41	84.16	74.23
20	63.28	84.37	72.32
25	65.62	88.42	75.33
30	71.09	90.1	79.47
40	71.87	92.93	81.05
50	70.31	90	78.94
60	74.22	89.62	81.19
70	78.9	90.18	84.16
80	79.68	88.69	83.94
100	82.81	86.88	84.79
120	84.37	87.09	85.71
140	83.59	85.6	84.58
160	84.37	85.04	84.70
200	85.16	85.83	85.49
240	85.94	85.94	85.94
280	85.16	85.83	85.49

TAB. B.2 – Performances obtenues en fonction du nombre de mots -thème Économie-

n	Rappel (%	Précision (%	F_1(%
5	75.78	39.11	51.59
10	85.94	48.03	61.62
15	85.16	50.23	63.19
20	85.94	50.46	63.58
25	87.5	52.09	65.30
30	97.65	62.19	75.98
40	89.06	57.57	69.93
50	86.72	57.81	69.37
60	85.94	61.11	71.42
70	87.5	67.06	75.93
80	85.94	67.48	75.60
100	83.59	69.93	76.15
120	85.15	72.18	78.13
140	84.37	73.97	78.83
160	84.37	77.14	80.59
200	85.16	77.86	81.34
240	85.94	79.14	82.40
280	85.94	78.57	82.10

TAB. B.3 – Performances obtenues en fonction du nombre de mots -thème Local-

n	Rappel (%	Précision (%	F_1(
5	75	60.37	66.89
10	85.94	67.48	75.60
15	89.06	68.67	77.55
20	94.53	72.45	82.03
25	95.31	73.94	83.27
30	85.15	66.06	74.40
40	98.44	78.75	87.50
50	99.22	80.38	88.81
60	99.22	82.47	90.07
70	99.22	85.81	92.03
80	99.22	88.19	93.38
100	98.44	91.3	94.73
120	98.44	94.74	96.55
140	98.44	96.92	97.67
160	98.44	97.67	98.05
200	98.44	98.44	98.44
240	98.44	99.21	98.82
280	98.44	99.21	98.82

TAB. B.4 – Performances obtenues en fonction du nombre de mots -thème Mondial-

Annexe C

Outils et implémentation

C.1 L'outil WinHTTrack

L'outil WinHTTrack nous a permis de collecter un nombre considérable de pages Web contenant les articles dont nous nous sommes servis pour la construction de notre corpus. Pour s'en servir il suffit de remplir le champ d'édition réservé à l'adresse URL du site choisi. Comme dans notre cas nous avons utilisé plusieurs adresses : www.akhbaralkhaleej.com, www.alwatan.com, www.elkhabar.com, etc. Il y a en outre la possibilité de sélectionner les options qui facilitent la tâche de téléchargement. Par exemple nous pouvons exclure des fichiers de type pdf et permettre aux fichiers html d'être téléchargés.

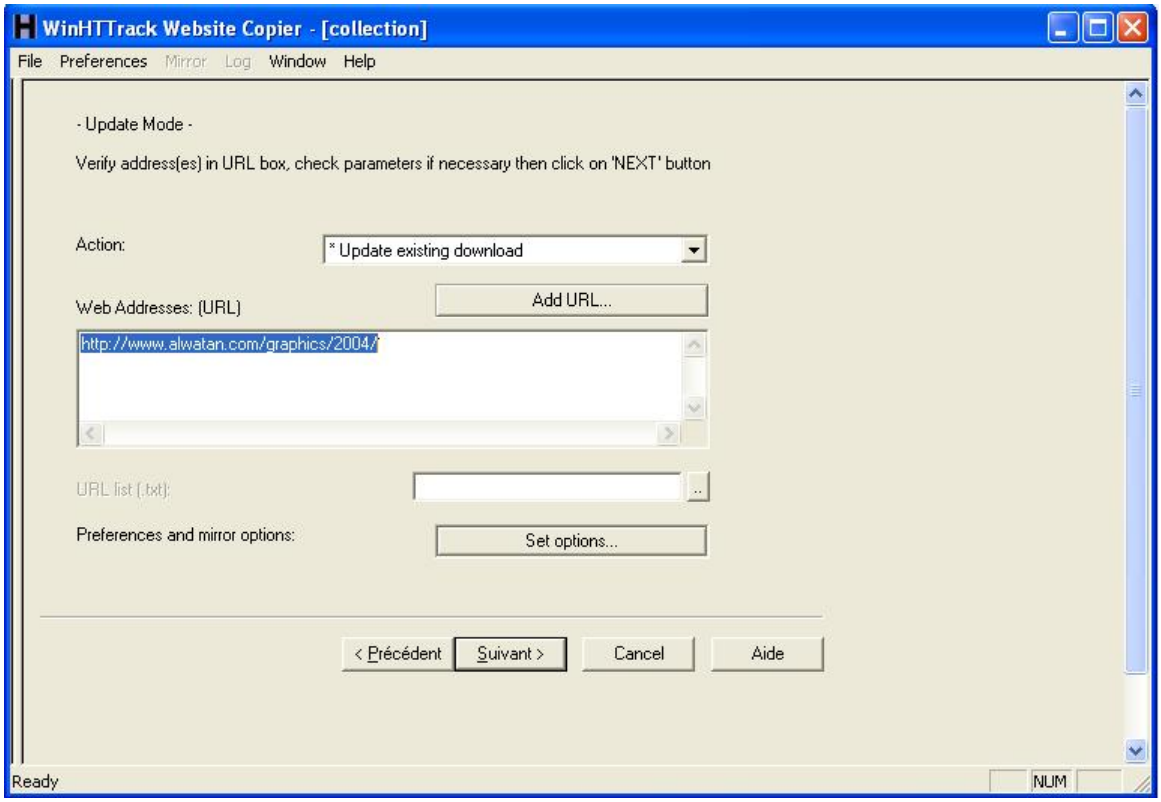


FIG. C.1 – L'outil Winhttrack utilisé dans la collecte de corpus

C.2 Logiciel d'identification de thèmes

La figure (C.2) présente l'interface du logiciel que nous avons implémenté. Il comporte tous les modules que nous avons programmé au cours de la préparation de cette thèse, en commençant par l'extraction des textes à partir des pages html, calcul des fréquences des mots, etc, jusqu'aux méthodes sélectionnées d'identification de thèmes.

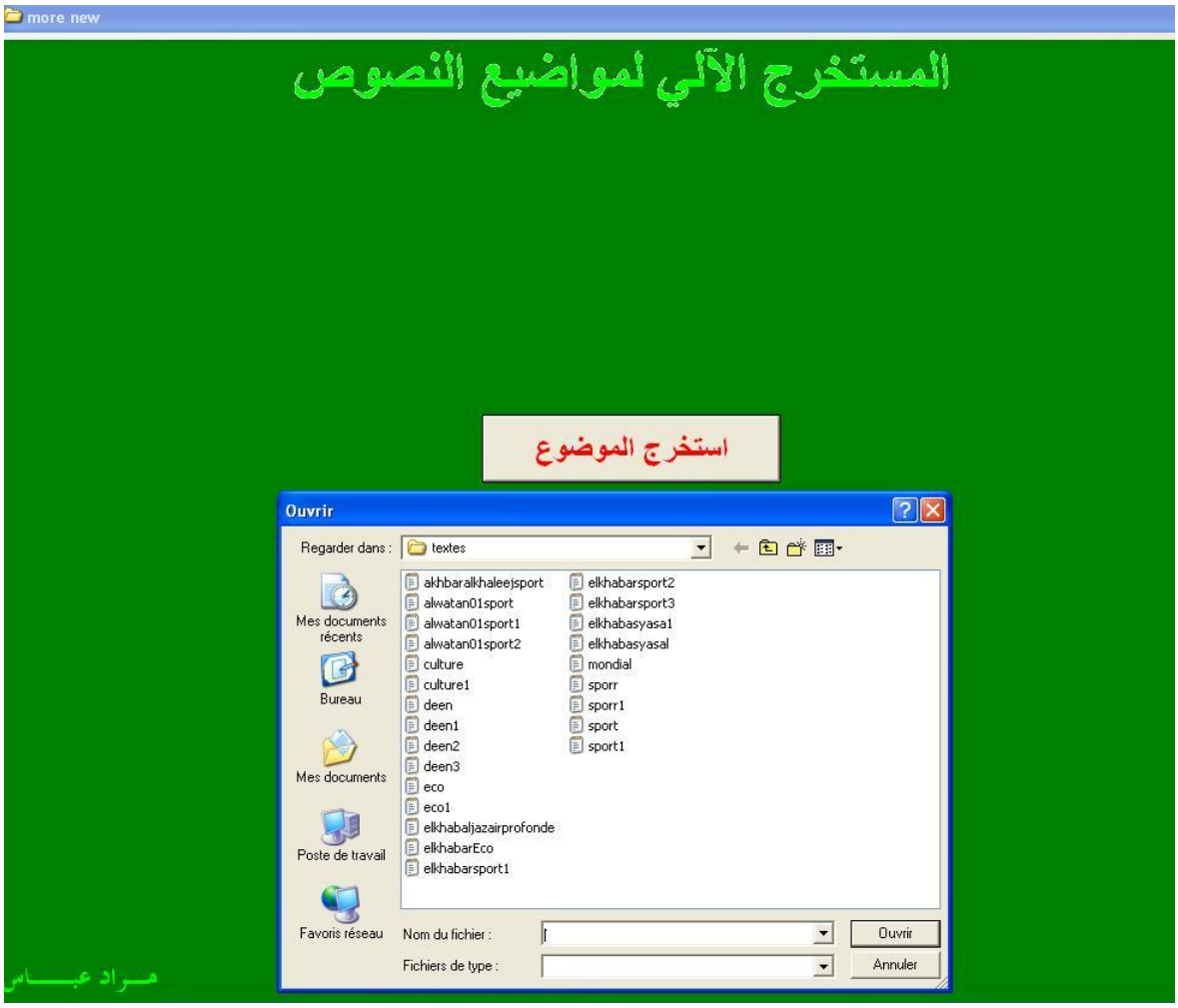


FIG. C.2 – Logiciel d'identification de thèmes

Annexe D

TR-classifier

D.1 Performances du TR-Classifier pour une taille 200 de chacun des vocabulaires thématiques

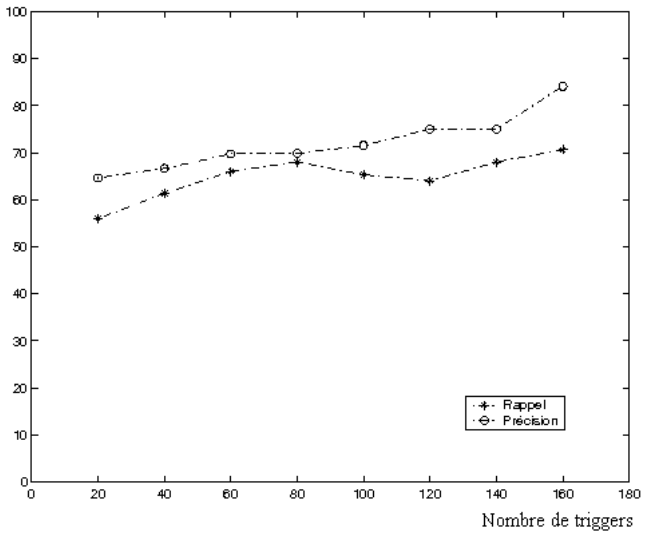


FIG. D.1 – Performances du TR-Classifier concernant le thème Culture -Taille du vocabulaire 200-

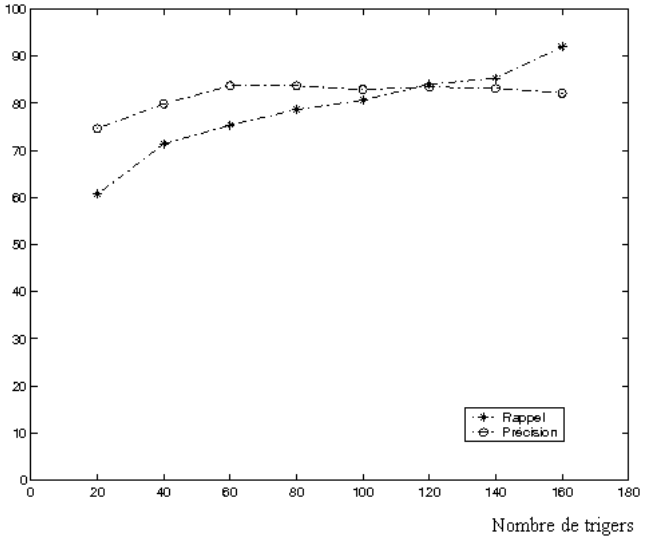


FIG. D.2 – Performances du TR-Classifier concernant le thème Religion -Taille du vocabulaire 200-

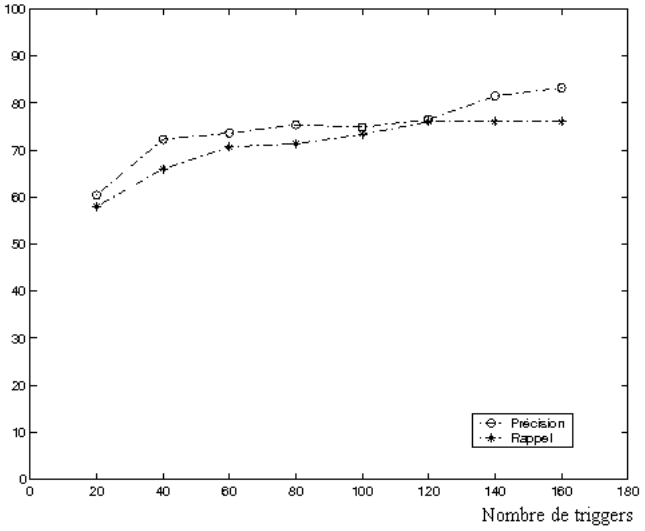


FIG. D.3 – Performances du TR-Classifier concernant le thème Économie -Taille du vocabulaire 200-

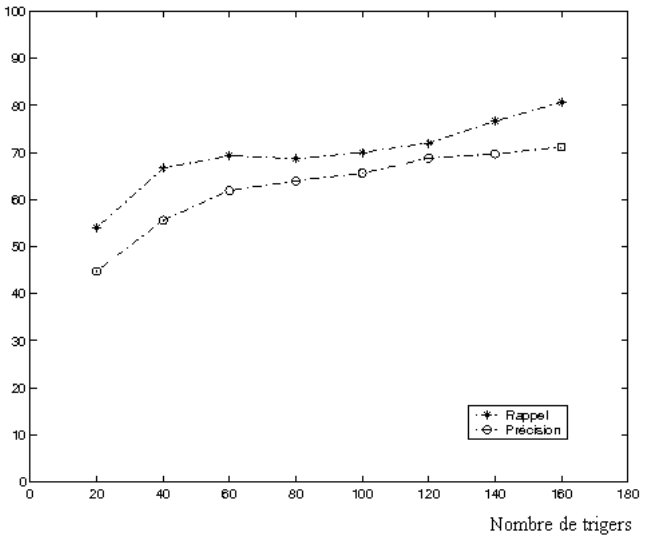


FIG. D.4 – Performances du TR-Classifier concernant le thème Local -Taille du vocabulaire 200-

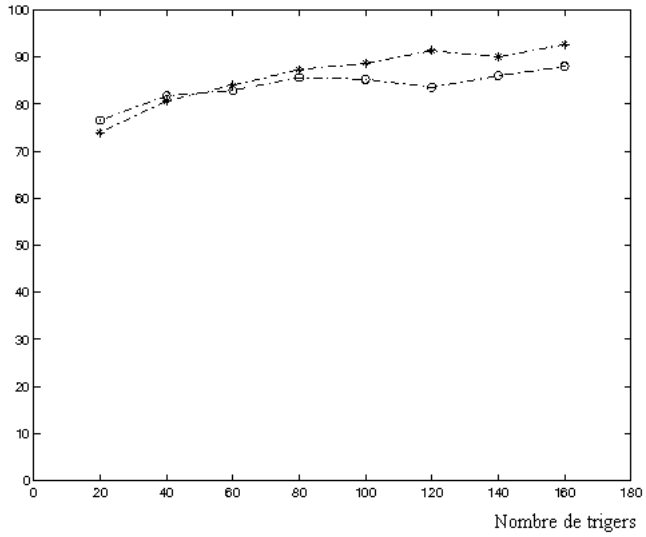


FIG. D.5 – Performances du TR-Classifier concernant le thème International -Taille du vocabulaire 200-

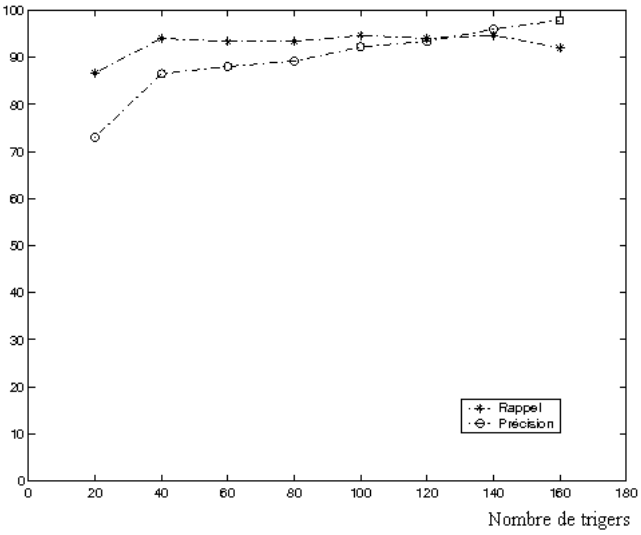


FIG. D.6 – Performances du TR-Classifier concernant le thème Sports -Taille du vocabulaire 200-

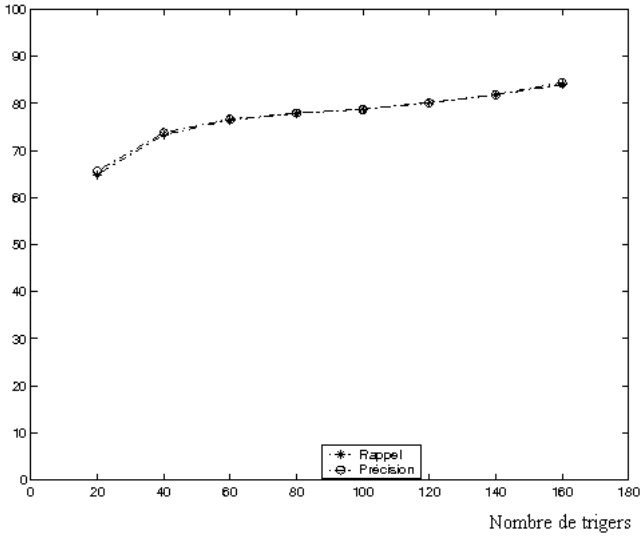


FIG. D.7 – Moyenne des valeurs de Rappel et de Précision -Taille du vocabulaire 200-

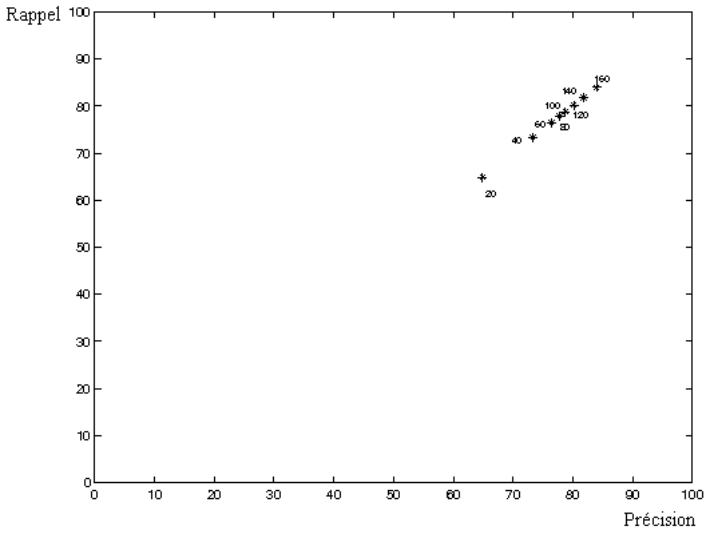


FIG. D.8 – Rappel en fonction de la Précision -Taille du vocabulaire 200-

Annexe E

Récapitulatif des performances des méthodes pour chacun des thèmes

Méthode	Rappel (%)	Précision (%)	F_1(%)
TFIDF	71.33	88.43	78.96
SVM	97.33	95.51	96.41
M-SVM	75	78	76.47
TR	82.66	80.55	81.60

TAB. E.1 – Récapitulatif des performances des méthodes utilisées -Thème culture-

Méthode	Rappel (%)	Précision (%)	F_1(%)
TFIDF	93.33	86.95	90.03
SVM	96.93	99.32	98.11
M-SVM	95	96	95.49
TR	96.33	83.56	89.50

TAB. E.2 – Récapitulatif des performances des méthodes utilisées -Thème Religion-

Méthode	Rappel (%)	Précision (%)	F_1(%)
TFIDF	83.33	80.64	81.96
SVM	96.26	96.57	96.41
M-SVM	83.5	75	79.02
TR	83.50	84.05	83.77

TAB. E.3 – Récapitulatif des performances des méthodes utilisées -Thème EconomieÉconomie-

Méthode	Rappel (%)	Précision (%)	F_1(%)
TFIDF	80	76.92	78.43
SVM	96.13	96.55	96.34
M-SVM	74	64	68.64
TR	86.25	82.53	84.35

TAB. E.4 – Récapitulatif des performances des méthodes utilisées -Thème Local-

Méthode	Rappel (%)	Précision (%)	F_1(%)
TFIDF	93.33	84.33	88.60
SVM	98.26	96.88	97.56
M-SVM	86.75	83	84.83
TR	93.33	90.66	91.97

TAB. E.5 – Récapitulatif des performances des méthodes utilisées -Thème International-