وزارة التعليم العالي و البحث العلمي

BADJI-MOKHTAR UNIVERSITY–ANNABA-

UNIVERSITÉ BADJI MOKHTAR–ANNABA-

جامعة باجي مختار
ـعنابةـ

Faculté : Sciences de l'ingénieur – Année 2009 –

Département : Informatique

MEMOIRE

Présentation en vue de l'obtention du diplôme de magister

Identification d'opinions dans les journaux arabes

Option : Texte, Parole et Imagerie

Par

Lazhar FAREK

DIRECTEUR DE MEMOIRE : MC. Tlili Guiassa Yamina

DEVANT LE JURY

PRESIDENT: Mr. Laskri Mohamed Tayeb Prof. Université Badji-Mokhtar - Annaba

EXAMINATEURS:

- Dr. Kholladi Mohemed Kheireddine MC. Université Mentouri - Constantine
- Dr. Merouani Hayet Farida MC. Université Badji Mokhtar - Annaba

Remerciement

Je tiens à remercier chaleureusement Madame Tlili Guiassa Yamina, mon encadreur pour ses précieux conseils au cours de la période de mon travail.

Un grand merci à Monsieur Laskri Mohamed Tayeb d'avoir accepté d'être le président de jury.

Un remerciement particulier à Madame Merouani Hayet Farida d'avoir accepté d'être un membre de jury.

Ma reconnaissance à Monsieur Kholadi Kheireddine d'avoir accepté d'être un membre de jury.

Un très grand merci pour mes amis Mechouma Toufik et Hachemi Samir pour avoir été présent pour m'aider, et m'encourager.

Dédicace

Je dédie ce modeste travail à mes chères parents, toute ma famille et à tous mes amis.

Résumé

Après une demande croissante en matière d'analyse de textes véhiculant des critiques, des opinions ou des jugements, le traitement automatique des langues a donné naissance à une nouvelle discipline appelée *fouille de données d'opinions* (opinion-mining). Cette discipline est un résultat de l'intersection de trois disciplines : le traitement automatique des langues, la linguistique et la philosophie. Elle n'est pas intéressée par l'étude des thèmes d'un document mais par les opinions exprimées dans les textes.

Dans ce présent mémoire, nous avons proposé une approche d'identification d'opinions basée sur une analyse symbolique des textes, qu'on testé sur un ensemble de textes journalistiques de la langue arabe. Cette dernière, malgré ses particularités syntaxiques, morphologiques et sémantiques, présente l'axe sur lequel s'articule notre travail. Nous avons inspiré notre approche, après une étude comparative de trois approches utilisées en classification de sentiments : l'approche symbolique basée sur l'analyse syntaxique des textes, l'approche statistique basée les techniques d'apprentissage automatique, et la troisième approche, est une hybridation des deux premières. Son fonctionnement est basé sur l'extraction des expressions subjectives qui reflètent des jugements personnels sur des sujets divers.

Notre approche, se base sur un modèle de représentation d'opinion, qui considère qu'une opinion est conformée de quatre éléments : *prédicat*, *source*, *sujet* et *contenu*. L'identification de chaque élément nécessite un ensemble règles linguistiques bien définies. Deux éléments appelés *polarité* et *intensité*, ont été ajoutés à ce modèle pour calculer l'orientation sémantique globale de l'opinion en fonction de ces constituants.

Les opinions identifiées sont ensuite classifiées selon leurs orientations sémantiques et leurs intensités, en cinq catégories : positive forte, positive faible, négative forte, négative faible, et neutre.

Notons, que notre travail a été l'un des 26 articles acceptés parmi les 62 articles soumis à la conférence *IC2009*.

Mots Clés : Langue Arabe, TALN, Opinion, Identification, Objectivité, Subjectivité, Classification.

ملخص

بعد الطلب المتزايد على تحليل النصوص التي تحمل انتقادات ، أراء أو أحكام أعطت المعالجة الآلية للغة ميلاد علم جديد سمي – البحث في الرأي – ، هذا العلم هو نتيجة لالتقاء ثلاثة علوم وهي : المعالجة الآلية للغة ، علم اللسانيات و الفلسفة ، لا يهتم هذا العلم بدراسة المواضيع المطروحة في الوثائق و لكن بدراسة الآراء المعبر عنها في النصوص .

في رسالتنا هذه اقترحنا طريقة لتحديد الآراء ترتكز على التحليل الرمزي للنصوص، حيث تم تجريبها على مجموعة من النصوص الصحفية المكتوبة باللغة العربية. هذه الأخيرة رغم ما تحويه من خصائص نحوية، شكلية و معنوية مميزة فهي تشكل المحور الرئيس الذي يرتكز عليه عملنا.

طريقتنا هذه تم استنتاجها بعد القيام بإجراء مقارنة بين ثلاثة طرق تستخدم لتصنيف الأحاسيس: الطريقة الرمزية تعتمد على المعالجة النحوية للنصوص، الطريقة الإحصائية تعتمد على تقنيات التلقين الآلي، و الطريقة الثالثة هي هجين بين الطريقة الرمزية و الإحصائية. يرتكز عمل طريقتنا على استخراج العبارات غير الموضوعية التي تعكس الأحكام الذاتية في مواضيع مختلفة.

اعتمدنا في طريقتنا على نموذج لتمثيل الرأي و يعتبر هذا النموذج أن الرأي يتكون من أربعة عناصر وهي : المسند ، المصدر ، الموضوع و المحتوى ، يتم تحديد كل عنصر من هذه العناصر وفقا لمجموعة من القواعد اللغوية مبنية على أسس صحيحة ،كما تم إضافة عنصرين لهذا النموذج وهما : القطبية و الشدة لحساب التوجه المعنوي الكلي للرأي وفقا للعناصر المشكلة له .

الآراء المعرفة يتم تصنيفها في الأخير وفقا لتوجهاتها المعنوية وشدتها إلى خمسة أصناف: موجبة قوية، موجبة ضعيفة، سالبة قوية، سالبة ضعيفة، و معتدلة.

للتذكير فإن عملنا هذا من الأعمال الستة و العشرون التي تم قبولها من بين اثنين و ستين عملا المسلمة إلى محاضرة IC2009.

الكلمات الرئيسية: اللغة العربية، المعالجة الآلية للغة، الرأي، تحديد، الموضوعية، الذاتية، التصنيف.

Abstract

After an increasing demand of analysis of texts conveying of criticisms, opinions or judgments, the automatic treatment of the languages gave rise to a new discipline called *Opinion-mining*. This discipline is a result of the intersection of three disciplines: automatic treatment of the languages, linguistics and philosophy. It is not interested by the study of the topics of a document but by the opinions expressed in the texts.

In this present thesis, we proposed an approach for identification of opinions, based on a symbolic analysis of the texts, which is tested on a set of journalistic texts of the Arab language. The latter, in spite of its syntactic, morphological and semantic characteristics, presents the axis on which our work is articulated. We inspired our approach, after a comparative study of three approaches used in classification of feelings: the symbolic approach based on the syntactic analysis of the texts, statistical approach based on techniques of machine learning, and the third approach, is a hybridization of the two first. Its operation is based on the extraction of the subjective expressions which reflect personal judgments on various subjects.

Notre approach, bases on a model of representation of opinion, which considers that an opinion is formed of four elements: *predicate*, *source*, *subject* and *contents*. The identification of each element requires a set of linguistic rules well defined. Two elements called *polarity* and *intensity* were added to this model to calculate the total semantic orientation of the opinion according to these components.

The identified opinions are then classified according to their semantic orientations and their intensities, in five categories: strong positive, weak positive, strong negative, weak negative, and neutral.

Note, that our work was one of the 26 articles accepted from the 62 articles submitted to the conference *IC2009*.

Key words: Arabic Language, TALN, Opinion, Identification, Objectivity, Subjectivity, Classification.

Liste des tableaux

Liste des figures

Table des matières

Introduction

Les commerçants vendant des produits sur le Web demandent souvent à leurs clients de passer leurs avis sur les produits qu'ils ont achetés et les services associés, les journalistes qui rédigent des articles, laissent souvent leurs avis ou les avis d'autres personnes ou d'autres organismes sur les pages des journaux.

Les journaux, les blogs politiques, les sites Web des consommateurs, et les forums de discussion, sont juste quelques exemples des ressources d'opinions textuelles disponibles aux lecteurs en format brut. Ces ressources sont considérées comme référence pour les sociétés qui veulent savoir les opinions de leurs clients sur un produit ou un service, et les sondeurs qui veulent savoir les opinions publiques sur un sujet politique ou autres,…etc.

Le problème considéré par tous les *consommateurs d'opinions*[1] est qu'il y a une telle richesse des textes à traiter, et qu'il est difficile de tout lire, ce qui pose le problème d'exploiter ces ressources en temps et en coût très élevé pour avoir ce qui est nécessaire, d'où une nécessité de traiter les textes bruts et l'extraction des expressions pertinentes qui peuvent être subjectives ou objectives, peut amener ces consommateurs de connaître l'orientation des rédacteurs d'opinions.

Les techniques sont maintenant développées pour exploiter ces ressources pour aider des organismes et des individus à obtenir les informations importantes facilement et rapidement [Hu et al., 2006]. Ces techniques sont les fruits des travaux de recherche dans le domaine de text-mining et plus précisent dans l'opinion-mining[2] soit pour l'extraction ou la classification d'opinions.

1. Problématique et objectifs

L'identification d'opinions est un ensemble de techniques qui fait partie du domaine du traitement automatique du langage naturel et plus précisément de la recherche de l'information, qui consiste à extraire les opinions disponibles sur des masses textuelles importantes, et les classifier en catégories. L'identification d'opinions est un axe d'actualité qui cherche à déterminer les avis ou les critiques des individus et des organismes a propos d'un sujet ou un d'un objet.

La disponibilité d'une masse textuelle sous forme d'articles de journaux en langue arabe en format électronique impose une technique d'exploration particulière. Notre sujet porte sur l'identification d'opinions dans les journaux arabes. Cette tâche se définie comme un processus :

[1] Ceux qui utilisent les opinions d'autres individus ou d'autres organismes, par exemple pour améliorer la qualité d'un service ou d'un produit.
[2] FODOP 2008, est la première conférence en fouille de données d'opinions.

- Il est nécessaire de prétraiter les données brutes pour pouvoir extraire les informations pertinentes ;

- La notion de similarité entre documents est fortement liée au choix de la méthode de représentation des textes ;

- La classification est une méthode d'analyse de données qui vise à regrouper en classes homogènes un ensemble d'observations.

Notre système d'identification est un système qui cherche à identifier et classifier les opinions, en commençant par le prétraitement des textes brutes, en les représentant pour qu'on puisse les manipuler, enfin nous regroupons les opinions similaires. Les résultats obtenus doivent êtres évalués et commentés.

2. Organisation du mémoire

Dans notre travail, nous nous s'intéressons par une ressource d'opinions classique, qui est les journaux, et plus précisément les journaux arabes, cette dernière est considérée comme une excellente ressource d'opinions, où les gens et les organismes expriment leurs points de vue sur presque n'importe quoi, et dans n'importe quel domaine : politique, économique, social, cultuel…etc.

Selon la nature de notre projet, nous avons organisé ce présent mémoire en six chapitres comme suit :

Le premier chapitre est sur l'identification d'opinions, dans lequel on explique les différents concepts en fouille de données d'opinons. Dans le deuxième chapitre, on explique les techniques utilisées pour prétraiter les textes arabes. Le troisième chapitre est dédié pour exposer les différentes approches de représentation de textes, en effectuant une comparaison entre elles. Le quatrième chapitre pour représenter les différentes techniques et algorithmes utilisés en classification d'opinions. Le cinquième chapitre pour expliquer notre approche d'identification d'opinons, l'architecture de notre système d'identification que nous avons appelés **IOJAR** (**I**dentification d'**O**pinions dans les **J**ournaux **AR**abes), en expliquant le modèle utilisé pour représenter une opinion, ainsi que l'approche utilisée pour classifier toutes les opinions extraites. Dans ce chapitre nous avons abordés, entre autres, les points suivants :

1. Comment représenter une opinion ?
2. Quels sont les différents constituants d'une opinion ?
3. Comment extraire ses constituants à partir du texte ?
4. Comment déterminer l'orientation sémantique d'une opinion ?
5. L'approche utilisée pour classifier les opinions identifiées.

Le sixième chapitre est pour présenter l'implémentation de notre système, en expliquant les outils de développement, ainsi que le fonctionnement de notre application, illustrée par des captures écran.

CHAPITRE 1

Identification d'opinions

1.1. Introduction

La fouille de données d'opinions est un domaine de recherche en plein essor. Elle devient essentielle, par exemple pour le développement de tâches de veille (technologique, marketing, concurrentielle, sociétale) qui peuvent se révéler cruciales pour les entreprises et trouve de très nombreux domaines d'applications. Nous pouvons citer, par exemple, les clients qui souhaitent connaître comment évaluer un produit avant de l'acheter, l'image que les clients peuvent se faire d'une entreprise, la détection de rumeurs[1] sur le web. Cependant, les approches traditionnelles de fouilles de données ne sont plus adaptées à un contexte dans lequel il faut appréhender non seulement de gros volumes de données mais s'intéresser à la qualité des données : comment déterminer des avis négatifs ou positifs dans des documents aussi divers que des blogs ou des journaux ? Comment valider/évaluer les résultats obtenus ? Quel type de données à utiliser ? Une approche pluridisciplinaire regroupant différentes communautés (fouille de données, aide à la décision, modélisation des connaissances, TAL, Linguistique, etc.) parait aujourd'hui essentielle au développement rigoureux de cette thématique.

Notons, que l'un des objectifs de l'opinion-mining est la classification de textes en fonction des jugements favorables ou défavorables qu'ils expriment.

L'identification d'opinions est une tâche de recherche en opinion-mining qui sert à extraire à partir d'un ensemble de documents, les opinions exprimées par une source sur différents objets[2]. Malgré la confusion existante autour de cette tâche, l'identification d'opinions a attiré beaucoup d'attention ces dernières années, et les travaux de recherche existants, convergent plus ou moins vers le but souhaité.

Nous présentons dans ce chapitre les concepts de base en fouille de données d'opinions (opinion-mining) en donnant quelques définitions utiles, en présentant les travaux de recherche en identification et classification d'opinions, ainsi que les domaines de recherche.

[1] Le volume important des documents disponibles sur le net, forme un obstacle, dont l'utilisateur ne peut pas trouver les informations pertinentes et réelles sur le produit qu'il va acheter.

[2] Le terme objet a été utilisé, pour dénoter une entité commentée, qui peut être un individu, une organisation ou un événement.

1.2. Définitions

Dans ce paragraphe, nous présentons la terminologie de base utilisée en fouille de données d'opinions, qui apparait utile pour nos étapes ultérieures dans ce mémoire.

- **Opinion-mining**

D'après [Esula, 2008], l'opinion-mining est une discipline récente, résultat de l'intersection de recherche de l'information et la linguistique, elle n'est pas concernée par l'étude des thèmes d'un document, mais par l'opinion qu'il exprime.

- **Opinion**

Une opinion est un avis, un jugement personnel que l'on s'est forgé sur une question ou un sujet en discussion qui ne relève pas la croissance rationnelle. L'opinion, même s'il est affirmé avec conviction, est un jugement qui n'est pas nécessairement juste. [Site3, 2009].

Sentiment de celui qui opine sur quelque affaire mise en délibération. [Site4, 2009].

- **Sentiment**

Conscience plus ou moins claire, connaissance comportant des éléments affectifs et intuitif [Robert, 2001].

- **Objectivité**

Est objectif ce qui existe en soi, indépendamment du sujet pensant. Plus généralement, est objectif ce qui fait référence à la réalité extérieure, indépendante des consciences. [TLFI, 2009].

- **Subjectivité**

Est subjectif ce qui est propre à un sujet déterminé, qui ne vaut que pour lui seul (synonyme : individuel) ; ou encore ce qui ne correspond pas à une réalité, à un objet extérieur mais à une disposition particulière du sujet qui perçoit. [TLFI, 2009].

- **Orientation sémantique**

L'orientation sémantique d'un mot ou d'une expression est sa polarité positive, négative ou neutre.

Exemple :

Quelques mots positifs : bon, excellent, correct, encourageant…etc.

Quelques mots négatifs : mauvais, faux, désespérant…etc.

Quelques mots neutres : livre, stylo, enfant, vert…etc.

Notons, que ces mots peuvent êtres des adjectifs, des verbes, ou des adverbes…etc.

Philosophiquement, il est nécessaire de distinguer deux dimensions distinctes de l'opposition de l'objectif et du subjectif. Si l'on définit l'objectivité comme ce qui renvoie à une réalité subsistant en elle-même, une réalité indépendante de toute connaissance, donc de tout sujet, alors la subjectivité désignera au contraire tout ce qui est de l'ordre de l'idée, de la perception, etc. autrement dit tout ce qui appartient au domaine de l'expérience. Si, au contraire, on prend pour définition de la subjectivité ce qui ne vaut que pour un individu, alors l'objectivité désignera ce sur quoi tous les individus s'accordent. En ce sens, ce qui est objectif au second sens (car partagé par tous) sera néanmoins considéré comme subjectif dans le premier sens (car dépendant de la connaissance).

A titre d'exemple, la science moderne a pu se présenter comme une tentative pour rendre compte des phénomènes objectifs au sens d'indépendants de toute expérience. C'est ainsi qu'Ampère affirmait que « les lois mathématiques du mouvement des astres réglaient leur mouvement depuis que le monde existe et bien avant que Kepler ne les ait démontrées ». Concernant le subjectif au sens de ce qui est propre à un individu, il n'est pas difficile de trouver des exemples ; pensons par exemple à toutes les données individuelles auxquelles nous accédons par introspection. Prenons enfin un exemple ce qui peut être subjectif en un sens et objectif en un autre. En philosophie, les *qualités premières* (étendue, figure, solidité) sont celles dont on fait l'expérience directe et dont on peut affirmer qu'elles appartiennent à l'essence de la chose. Au contraire les *qualités secondes* dépendent étroitement de l'expérience que nous en faisons par l'entremise de nos sens, de nos modifications internes ; à ce titre on peut juste faire de l'objet la cause de ces qualités mais non pas dire que ces qualités appartiennent à son essence. Les qualités secondes (couleur, odeur, son, etc.) sont subjectives au sens où elles n'existent pas en dehors du sujet percevant mais elles peuvent également être objectives si l'on considère qu'elles ne diffèrent pas d'un individu à l'autre (ce qui bien évidemment n'est qu'une hypothèse).

On peut défendre l'idée que la science ne consiste pas tant dans une découverte et une explication de phénomènes objectifs ayant existés de tout temps que dans des *processus d'objectivation*. On affirme alors que la science est une construction humaine, cela ne menaçant en rien son objectivité. C'est ce qu'on appelle le *constructivisme*. [Site5, 2009].

1.3. Travaux de recherche

Parmi les travaux de recherche dans le domaine de fouille de données d'opinions (opinion-mining), nous citons :

- Dans [Hu et al., 2006], les auteurs ont utilisé le Web comme une ressource pour l'extraction des opinions disponibles sur les forums, les blogs, et les sites de e-commerce, et ont proposé d'étudier les problèmes suivants :

1. Identification des caractéristiques des produits sur lesquelles, les clients ont donné leurs opinions ;

2. Pour chaque caractéristique, identifier les phrases qui portent des opinions positives ou négatives ;

3. Construire un résumé sur le produit en utilisant les opinions extraites sur les caractéristiques.

L'exemple suivant a été extrait de leur article : le produit commenté par les clients est un «Appareil-photo», les caractéristiques du produit commentées sont « la qualité d'image » et « la taille ». Les opinions collectées, permettent de donner un résumé sur ce produit, illustré comme suit :

- pour la qualité d'image, le nombre d'opinions positives : 256, le nombre d'opinions négatives : 6.

- pour la taille, le nombre d'opinions positives : 134, le nombre d'opinions négatives : 10.

L'utilisation de tels résumés sur des produits différents, permettent de comparer les opinions des clients sur les produits concurrents où les clients peuvent voir visuellement les forces et les faiblesses de chaque caractéristique, en utilisant bien sûr des représentations graphiques[3].

➢ Dans [Ding et al., 2007], les auteurs ont utilisé le Web comme source d'opinions et ont travaillé sur les critiques des clients disponibles en ligne sur des produits commerciaux, ils ont posé le problème des mots d'opinions qui dépendent du contexte, par exemple le mot « *petit* » peut indiquer une opinion positive ou négative d'une caractéristique d'un produit, dépend de la caractéristique elle-même.

Pour remédier à ce problème, les auteurs ont proposé d'utiliser une fonction d'agrégation d'opinions : une phrase S peut contenir multiple caractéristiques[4] ($f_1,\dots f_m$) sur un produit et multiple mots d'opinions ($w_1,\dots w_n$).

L'objective est de déterminer l'orientation de l'opinion exprimée sur la caractéristique f_i dans S, i.e., le pair (f_i,S). Etant donnée une liste de mots d'opinions positives et négatives dépendant du contexte, y compris des phrases et idiomes. Le système travaille comme suit :

- Segmenter la phrase S, en utilisant des mots clés comme « mais », « à l'exception de… », …etc.

- Supposer que la caractéristique f_i est dans un segment S_k.

- Calculer le score de l'orientation d'opinion pour la caractéristique f_i, dans le segment S_k :

$$score(f_i,S_k) = \sum_{w_j \in S_k} \frac{w_j.SO}{d(w_j,f_i)}$$

Où : w_j, est un mot d'opinion dans le segment S_k, qui est le segment de phrase qui contient la caractéristique f_i, $d(w_j,f_i)$ est la distance entre le mot w_j et la caractéristique f_i. $w_j.SO$ est l'orientation sémantique du mot w_j.

[3] Pour chaque Objet, les auteurs ont utilisés un diagramme en bâtons, où l'axe des ordonnées représente les caractéristiques, et l'axe des abscisses représente la somme des opinions positives dans le sens positif, et la somme des opinions négatives dans le sens négatif.

[4] Nous désignons par *caractéristique*, un sous-composant de l'objet (*feature* en anglais).

Notons qu'à un mot positif est assigné une orientation sémantique de score est égale à 1, et à un mot négatif est assigné une orientation sémantique de score est égale à -1. Une simple sommation a été employée pour calculer le score final.

Si le score final est positif, alors l'opinion sur la caractéristique f_i dans la phrase *S* est positive. Si le score final est négatif, alors l'opinion sur la caractéristique f_i dans la phrase *S* est négative. Elle est neutre autrement.

Dans le même projet, les auteurs ont proposé une liste de règles linguistiques pour impliquer les opinions :

(1) Règle de conjonction intra-phrase

Par exemple, nous avons la phrase suivante :

« La durée de vie de la batterie est très *longue* ».

Il n'est pas clair que le mot « *longue* » exprime une opinion positive ou négative. L'algorithme proposé dans ce contexte essaye de déterminer si « *longue* » exprime une opinion positive ou négative à partir des critiques des clients. Par exemple dans d'autres critiques, on trouve « cet appareil-photo prend de *belles* photos et la batterie à une durée de vie *longue* ». A partir de cette phrase, nous pouvons découvrir que « *longue* » est positif pour « la durée de vie de la batterie » parce qu'il conjointe avec le mot positif « *belle* ».

Cette règle est appelée *règle de conjonction*, qui signifie qu'une phrase exprime seulement l'orientation d'une opinion sans la présence des mots comme « mais » qui changent la direction. La phrase suivante est peu probable : « cet appareil-photo prend de *belles* photos et la durée de vie de la batterie est *courte* ».

(2) Pseudo règle de conjonction d'intra-phrase

Parfois, on ne peut pas employer une conjonction explicite « et ». Employons la phrase suivante :

« La durée de vie de la batterie est *longue* ».

Nous n'avons aucune idée si « *longue* » est positif ou négatif pour « la durée de vie de la batterie ». Une stratégie semblable peut être appliquée. Par exemple, dans une autre critique on peut avoir:

« La durée de vie de la batterie est *longue*, c'est *magnifique* ».

La phrase indique que l'orientation sémantique de « *longue* » pour la « durée de vie de la batterie » est positive dû à « *magnifique* », sans explicitement utiliser « et ».

(3) Règle de conjonction d'inter-phrases

La règle de conjonction peut également être généralisée sur des phrases voisines. L'idée est qu'il est possible d'exprimer une opinion dans un ensemble de phrases consécutives. Des changements d'opinions sont indiqués par des mots comme « *mais* », « *cependant* »…etc. par exemple, les passages suivants sont normaux : « La qualité d'image est *bonne*. La durée de vie de la batterie est *longue* » et « la qualité d'image est *bonne*. Cependant, la durée de vie de la batterie est *courte* ». Cependant, le passage suivant n'est pas normal : « La qualité d'image est *bonne*. La durée de vie de la batterie est *courte* ». Bien que nous ne sachons pas si « *long* » (ou « *court* ») est positif ou négatif pour la « durée de vie de la batterie », si nous savons que « *bonne* » est positif puis nous pouvons impliquer que « *longue* » est positif et « *court* » est négatif pour la « durée de vie de la batterie».

(4) Règle des synonymes et antonymes

Si un mot s'avère positif (ou négatif) dans un contexte pour une caractéristique, ses synonymes sont également considérés positifs (ou négatifs), et ses antonymes sont considérés négatifs (ou positifs).

Par exemple, dans l'exemple de la règle précédente, nous savons que « *longue* » est positif pour la « durée de vie de la batterie ». Alors nous savons également que « *courte* » est négatif pour la « durée de vie de la batterie ».

1.4. Domaines de recherche

Il y'a trois domaines de recherches principaux [Esula, 2006]:

- Développement des ressources linguistiques pour l'opinion-mining, par exemple, construction automatique d'un lexique des termes subjectifs : le travail de recherche dans ce domaine dépend sur trois tâches principales :

  - Détermination de l'orientation des termes : déterminer si un terme subjectif a une orientation positive ou négative.

    Hypothèse 1

    Les adjectifs coordonnés par « *et* » ont une orientation similaire, et les adjectifs employés avec « *mais* » ont une orientation opposée [Esuli, 2006].

    Hypothèse 2

    Les termes ont une orientation semblable tendent d'êtres co-occurrents dans le document [Esuli, 2003].

  - Détermination de la nature des termes: Déterminer si un terme exprime la subjectivité (subjectif) ou pas (objectif).

Dans [Esuli, 2006], l'auteur a schématisé l'orientation sémantique d'un terme comme suit :

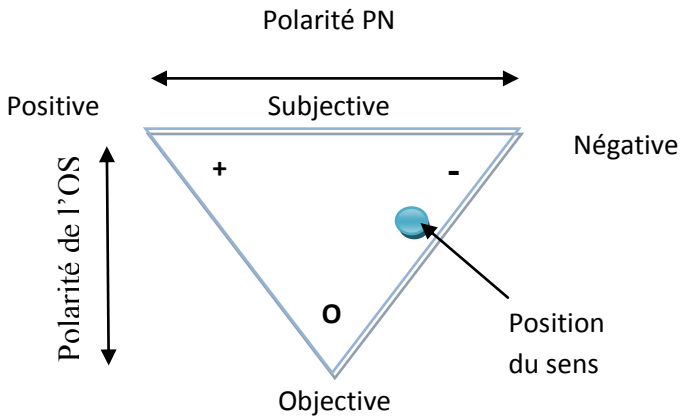

Figure 1 : Représentation de l'orientation sémantique d'un terme

- Détermination de la force des termes, comme en attribuant aux termes des degrés (à valeurs réelles) de positivité ou de négativité.

Exemple :

Bon, excellent, meilleur → termes positifs ;

Mauvais, faux, le plus mauvais → termes négatifs ;

Vertical, jaune, liquide → termes objectifs.

NB :

- Quelques termes posent le problème d'ambigüité entre le sens objectif et le sens positif, exemple : le terme « *estimable* » ambiguë avec le sens objectif (i.e. *mesurable*), et le sens positif (*mérite le respect*).

- Il faut prendre en considération, les expressions multi-termes, exemple : *pas entièrement satisfaisant* → expression négative.

➢ Classification des textes (documents, phrases entières) par leurs contenus d'opinions, par exemple, classifier les avis sur un film en positifs et en négatifs.

➢ Extraction d'expressions d'opinions à partir des textes, pour l'identification d'opinions.

Hypothèse 3

L'orientation sémantique de tout le document est égale à la somme des orientations sémantiques de toutes ces parties [Esuli, 2006].

L'orientation sémantique d'une opinion est calculée comme la moyenne des orientations sémantiques des adjectifs et des verbes contenus dans l'opinion.

1.5. Conclusion

Dans ce chapitre, nous avons présentés les notions fondamentales en opinion-mining, en vue de simplifier la compréhension des concepts de base qui ont généralement une orientation philosophique que linguistique ou informatique.

L'identification d'opinions, cette récente sous-discipline a connu beaucoup d'applications importantes : extraction et classification des critiques des clients sur un produit commercialisé en ligne, suivre les opinions du général public sur un candidat politique par la fouille en ligne dans les forums et les blogs…etc.

En identification d'opinions, plusieurs sous-tâches sont imposées come la détermination de la subjectivité dans un texte, détermination si un texte ou un mot subjectif a une orientation positive, négative ou neutre.

CHAPITRE 2
Prétraitement de textes arabes

2.1. Introduction

Afin de pouvoir construire un processus de traitement de textes adapté à tous les cas possibles, il est nécessaire de prétraiter les données brutes pour pouvoir extraire les informations pertinentes.

Dans notre thème, nous nous intéressons aux données textuelles de la langue arabe, cette dernière est caractérisée par sa richesse linguistique qui nécessite un prétraitement particulier vis-à-vis de la complexité morphologique et syntaxique qui constitue des ambiguïtés sémantiques, lexicales et structurelles, d'où un prétraitement profond.

Ces prétraitements peuvent êtres effectués avec des techniques d'analyse linguistique plus ou moins poussées et des coûts en temps et en ressources très variables. Elles consistent à normaliser les diverses manières d'écrire un même mot, à corriger les fautes d'orthographe évidentes ou les incohérences et expliciter les ambiguïtés.

Les incohérences et les ambiguïtés sont deux grands types de difficultés à traiter dans les textes bruts: Celles-ci doivent être traitées avant toute extraction d'informations. Les incohérences doivent être corrigées afin de permettre un traitement unifié lors des étapes ultérieures d'extraction d'informations.

Le format variable d'encodage des textes, par exemple des caractères encodés de manières différentes d'un texte à l'autre, constitue un premier type d'incohérence. Ce type d'incohérence est à traiter en premier lieu car le problème se situe au niveau du caractère et il faut décider d'un encodage unique avant de pouvoir effectuer des traitements au niveau des mots.

La présence des fautes d'orthographe constitue un autre type d'incohérence. Ce traitement devrait être effectué avant celui des ambiguïtés lexicales pour limiter les erreurs dues aux données erronées.

Les ambiguïtés lexicales nécessitant parfois des ressources externes aux textes afin de les lever, par exemple un lexique d'abréviations, constitue un premier type d'ambiguïté. Ce traitement devrait être effectué avant celui de la structure des textes pour ne plus avoir le problème des abréviations.

2.2. Problèmes du traitement automatique de la langue arabe

Un des aspects complexes de la langue arabe est l'absence des voyelles dans le texte, qui risque de générer certaines ambigüités à deux niveaux [Douzidia, 2004] : sémantique et syntaxique.

Nous citons dans cette section quelques problèmes rencontrés lors du traitement automatique de la langue arabe :

2.2.1. La vocalisation

Dans les textes arabes, comme c'est le cas en Hébreu et dans d'autres langues sémitiques dont le système graphique est issu de l'alphabet phénicien, un nombre important de signes ne sont pas notés. Il s'agit pour l'essentiel, en arabe, des voyelles brèves, de la gémination des consonnes et de certaines marques casuelles et de détermination. On le voit, ces signes peuvent être graphémiques (voyelles brèves, consonnes géminées). Lorsqu'ils sont notés (par exemple dans les éditions du texte coranique ou dans les éditions d'apparat de textes religieux ou de poésie ancienne), ils sont réalisés sous la forme de signes secondaires [Abbès, 2004].

En arabe nous comptons six voyelles, et une voyelle muette.

Signe	**Désignation**
◌َ	Fatha
◌ِ	Kasra
◌ُ	Dama
◌ً	Double Fatha
◌ٍ	Double Kasra
◌ٌ	Double Dama
◌ْ	Muette

Tableau 2.1 : Tableau de voyelles en arabe

Plusieurs cas de déclinaison ou de détermination sont illustrés par une simple voyelle. En l'absence de ce signe diacritique, nous perdons l'information qu'il véhicule. Voici à titre d'exemple le modèle de déclinaison du nom « نملة » (fourmi), où toutes les déclinaisons du nom sont déterminées par une voyelle.

Mode/Cas	**Nominatif**	**Accusatif**	**Génitif**
Indéterminé	نَمْلَةٌ	نَمْلَةً	نَمْلَةٍ
Déterminé par annexion	نَمْلَةُ	نَمْلَةَ	نَمْلَةِ
Déterminé par l'article	اَلنَّمْلَةُ	النَّمْلَةَ	اَلنَّمْلَةِ

Tableau 2.2 : Tableau de déclinaison du nom نملة

Les signes diacritiques sont souvent des déterminants pour le sens du mot. En dehors des problèmes de découpage. En voici quelques exemples, dans les deux cas nous ne noterons pas la voyelle finale pour regrouper tous les cas de déclinaisons. Le premier exemple donne deux sens différents mais les mots ont la même racine, le second cas donne deux mots qui n'ont pas la même vocalisation :

Exemple :

Soit le mot non vocalisé,الحزب selon la vocalisation il peut être :

- اَلْحُزْب, le nom verbal du verbe حَزَبَ – يَحْزُبُ (racine حزب) donnant le sens de « la situation difficile, contraignante, problématique »;

- اَلْحِزْب, le nom signifiant «parti» (racine حزب)

Toute correction ou harmonisation doit passer par une vocalisation partielle ou totale des textes, sachant que ceci ne résoudra pas l'intégralité des problèmes mais il les réduit à une complexité abordable par un éventuel outil automatique.

La vocalisation ne fait pas partie des mœurs d'écriture chez les arabes. En dehors du texte coranique et des textes d'apprentissage élémentaires de la langue, rare sont les textes comportant des voyelles. Aujourd'hui encore il n'y a aucune raison de s'attendre à des textes dactylographiés vocalisés, d'autant que la saisie sur le clavier reste moins évidente que l'écriture.

2.2.2. La Chadda

Il existe aussi un signe particulier associé aux « voyelles », c'est la Chadda ّ, elle est aussi placée au dessus des consonnes sans être une simple marque vocalique. La Chadda illustre souvent le doublement d'une consonne, sa présence et du même ordre que celle des consonnes. Seulement, dans la tradition d'écriture chez de nombreux arabes, elle est omise au même titre que les autres signes-voyelles.

L'absence de la Chadda pose un problème totalement différent de celui des voyelles. Spéculer sur sa présence revient souvent à vérifier s'il ne manque pas une lettre au mot graphique. La Chadda comme signe de vocalisation peut augmenter le taux d'ambigüité.

2.2.3. La confusion dans l'écriture de certaines lettres

Avec la Chadda, nous avons soulevés le problème qu'engendre l'absence du signalement graphique d'une consonne. Dans ce paragraphe nous allons présenter un autre problème touchant les consonnes, la confusion dans l'écriture de certaines lettres.

Les cas que nous allons citer ne font pas office d'exception. Leurs fréquences d'utilisation les mettent dans le rang de la pratique d'écriture répandue chez les écrivains arabes.

Les confusions d'écriture des consommes peuvent poser des incohérences au niveau de texte, qui pose un traitement supplémentaire. Exemples :

- **La Hamza et le Alif**

Les textes arabes confondent les lettres أ et إ au début et au milieu des mots. Ils les notent indifféremment en tant que ا (*alif*). Ce qui en plus d'être une erreur d'orthographe présente une grande source d'ambigüité.

Exemple :

- سال signifie « il s'est écoulé » « il a coulé », du verbe يسيل – سال (racine سيل).

- سأل signifie « il a questionné » ou « il a posé une question », du verbe يسأل -سأل (racine سأل).

Beaucoup de mots minimaux ont un homographe avec et vice-versa, et ils ne sont pas toujours de la même famille. Exemple :

- Le nom أمارة (*amarat*) signifie « un marque de… » ou « signifie de »

- Le nom إمارة (*imarat*) signifie «Emirat »

De surcroît, avec une voyelle au-dessus ou au-dessous, le Alif peut faire partie du mot. Comme dans le déverbal انحراف (*inhiraf*). « Déviation » ou dans la particule de détermination ال (*al*). Exemple :

- قران «liée deux… » ou « mariage ».

- قرآن « le Coran ».

- **Le Ya et le Alif Maqsûra**

Dans l'usage typographique égyptien par exemple, les auteurs notent la lettre ي à la fin des mots sans point dessous, ce qui la rend équivalente à ى Alif maqsûra. L'absence des deux points change totalement le mot, et pose un vrai problème de reconnaissance de la forme écrite. A l'image du premier point, la plus part des mots se terminent avec Alif Maqsûra ont un homographe avec le Ya. Exemple :

- Le mot outils على.

- Le nom propre علي.

Remarque : certains mots peuvent cumuler les deux confusions, Hamza avec Alif d'un côté et Ya avec Alif Maqsûra de l'autre. Exemple : الاولى (*alawla*).

2.2.4. Le Tanwin

Le Tanwin pose la difficulté de variation de sa position au cas direct à la fin des mots. Les terminaisons ـًا, sont très souvent notées ـاً.

Il s'agit d'une variation typographique, empruntée à l'usage des calligraphes qui peuvent noter le signe du Tanwin «ً » avant, au dessus ou après le alif « ا ».

2.2.5. Le caractère '-'

Les typographes font un usage fréquent du caractère '-' (appelé Kashida), qui permet l'allongement du trait au milieu des mots, pour une meilleure lisibilité, pour limiter les espaces blancs sur une ligne justifiée, voire pour des raisons purement esthétiques. Or cet usage peut nuire aux analyses automatiques : ce caractère ne fait pas partie de l'alphabet arabe, il est considéré comme un intrus par le système d'analyse automatique. Il faut donc recourir à un sous-programme particulier afin de l'éliminer. Exemple : le mot الكتاب : peut être écrit de plusieurs façons : الـكتاب, الكتـاب, الــكتاب , …etc. [Abbès, 2004]

2.2.6. Mots étrangers translittérés en arabe

Les translittérations en arabe de mots étrangers posent un problème, puisqu'ils n'ont pas de racine en arabe. Les mots translittérés sont considérés comme inconnus par l'analyseur.

Quelques items étrangers méritent une attention particulière en raison de leurs fréquences élevées. Exemple: أورو, دولار …etc.

.

2.2.7. L'ambigüité

Les mots peuvent être ambigus aux niveaux lexical et grammatical. Le mot « ذهب » est ambigu lexicalement. Il peut désigner l'or en français ou encore le verbe aller. «كاتب », quant à lui, est ambigu grammaticalement. Il peut appartenir à plusieurs catégories grammaticales différentes; فعل, اسم, اسم فاعل Le sens de ce mot sera très différent selon sa catégorie, nom : « écrivain » ou verbe : « écrit » ; il peut appartenir à quatre catégories grammaticales différentes. Le nombre de catégories auxquelles un mot peut appartenir, dépend du jeu d'étiquettes choisies. En moyenne, le nombre d'étiquettes par unité lexicale voyellée est de 9, alors que pour les unités lexicales non voyellées, elle peut atteindre 12 étiquettes par unité lexicale. Le mot أَقَالَ signifie « est-ce qu'il a dit ? », « est-ce qu'il a fait la sieste ? » et « il a démis quelqu'un (de ses fonctions) ». A chaque cas correspond une racine différente. Un simple listing peut occulter un ou plusieurs de ces sens.

Un lexique est indispensable à tout étiqueteur de mot arabe qui doit attester à la fois de la bonne orthographe du mot et aussi de son appartenance à la langue.

L'ambigüité lexicale est due au fait que le dictionnaire permet d'attribuer plusieurs valeurs d'étiquettes pour une même entrée lexicale.

Le taux d'ambigüité évolue dans le même sens que la surface de la langue couverte par le lexique. Plus le lexique est riche plus il génère de possibilités d'analyse. Le nombre d'entrées et le nombre d'informations morpho-syntaxiques attachées aux lexèmes influent directement sur le taux des ambigüités lexicales.

Pour une langue agglutinante, aussi flexionnelle que l'arabe et de surcroit non vocalisé, nous devons nous attendre à plusieurs situations où les mêmes formes graphiques n'ont pas la même analyse morpho-syntaxique [Abbès, 2004].

2.3. L'analyse morphologique en arabe

L'analyse morphologique en arabe s'intéresse, comme les autres langues, aux formats du mot. Mais étant donnée la richesse du mot graphique, l'opération englobe rapidement des aspects formels de la langue, reléguant les traits sémantiques et pragmatiques en arrière plan.

Dans cette partie nous essayerons de présenter les difficultés posées lors d'une analyse morphologique ou morpho-syntaxique en détaillant les composants du mot graphique en arabe.

2.3.1. Difficulté de l'analyse morphologique de l'arabe

En analyse morphologique, le principal problème à résoudre est l'ambigüité. Il existe différents types d'ambigüités. D'abord, les mots peuvent être ambigus aux niveaux lexical et grammatical. Le mot « ذهب » est ambigu lexicalement. Il peut désigner l'or en français ou encore le verbe aller. « كاتب », quant à lui, est ambigu grammaticalement. Il peut appartenir à plusieurs catégories grammaticales différentes : اسم فاعل, فعل, اسم. Le sens de ce mot sera très différent selon sa catégorie, nom : « écrivain » ou verbe « écrit » ; il peut appartenir à plusieurs catégories grammaticales différentes.

Le nombre de catégories auxquelles un mot peut appartenir dépend du jeu d'étiquettes choisies. En moyenne, le nombre d'étiquettes par unité lexicale voyellée est de 9, alors que pour les unités lexicales non voyellées, elles peuvent atteindre 12 étiquettes par unité lexicale. Un autre facteur pour la langue arabe, est le nombre d'unités lexicales ambigües dans un texte. Contrairement au français et à l'anglais, il peut concerner 66% des unités lexicales qui composent le texte. Ces difficultés peuvent poser plusieurs ambigüités au niveau sémantique et syntaxique [Abbès, 2004].

2.3.1.1. Ambiguïté dérivationnelle et flexionnelle

La flexion est la variation de la forme des mots en fonction de facteurs grammaticaux telle que la conjugaison pour les verbes (exemple : le mot يتأثرون (ils s'influencent) est le résultat de la concaténation du préfixe « ي » indiquant le présent et du suffixe «ون» indiquant le masculin pluriel du verbe « تأثر »).

Le problème en analyse morphologique de l'arabe se rapporte surtout au niveau de la dérivation qui est un phénomène plus complexe que la flexion. En effet, la dérivation est la formation de nouveaux mots à partir de mots existants. Dans le cas de la langue arabe, la plupart des mots sont dérivés à partir de racines trilitères ou quadrilitères. Le mot arabe n'est pas le résultat d'une simple concaténation de morphèmes comme c'est le cas en anglais (exemple : unfailingly = un+fail+ing+ly), mais c'est à partir d'une racine, d'une combinaison de voyelles, de préfixes, d'infixes, de suffixes et d'un schème morphologique qu'on obtient un mot (exemple : à partir de la racine «أثر », on peut dériver plusieurs verbes tel que «تأثر » (s'influencer) et plusieurs noms tel que « متأثر » (ému)).

2.3.1.2. Ambiguïté d'agglutination

Contrairement aux langues latines, en arabe, les articles, les prépositions, les noms, etc. collent aux adjectifs, noms, verbes et particules auxquels ils se rapportent. Comparé au français, un mot arabe peut parfois correspondre à une phrase française, (exemple : le mot en arabe «أتتذكرونا » correspond en français à la phrase «Est-ce que vous vous souvenez de nous). Cette caractéristique engendre une ambiguïté morphologique au cours de l'analyse. En effet il n'est pas toujours facile de distinguer un proclitique ou un enclitique d'un caractère original du mot. Par exemple le mot « وفتح » (et il a ouvert), il s'agit plutôt d'une proclitique.

2.3.1.3. Ambiguïté due à la non voyellation

La morphologie arabe est assez régulière lorsque les mots sont présentés sous leurs formes non voyellées. Cependant, la majorité des documents arabes sont non voyellés sauf pour le Coran et pour certains ouvrages scolaires pour les débutants. En fait, les mots non voyellés engendrent beaucoup de cas ambigus au cours de l'analyse (exemple : le mot non voyellés «فصل » pris hors contexte peut être un verbe au passé conjugué à la troisième personne du singulier « فَصَّل » (il a licencié), ou un nom masculin singulier « فَصْلٌ » (chapitre/saison), ou encore une concaténation de la conjonction de coordination «فَ » (puis) avec le verbe « صل » : impératif du verbe lié conjugué à la deuxième personne du singulier masculin).

L'absence des voyelles dans le texte représente un des aspects complexes de la langue arabe est, qui risque de générer certaines ambiguïtés à trois niveaux :

- Ambiguïté sémantique : difficulté à identifier le sens du mot :

Nous citons à titre d'exemple l'énoncé suivant :

زهير أشعر أهل الجاهلية

L'ambiguïté vient du mot أشعر qui signifie *informer (avertir…)* ou *le meilleur poète* alors que voyellé on aura أَشْعَرَ pour *informer,* et أَشْعَرُ pour *le meilleur poète.*

Un texte arabe non voyellé, est fortement ambigu.

Cette ambiguïté pourrait, dans certains cas, être levée soit par une analyse plus profonde de la phrase ou des statistiques (par exemple il est plus probable d'avoir زهير أشعرُ أهل الجاهلية (*Zuhayr est le meilleur poète* que *Zuhayr a informé*…).

- Ambiguïté syntaxique : difficulté à identifier sa fonction dans la phrase.

A titre d'exemple, soit l'énoncé suivant :

وضع الولد صورة (العصفور فوق الكتاب)

- Ambiguïté structurelle : La phrase arabe peut-être :

- Soit simple se limitant à contenir un verbe et un sujet, exemple : أكتبُ (j'écris) au quelle, il peut-être ajouté un ou plusieurs compléments d'objet si le verbe est transitif.

- Soit complexe permettant de relier entre deux ou plusieurs membres ou propositions par des conjonctions ou par le sens, nous citons à titre d'exemple l'énoncé suivant :

أخذ زهير ابن أبي سلمى بثأر أبيه الذي قتل, رجل من بني أسد قتله

Ceci peut influencer les fréquences des mots, étant donné qu'elles sont calculées après la détection de la racine ou la lemmatisation des mots, qui est basée sur la suppression de préfixes et suffixes. Lors du calcul des scores à partir des titres, il peut arriver que des mots soient considérés comme dérivants d'un même concept alors qu'ils ne le sont pas.

De plus la capitalisation n'est pas employée dans l'arabe ce qui rend l'identification des noms propres, des acronymes, et des abréviations encore plus difficile. Comme la ponctuation est rarement utilisée en arabe, nous devons ajouter une phase de segmentation de texte pour l'analyse d'un texte que nous allons détailler dans le paragraphe 2.3.4.1.1.

2.3.2. Le modèle du mot dans les prétraitements en arabe

Un mot maximal est composé d'une base, pré-base et post-base. La pré-base est elle-même composée d'un proclitique et d'un préfixe. La post-base est composée d'un enclitique et d'un suffixe. A partir de la base ; nous pouvons extraire la racine et le schème. La figure suivante illustre ce modèle :

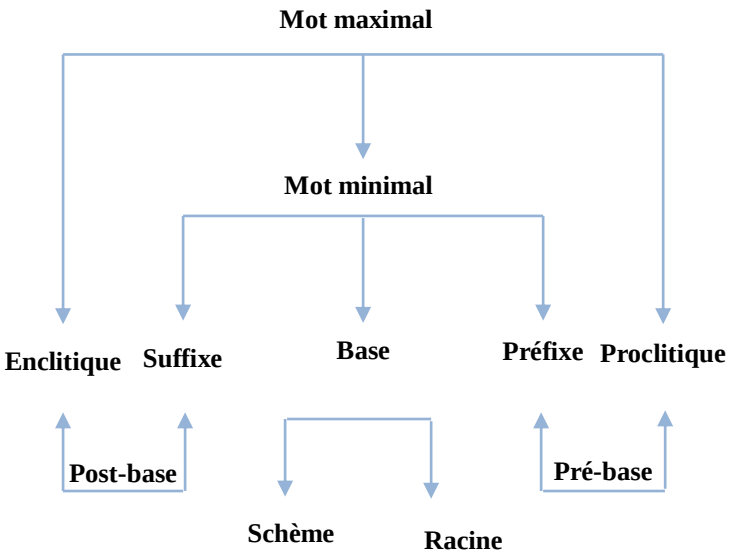

Figure 2.1 : Modèle du mot graphique en arabe

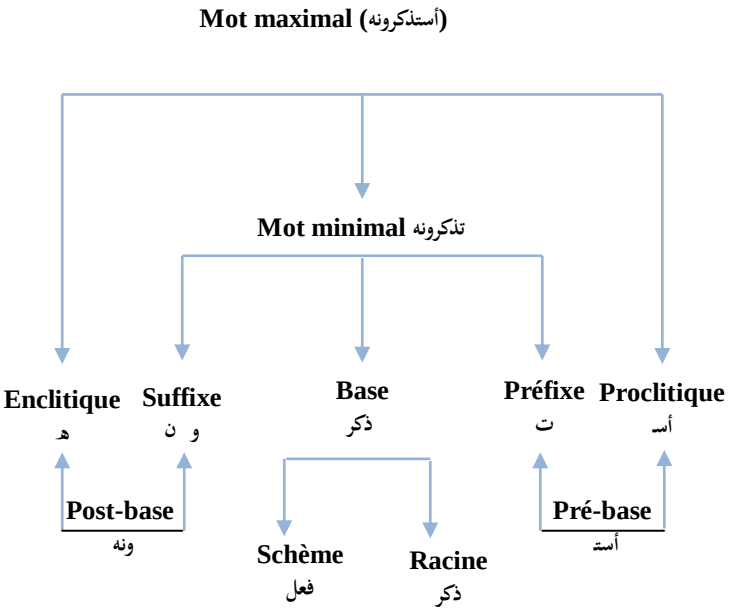

Figure 2.2 : Modèle du mot graphique en arabe appliqué pour le mot « أستذكرونه **»**

La base, pour la partie du lexique qui relève du système dérivationnel propre aux langues sémitiques de la même famille que l'arabe, s'analyse en une racine et un schème. On notera toutefois, qu'un sous-ensemble important des noms ne peut être analysé ainsi. Ces noms correspondent à des pro-bases.

Exemple :

ياسمين « *Jasmin* », إبراهيم « *Abraham* »,...etc.

Bases et pro-bases sont le noyau lexical du mot graphique (ou format noyau), les autres constituants étant des extensions (ou formats-extensions).

La structure de l'unité lexicale en arabe peut se présumer ainsi :

- Le mot est une unité linguistique dont la manifestation la plus aisément observable est le mot graphique.

- Les morphèmes constitutifs de l'unité-mot sont appelés des formants de mots, c'est-à-dire, des signes linguistiques minimaux dont les relations de contextualisation sont limitées aux autres morphèmes inclus dans l'unité composée que constitue le mot dans sa manifestation graphique.

2.3.3. Le lexique utilisé en analyse morphologique

Un lexique pour le traitement automatique de l'arabe doit contenir les informations nécessaires pour toute entreprise du domaine. Nous pouvons diviser les données en deux catégories principales, d'un point de vue de traitement automatique. D'une part, celles nécessaires au découpage du mot, pour identifier ses différents composants décrits dans la

structure du mot graphique, la base et les particules. D'autre part, les traits morpho-syntaxiques attachés aux différents composants du mot.

2.3.3.1. Les particules

La procédure d'analyse est basée sur un découpage en pré-base, base et post-base. Les pré- et post-bases sont stockés séparément de leurs vocalisations, pour traiter les mots indépendamment de la vocalisation. Les affixes et les clitiques sont de simples listes de mots, référencées par des numéros.

A chaque pro-base nous associons le numéro de l'affixe et de l'enclitique qui la compose. A chaque base nous associons la liste des affixes et de pré- et post-bases qu'elle peut prendre.

2.3.3.1.1. Les préfixes

Les préfixes ne sont utilisés qu'à l'inaccompli. Ils sont en inventaire fini et ne se combinent pas entre eux. En dehors de certaines règles d'écriture, la hamza en l'occurrence, les préfixes ne suscitent pas d'information sur les types de verbes.

Les préfixes de l'arabe sont :

N° Préfixe	Préfixe
1	
2	أَ
3	أُ
4	تَ
5	تُ
6	تِ
7	تْ
8	يَ
9	يُ

Tableau 2.3 : Liste des préfixes de l'arabe

2.3.3.1.2. Les suffixes

Ils sont en inventaire fini et ne se combinent pas entre eux. Contrairement au préfixes, les suffixes peuvent s'ajouter aux verbes et aux noms y compris les déverbaux. Tous, ces ensembles ne sont pas entièrement distincts, mais comporte une intersection.

Dans le cas des verbes, les suffixes ne dépendent pas uniquement de l'aspect et du pronom mais aussi du type de verbe. Les verbes sont regroupés dans des familles suivant chacune un modèle de conjugaison. Du côté des déverbaux les suffixes dépendent du mode et du cas de déclinaison. Certains suffixes, exclusivement nominaux, ne prennent pas d'enclitiques, les

autres prennent les enclitiques possibles avec les mots ils sont attachés. Voici la liste exhaustive de tous les suffixes dans la langue arabe :

N°	Suffixe
1	
2	ً
3	ًا
4	ٌ
5	ٍ
6	َ
7	َا
8	َاتُ
9	َاتٍ
10	َاتُ
11	َاتِ
12	َانِ
13	َانْ
14	َةً
15	َةُ
16	َةٍ
17	َةَ

N°	Suffixe
18	َةُ
19	َةِ
20	َتَا
21	َتَانِ
22	َتَيِّ
23	َتَيْنِ
24	َتْ
25	َنَّ
26	َنْ
27	َوْا
28	َوْنَ
29	َيْ
30	َيْنَ
31	َيْنِ
32	ُ
33	ُنَّ
34	ُنْ

N°	Suffixe
35	ُو
36	ُوا
37	ُونَ
38	ِ
39	ِنَّ
40	ِنْ
41	ِي
42	ِينَ
43	ْ
44	ْتُ
45	ْتُ
46	ْتُمَا
47	ْتُمْ
48	ْتُنَّ
49	ْتِ
50	ْنَ
51	ْنَا

N°	Suffixe
52	ْنَانِّ
53	تَ
54	تُ
55	تُمَا
56	تُمْ
57	تُنَّ
58	تِ
59	نَ
60	نَا
61	نَانِّ
62	ِيًّا
63	ِيَّ
64	ِيِّ
65	ِيَّ
66	ِيَّ
67	ِيِّ

Tableau 2.4 : Liste des suffixes de l'arabe [Abbès, 2004]

2.3.3.1.3. Les proclitiques

Les proclitiques sont en inventaire fini, et se combinent antre eux pour donner les traits syntaxiques, coordonnant, déterminant…qui peuvent accompagner le mot arabe.

Dans le cas des verbes, les proclitiques dépendent exclusivement de l'aspect verbal. Ils prennent donc tous les pronoms et par conséquent ils sont compatibles avec tous les préfixes pris par l'aspect. Dans le cas des noms et les déverbaux, le proclitique dépend du mode et du cas de déclinaison.

Voici quelques proclitiques simples :

- La coordination par les coordonnants ف *fa* et و *wa*
- L'interrogation par le morphème أ *a*.
- La marque du futur س *sa*.

- L'article ٱلْ *al.*
- Les prépositions par lettres بِ *bi* et لِ *li*.
- Les particules du subjonctifs فَ *fa*, لِ *li* et وَ *wa*.
- Le marqueur de coordination لَ *la*…
- Les particules du jussif جزم par la lettre لِ *li*.

2.3.3.1.4. Les enclitiques

Ils sont en inventaire fini et peuvent se combiner entre eux, toutefois avec certaines restrictions sur les pronoms. Le lien entre les verbes et les enclitiques dépend du caractère de transitivité du verbe et du pronom. Les verbes au passif ne prennent aucun enclitique. Notons que certains suffixes ne peuvent être suivis d'enclitiques en aucun cas (généralement ceux qui se terminent par une voyelle double).

Voici dans le tableau suivant un récapitulatif de la liste des enclitiques retenue, leurs catégorisations et des informations utiles pour les compatibilités.

Remarque : De facto les verbes transitifs ne prennent pas d'enclitiques.

Enclitique	**Description**
ني	1ère personne masculin féminin ou singulier.
ان	1ère personne masculin féminin, duel et pluriel
كَ	2ème personne masculin, singulier.
كِ	2ème personne féminin, singulier.
كما	2ère personne masculin féminin, duel
كم	2ème personne masculin, pluriel
كن	2ème personne féminin, pluriel
هُ	3ème personne masculin, singulier
اه	3ème personne féminin, singulier
هما	3ème personne masculin, féminin, duel
هم	3ème personne masculin, pluriel
هن	3ème personne, féminin, pluriel
هِ	3ème personne masculin, singulier

Tableau 2.5 : Liste des enclitiques [Abbès, 2004]

2.3.3.1.5. Les pré-bases

Les pré-bases résultent de la combinaison entre proclitique(s) et préfixe. La génération des pré-bases se fait d'une manière semi-automatique. A l'aide des requêtes nous

sélectionnons les préfixes et les proclitiques ayant les mêmes aspects et les pronoms dans le cas des verbes. Quant aux noms, ils ne prennent pas de préfixes, donc toutes les pré-bases sont formées uniquement de proclitiques. Nous ajoutons aussi la liste dans le cas où le verbe ne prendrait pas de proclitique.

2.3.3.1.6. Les post-bases

Les post-bases sont obtenues par la combinaison entre suffixe et enclitique(s). Les compatibilités dépendent des pronoms décrits par chacune des particules.

- Les suffixes de la première personne prennent les enclitiques de la deuxième et de la troisième personne.

- Les suffixes de la deuxième personne prennent les enclitiques de la première et de la deuxième personne.

- Enfin, les suffixes de la troisième personne prennent indifféremment des enclitiques des trois personnes.

Certains suffixes, nominaux principalement, jouent un rôle terminal dans le mot arabe, ils ne peuvent pas prendre d'enclitiques.

2.3.3.1.7. La particule vide

La présence de toutes les particules dans le mot arabe n'est pas obligatoire. Dans certains cas elles ne peuvent pas être concaténées au mot. Le verbe à l'accompli ne prend pas de préfixes, les mots se terminent par un Tanwin ne prennent pas d'enclitiques…

La particule vide rentre dans la composition des pré- et post-bases. Le préfixe solitaire, par exemple, formera une pré-base construite par un proclitique vide et un préfixe. La pro-base vide est la concaténation des particules vides.

2.3.3.2. Les lexèmes

Par élimination, après le travail fait sur les particules, la base s'impose comme le meilleur lexème pour une application du traitement automatique de l'arabe. La base est une unité minimale dans le mot graphique, privé de toute marque flexionnelle.

La base est associée aux affixes et autres informations utiles pour fabriquer les modèles de conjugaisons des verbes et les modèles de déclinaisons des noms et des déverbaux. Par l'intermédiaire de relations plus complexes nous retrouvons les racines, les schèmes, les nombres, les modes, les cas…la transitivité.

L'unité lexicale (entrée du lexique) sera formée par la <base> ou <base+formant extension lexical> selon les cas.

Il y a deux types de variation morphologique par exemple : Pluriel-singulier pour les noms ; Accompli-inaccompli pour les verbes ; Verbe-forme infinitive (masdar).

- Par suffixation et préfixation (pour les verbes) ou par suffixation (pour les noms) :مُعَلّم («instituteur » masculin-singulier) ;

  مُعَلّمُونَ = وْنَ + مُعَلّم (masculin pluriel).

  مُعَلّمَات = اَت + مُعَلّم (féminin pluriel).

- Par dérivation interne (changement de schème, la racine reste constante) exemple : مَعْمَل (« atelier, usine », singulier) مَعَامِل (Pluriel). Dans ce cas on parle de fléchage. Le fléchage peut avoir lieu entre une <base+formant extension lexicalisé> et une autre base. Le fléchage peut avoir lieu entre une <base+formant extension lexicalisée> et une autre base, exemple :

  لَوْلَيّ <base + formant لولب extension lexicalisé يّ >, (pluriel لوالب).

Ainsi le lexique est greffé à la méthode d'analyse. L'avancement dans les profondeurs du lexique est déterminé par la complexité du mot à découper et le degré de l'analyse désiré. Avec cette organisation, le parcours est abandonné à la première discordance avec la source (en l'occurrence le mot en entrée) évitant ainsi les parcours inutiles et générant un énorme gain en temps de traitement et surtout, éviter la génération d'analyse inexistante dans la langue.

2.3.3.3. Les mots outils

En raison de leurs nombres réduits et de la grande fréquence d'utilisation de ces mots, nous les stockons dans une liste fermée.

Toutefois certains mots de la langue peuvent être confondus avec des mots outils. Exemple : أم peut signifier le mot outil « ou » ou « une mère ».

2.3.4. Désambigüisation

De nombreuses recherches ont été entreprises dans le but de résoudre les problèmes d'ambiguïté. La désambiguïsation attribue à chaque unité lexicale une étiquette unique (classe syntaxique et information morphologique de base) en contexte. Parmi les expériences de désambiguïsation tentées depuis quelques années, les chercheurs ont eu recours à de nombreuses techniques. Parmi celles-ci, soulignons la sélection « statistique » des mots traduits de la requête, le calcul basé sur la moyenne relative de la fréquence des termes, l'utilisation de plusieurs critères afin de déterminer le sens d'un mot dans un contexte, y compris les valeurs syntaxique, sémantique et pragmatique, de même que les relations de cooccurrences syntaxiques, le développement des requêtes utilisant la mise en grappes des termes et des documents, etc. [Belguith et al., 2006],[Adubert, 2003],[Crestan et al., 2003].

Les techniques de désambiguïsation existantes pour la réduction des ambigüités qui nécessitent un traitement approfondi, débordent le cadre de notre travail.

2.3.4.1. Quelques techniques de désambigüisation

Comme nous avons mentionnés précédemment, nous détaillons deux techniques différentes, utilisées pour réduire les ambigüités.

2.3.4.1.1. Segmentation des textes

La plupart des méthodes de segmentation sont destinées à la reconnaissance de frontières thématiques et à la création automatique de résumés.

La segmentation, sert à isoler les parties pertinentes des parties non pertinentes des textes :

Il existe plusieurs techniques de segmentation :

- Segmentation basée sur les signes de ponctuations :

Il s'agit de séparer les phrases reliées par des signes graphiques tels que : le point, le point virgule, la virgule…et.

En effet la langue arabe n'est pas basée principalement sur les signes de ponctuations et pour la séparation entre les phrases ; ces derniers ont généralement un rôle pausale. Par conséquent, nous pouvons trouver tout un paragraphe qui ne contient aucun signe de ponctuation à part un point à la fin. La séparation entre les phrases peut-être avec certaines particules et certains mots connecteurs.

- Segmentation basée sur les mots connecteurs

Certains particules, exemple : 'و' (waw), 'ف' (fa), 'ثم' (thuma) jouent un rôle principal dans la séparation des phrases. Nous citons à titre d'exemple l'énoncé suivant :

كان امرؤ القيس وطرفة رجلين طائشين وحياتهما غير منضبطة، وماتا ميتة عنيفة في عز شبابهما. بينما عاش زهير حياة طويلة ونال احترام الجميع لحكمته وأخلاقه العالية ,ثم انه لم يكن بحاجة للآخرين.

En Arabe, les mots connecteurs sont au nombre de 7:

من (Men), عن (Aan)على , (Ala), ب (Ba) , في(Fi) , ك (Kef), ل (Lem) .

- Segmentation basée sur les marqueurs linguistiques

Il s'agit de séparer les phrases reliées par des connecteurs (aussi appelés mots-outils, indicateurs ou signaux) tels que :

إذن, في البداية, مقدمة, بمعنى....

Ces marqueurs sont généralement utilisés pour effectuer un filtrage sémantique sur le texte. Ils permettent d'accéder au contenu sémantique d'un texte sans avoir recours à des analyses syntaxiques profondes ou à des connaissances extérieures. Nous citons à titre d'exemple l'énoncé suivant :

...وكان من أوائل من دخل الإسلام. ورد في كتاب الأغاني أن الرسول قابل زهير وهو في سن المائة وقال : " اللهم أعذني من شيطانه ". ويقال إنه توفي قبل أن يغادر الرسول البيت .في رواية أخرى أن زهير تنبأ بقدوم الرسول وذكر ذلك لابنيه كعب وبوجير، ونصحهم بالاستماع إلى كلام الرسول عند قدومه، وهذا يعني أنه توفي قبل ظهور الرسالة. كتاب الأغاني

كتاب ⟶ Document.

إن, أن ⟶ Subordonnant.

ذلك, هذا ⟶ Démonstratif.

يعني ⟶ Explication.

2.4.1.2. Détection de la racine

Le lexique arabe comprend trois catégories de mots : verbes, noms et particules. Les verbes et noms sont le plus souvent dérivés d'une racine à trois consonnes radicales, Une famille de mots peut ainsi être générée d'un même concept sémantique à partir d'une seule racine à l'aide de différents schèmes. Ce phénomène est une caractéristique à la morphologie arabe. On dit donc que l'arabe est une langue à racines réelles à partir desquelles, on déduit le lexique arabe selon des schèmes qui sont des adjonctions et des manipulations de la racine [Douzidia, 2004].

Le tableau suivant, donne quelques exemples de schèmes appliqués au mot كتب (écrire). On peut ainsi dériver un grand nombre de noms, de formes et de temps verbaux.

Schèmes	**كتب**	**Notion d'écrire**
فَاعِلٌ	كَاتِبٌ	écrivain
فَعَلَ	كَتَبَ	a écrit
مَفْعَلٌ	مَكْتَبٌ	bureau
فُعِلَ	كُتِبَ	a été écrit

Tableau 2.6 : Exemple de schèmes pour le mot كتب (écrire).

Pour détecter la racine d'un mot, il faut connaître le schème (الوزن) par lequel il a été dérivé et supprimer les éléments flexionnels (antéfixes, préfixes, suffixes, post fixes) qui ont été ajoutés, parce que en arabe un mot peut signifier toute une phrase grâce à sa structure composée qui est une agglutination d'éléments de la grammaire. Le modèle qu'on présenté précédemment (paragraphe 2.3.2) schématise une structure possible d'un mot. Notons que la lecture et l'écriture d'un mot se fait de droite vers la gauche.

Enclitique	**Suffixe**	**Schème**	**Préfixe**	**Proclitique**

- Proclitiques sont des prépositions ou des conjonctions.

- Préfixes et suffixes expriment les traits grammaticaux et indiquent les fonctions : cas du nom, mode du verbe et les modalités (nombre, genre, personne,…)

- Enclitiques sont des pronoms personnels.

Exemple :

أستذكرونه

Ce mot exprime la phrase en français : "Est ce que vous vous souvenez de lui ?"

La segmentation de ce mot donne les constituants suivants :

Proclitique : أ‌س conjonction d'interrogation

Préfixe : ت Préfixe verbal du temps de l'inaccompli.

Corps schématique:تَذَكَّرَ dérivé de la racine: ذَكَرَ selon le schème فَعَلَ

Suffixe : ون suffixe verbal exprimant le pluriel.

Enclitique : ها pronom suffixe complément du nom.

La phase d'analyse morphologique détermine un schème possible. Les préfixes et suffixes sont trouvés en enlevant progressivement des préfixes et des suffixes et en essayant de faire correspondre toutes les racines produites par un schème afin de retrouver la racine .Lorsqu'un mot peut être dérivé de plusieurs racines différentes, la détection de la racine est encore plus difficile, en particulier en absence de voyelles.

Lorsqu'un mot peut être dérivé de plusieurs racines différentes, la détection de la racine est encore plus difficile, en particulier en absence de voyelles.

2.4. Etapes de processus du prétraitement

A partir de ce que nous avons vue précédemment, un processus de prétraitement peut contenir toutes les étapes suivantes :

1. Encodage unique des textes ;
2. Normalisation des textes ;
3. Suppression des mots vides ;
4. Correction des fautes d'orthographes et des incohérences ;

5. Traitement des ambigüités.

Nous détaillons ces étapes dans le chapitre 5.

Parmi les applications existantes qui se basent sur la phase de prétraitement de textes, nous citons : *Aramarph, GATE*, *NOOJ*. Une des fonctionnalités de ce dernier est l'extraction d'informations à partir du corpus « brut », *i. e.* ne comportant pas de balises spécifiques. *NOOJ* contient un analyseur morphologique qui permet d'effectuer des recherches et des traitements dans les textes à partir d'expressions régulières intégrant des formes, des lemmes, des catégories syntaxiques ou toute information lexicale [Site6 , 2009]. A titre d'exemple :

L'expression suivante permet de rechercher dans un corpus tous les noms qui finissent par *-ion*, *-age* ou *-ment*.

<N+MP= "(ion|age|ment)$">

Pour telles requêtes, NOOJ, utilise toutes les techniques de prétraitement de textes, pour effectuer des recherches dans des corpus de textes buts.

2.5. Conclusion

Dans ce chapitre, nous avons présenté les difficultés du traitement automatique de la langue arabe ainsi que les différents types d'ambigüités qui peuvent êtres imposées, et quelques techniques de désambigüisation comme la segmentation de textes et la détection des racines.

Le lexique utilisé (les préfixes, les suffixes, les enclitiques, les proclitique, le pré-bases, les postes-bases) en traitement automatique de l'arabe, nous permet de segmenter les mots pour nous permettre de séparer les unités lexicales étiquetés.

Enfin, l'étiquetage des unités lexicales nous permet de calculer facilement la fréquence des unités lexicales figurant dans le texte. Le prétraitement des textes est considéré comme pré-requis pour les représenter sous forme manipulable par la machine, nous détaillons dans le chapitre 3, les méthodes de représentation textuelle.

CHAPITRE 3
Représentation de textes

3.1. Introduction

La notion de similarité entre documents est évidement fortement liée au choix de la méthode de représentation des textes. La représentation la plus utilisée est la représentation vectorielle (mise en œuvre en particulier, dans les systèmes de recherche documentaire), dans le cadre de laquelle un document est représenté par un vecteur dans un espace vectoriel dont les dimensions sont associées à des unités linguistiques spécifiques (mots, stems, lemmes,.. etc.). La similarité entre documents est alors évaluée par mesure de similarité définie sur cet espace vectoriel.

Nous allons présenter dans ce chapitre les notions principales de la représentation textuelle pour le traitement du langage naturel et la recherche d'informations. Nous décrirons notamment les concepts de base sur les espaces vectoriels des données, ainsi que les différentes approches en représentation de textes : la méthode vectorielle, la méthode LSA, et ExpLSA.

3.2. Concepts fondamentaux

Dans cette partie nous présentons quelques concepts utilisés lors de notre étude pour la représentation textuelle des corpus.

3.2.1. Les espaces vectoriels

Il est possible de présenter les espaces vectoriels comme une généralisation de l'espace géométrique ordinaire à trois dimensions. Un espace vectoriel peut avoir un nombre quelconque de dimensions, mais ce n'est qu'au milieu du 19ème siècle que les mathématiciens commencèrent à accepter l'idée d'espaces à plus de trois dimensions. Le nombre de dimensions (la dimension) d'un espace vectoriel est alors le nombre minimal d'axes de coordonnées nécessaires pour définir tout point de cet espace. De tels axes sont indépendants entre eux, et la notion d'indépendance linéaire, fondamentale en algèbre linéaire, est également cruciale pour l'étude des espaces vectoriels.

La théorie des espaces vectoriels est souvent exposée de manière purement axiomatique et formelle. Ainsi un espace vectoriel se définit comme un ensemble d'éléments (les vecteurs) muni de deux opérations internes particulières (l'addition vectorielle et la multiplication par un nombre scalaire). L'ensemble est fermé pour ces opérations, qui redonnent toujours des éléments de l'ensemble, c'est-à-dire des vecteurs. Cette définition formelle a l'avantage que les vecteurs peuvent être des objets très variés, comme des polynômes ou des fonctions.

En ajoutant ensuite à cette structure algébrique une opération telle que le produit scalaire (défini plus loin), on munit un espace vectoriel d'une mesure de distance entre vecteurs. Cette mesure permet une interprétation géométrique de l'espace vectoriel, point de vue qui se révèle souvent très intuitif et heuristique dans de nombreux problèmes. [Memmi, 2000]

3.2.2. Les vecteurs

Un vecteur est un ensemble de valeurs, ou composantes, représentant typiquement un objet ou un individu par des traits numériques. Par exemple, on peut décrire les habitants d'une ville par leur âge, revenu, niveau d'éducation, nombre d'enfants... Des traits qualitatifs (non numériques) comme le sexe, le statut marital, la profession, peuvent se traduire aisément en valeurs binaires, donc également numériques. Les traits peuvent être pondérés selon leur importance, mais ne sont pas autrement structurés entre eux.

En prenant ces valeurs comme des coordonnées dans un espace multidimensionnel, on retrouve la conception géométrique du vecteur : un point dans un espace à n dimensions. Ce point correspond à un segment de droite dirigé (une flèche) à partir de l'origine des coordonnées, ce qui est une représentation familière (bien que simpliste) des vecteurs. Le nombre de traits choisis pour décrire les individus en jeu est la dimension de l'espace vectoriel. Cette représentation vectorielle a l'avantage de récupérer dans une certaine mesure (demandant des précautions) notre intuition de l'espace physique habituel à trois dimensions, tout en permettant un nombre de dimensions quelconque, de plusieurs milliers si nécessaire. Le caractère à la foi algébrique et géométrique de l'algèbre linéaire en font ainsi un domaine très fructueux. La théorie est maintenant très bien comprise et formalisée, et on en a tiré de nombreuses applications.

En bref, on peut voir plus ou moins intuitivement un espace vectoriel comme un espace abstrait à nombre quelconque de dimensions. Une fois que des objets (par exemple des documents) auront été représentés par de vecteurs dans un espace vectoriel approprié, on pourra les traiter grâce aux opérations usuelles sur les vecteurs. Les opérations de base sont l'addition vectorielle et la multiplication par u scalaire, qui sont des généralisations de l'addition et de la multiplication ordinaires.

On additionne deux vecteurs en additionnant leurs composantes terme à terme (les deux vecteurs doivent avoir la même dimension), on multiplie un vecteur en multipliant chacune de ses composantes par un nombre. D'autres opérations s'en déduisent aisément, comme la soustraction vectorielle et la division par un scalaire.

Mais on cherche aussi très souvent à mesurer la ressemblance ou similitude de deux vecteurs, et on dispose pour cela d'opérations précises et faciles à calculer comme le produit scalaire. Ce dernier permet de mesurer des notions géométriques comme longueur, angle ou distance.[Memmi, 2000].

3.3. Méthodes de représentation des textes

D'après [Jaillet, 2004], il existe trois méthodes de représentation de textes : la représentation statistique, la représentation conceptuelle, et une autre méthode qui réunit les avantages des représentations statistiques et conceptuelles, appelée représentation mixte.

3.3.1. Représentation statistique

Le formalisme le plus utilisé pour représenter les textes est le formalisme vectoriel. Dans ce formalisme, chaque dimension de l'espace vectoriel correspond à un mot, que l'on nomme terme d'indexation. La représentation vectorielle consiste à associer à chaque mot une dimension au sein de l'espace. Cette représentation offre l'avantage de représenter chaque sens sur une dimension propre de l'espace.

L'utilisation des mots, est possible mais pose toutefois un certain nombre de problèmes. En effet, il existe plusieurs dizaines de milliers de mots et associer à chacun de ces mots un sens, c'est à dire une dimension de l'espace, est maladroit. Un prétraitement linguistique en amont de la représentation est la plupart du temps mis en place afin de résoudre ce problème. Les deux prétraitements les plus connus sont : la radicalisation et la lemmatisation.

3.3.2. Représentation conceptuelle

Une autre méthode de représentation, bien que se basant aussi sur le formalisme vectoriel pour représenter les documents, reste fondamentalement différente de la représentation précédente. Les dimensions de l'espace vectoriel ne sont pas associées ici à des termes d'indexation mais à des concepts. Pour permettre une telle représentation des documents, il est nécessaire de pouvoir projeter n'importe quelle lexie du dictionnaire sur l'espace généré par l'ensemble des concepts prédéfinis.

Exemple : Les concepts : *pic*, *cime*, *sommet*, *crête* peuvent être réduits à un seul.

3.3.3. Représentation mixte

L'avantage de la représentation conceptuelle est en particulier, de réduire les effets synonymiques du vocabulaire Par exemple, "pic", "cime", "sommet", "crête" possèdent des sens en commun. Lors d'une représentation statistique, chaque mot sera associé à une dimension. Il n'y aura donc aucune ressemblance entre des textes utilisant ces différents mots. L'avantage de la représentation conceptuelle est que des mots synonymes partagent au moins un concept. Cependant, l'inconvénient majeur de la représentation conceptuelle est que les noms propres du document ne sont pas pris en compte. En effet les noms propres, étant sémantiquement vides par définition, ne possèdent pas de représentation au sein du thésaurus. Par exemple les mots « Ferrari » et « Renault » sont définis comme des vecteurs « nuls » alors qu'ils peuvent être utile lors d'un processus de catégorisation, notamment pour des catégories de type : « Automobile ».

L'idée de la représentation mixte est donc d'allier, à une représentation conceptuelle pure, une dimension statistique supplémentaire. Cette double représentation des textes a pour avantage de fournir deux informations différentes et complémentaires à un processus de catégorisation. En effet, la représentation statistique permet de mettre en évidence le vocabulaire discriminant tandis que la représentation conceptuelle permet quant à elle, d'obtenir une vision plus globale du texte en projetant ce dernier sur un ensemble de concepts. Cette projection permet d'en déduire le « champ sémantique » du texte en question. Chaque document dans le processus de catégorisation mixte sera représenté par un vecteur défini comme la concaténation des deux vecteurs suivants :

- La première moitié du vecteur mixte correspondra au vecteur statistique.
- La seconde moitié du vecteur mixte correspondra au vecteur conceptuel.

3.4. Représentation sémantiques des textes

Une des difficultés majeures de la catégorisation concerne la dimension extrêmement élevée de l'espace de représentation. Celui-ci se compose en effet d'un ensemble de termes uniques (mots ou phrases) dont la dimension peut atteindre plusieurs centaines de milliers pour une collection de textes relativement modérée, or seuls quelques traitements basés sur les réseaux neuronaux sont actuellement capables de traiter un si grand nombre de nœuds. Il est donc hautement souhaitable de réduire la dimension de l'espace d'origine, mais sans sacrifier pour autant la précision de la classification.

Au cours de dernières années, plusieurs recherches ont été axe pour remédier ce problème.

Nous présentons dans cette partie les deux célèbres méthodes de représentation sémantique des textes : LSA et ExpLSA.

3.4.1. La méthode LSA (Latent Semantic Analysis)

La méthode LSA est fondée sur le fait que des mots qui apparaissent dans un même contexte sont sémantiquement proches. Le corpus est représenté sous forme matricielle. Les lignes sont relatives aux mots et les colonnes représentent les différents contextes choisis (un document, un paragraphe, une phrase, etc.). Chaque cellule de la matrice représente le nombre d'occurrences des mots dans chacun des contextes du corpus. Deux mots proches au niveau sémantique sont représentés par des vecteurs proches. La mesure de proximité est généralement définie par le cosinus de l'angle entre les deux vecteurs. [Béchet, et al., 2008].

	Contexte 1		Contexte j		Contexte P
Mot 1					
Mot j			a_{ij}		
Mot N					

Figure 3.1 : Représentation matricielle d'un corpus de texte

a_{ij} : représente le nombre d'occurrence du mot i dans le contexte j.

3.4.1.1. Limites de LSA

LSA offre des avantages parmi lesquels, la notion d'indépendance par rapport à la langue du corpus étudié, le fait de se dispenser de connaissances linguistiques ainsi que de celles du domaine, tels que des thésaurus. Bien que cette approche soit pertinente pour les tâches de classification, il n'en demeure pas moins que son utilisation soulève des contraintes. Notons tout d'abord l'importance de la taille des contextes choisis. [Rehder et al. 1998] ont montré lors de leurs expérimentations que si les contextes possèdent moins de 60 mots, les résultats s'avèrent être décevants.

Il a également été mis en évidence par [Roche et al., 2006] que l'efficacité de LSA est influencée par la proximité du vocabulaire utilisé. En effet, l'homogénéité des corpus sur le plan thématique donne des résultats décevants avec LSA. [Béchet, et al., 2008].

Pour résoudre de tels problèmes, une des solutions peut consister à ajouter des connaissances syntaxiques à LSA, comme cela est décrit dans la section suivante.

3.4.1.2. L'ajout de connaissances syntaxiques à LSA

[Landauer et al., 1997] posent le problème du manque d'informations syntaxiques dans LSA en comparant cette méthode à une évaluation humaine. Il est question de proposer à des experts humains d'attribuer des notes à des essais sur le cœur humain de 250 mots rédigés par des étudiants. Un espace sémantique a été créé à partir de 27 articles écrits en anglais traitant du cœur humain « appris » par LSA. Les tests effectués concluent que la méthode LSA obtient des résultats satisfaisants comparativement à l'expertise humaine. Il en ressort que les mauvais résultats étaient dus à une absence de connaissances syntaxiques dans l'approche utilisée. Ainsi, les travaux qui sont décrits ci-dessous montrent de quelle manière de telles connaissances peuvent être ajoutées à LSA.

La première approche de [Wiemer-Hastings et al.,2001], utilise des étiquettes grammaticales Brill, 1994 appliquées à l'ensemble du corpus étudié (corpus de textes d'étudiants). Les étiquettes étant rattachées à chaque mot avec un blanc souligné (« _ »), l'analyse qui s'en suit via LSA considère le mot associé à son étiquette comme un seul terme. Les résultats de calculs de similarités obtenus avec une telle méthode restent décevants.

Notons que de telles informations grammaticales ne sont pas des connaissances syntaxiques proprement dites contrairement à la seconde approche de [Wiemer-Hastings et al.,2001] décrite ci-dessous.

Cette seconde approche se traduit par l'utilisation d'un analyseur syntaxique afin de segmenter le texte avant d'appliquer l'analyse sémantique latente. Cette approche est appelée « LSA structurée » (SLSA). Une décomposition syntaxique des phrases en différents composants (sujet, verbe, objet) est tout d'abord effectuée. La similarité est ensuite calculée en traitant séparément par LSA les trois ensembles décrits précédemment. Les similarités (calcul du cosinus) entre les vecteurs des trois matrices formées sont alors évaluées. La moyenne des similarités est enfin calculée. Cette méthode a donné des résultats satisfaisants par rapport à « LSA classique » en augmentant la corrélation des scores obtenus avec les experts pour une tâche d'évaluation de réponses données par des étudiants à un test d'informatique. [Kanejiya et al., 2003], proposent un modèle appelé SELSA. Au lieu de générer une matrice de co-occurrences mot/document, ils ont proposé une matrice dans laquelle chaque ligne contient toutes les combinaisons mot/étiquette et en colonne les documents. L'étiquette « préfixe » renseigne sur le type grammatical du voisinage du mot traité. Le sens d'un mot est en effet donné par le voisinage grammatical duquel il est issu. Cette approche est assez similaire à l'utilisation des étiquettes de [Brill, 1994] présentée dans les travaux de [Wiemer-Hastings et al., 2001].

Mais SELSA étend ce travail vers un cadre plus général où un mot avec un contexte syntaxique spécifié par ses mots adjacents, est considéré comme une unité de représentation de connaissances. L'évaluation de cette approche a montré que la méthode LSA était plus pertinente que SELSA dans un test de corrélation avec des experts. Cependant, SELSA se

révèle plus précise pour ce qui est de tester les bonnes et mauvaises réponses (SELSA fait moins de fautes que LSA mais en retourne de plus nuisibles). [Béchet, et al., 2008].

3.4.2. La méthode ExpLSA (Expansion Latent Semantic Analysis)

L'approche ExpLSA place dans un contexte différent. L'utilisation de ressources lexicales et sémantiques pour enrichir des contextes est un concept répandu dans le domaine de la recherche d'informations textuelles, pour des tâches d'indexation ou d'expansion de requêtes. La plupart de ces approches utilisent des ressources lexicales, en ajoutant des termes reliés sémantiquement aux termes d'origine.

3.4.2.1. Principe de la méthode ExpLSA

L'approche ExpLSA propose d'enrichir un corpus lemmatisé en effectuant une expansion des phrases. Cette expansion se fonde sur une méthode syntaxique afin de compléter les mots du corpus avec des mots jugés sémantiquement proches.

Dans ce qui suit, nous résumons le principe de cette approche qui sera appliquée pour une tâche de classification de textes.

3.4.2.1.1. Utilisation d'un analyseur syntaxique

Dans [Béchet, et al., 2008], les auteurs ont utilisés l'analyseur syntaxique SYGFRAN [Chauché, 1984] afin d'extraire du corpus, les relations syntaxiques Verbe-Objet (Verbe_Préposition_Complément, Verbe_COD). Soit la phrase « *L'accompagnement nécessite des professionnels*» la relation syntaxique » verbe : nécessiter, COD : professionnels ». Une fois la totalité des relations syntaxiques extraites, ils ont lemmatisés le corpus en utilisant le système SYGMART [Chauché, 1984].

3.4.2.1.2. Regroupement des objets en fonction de la proximité des verbes

Nous évaluons ensuite la proximité sémantique entre les verbes. Cette mesure considère deux verbes comme proches s'ils possèdent un nombre important d'objets en commun en fonction du nombre total d'objets de chaque verbe.

Exemple 1 :

Soit les énoncés suivants :

ناقش الطرفان أزمة الصحراء الغربية و الوضع في العراق

تحادث الطرفان في ما يخص الوضع في العراق و أزمة الصحراء الغربية

Nous remarquons que les verbes :

تحادث, ناقش: possèdent plusieurs objets en commun, donc ils sont sémantiquement proches.

Exemple 2 :

Soit les énoncés suivants :

العمل في مثل هذه الظروف يحتاج إلى الصبر, المثابرة و الشجاعة

العمل في مثل هذه الظروف يتطلب الصبر, المثابرة و الشجاعة

Après avoir étudié la proximité sémantique entre les verbes en évaluant chaque verbe du corpus avec tous les autres, nous ne conservons que le couple de verbes ayant obtenu le meilleur score de similarité. Citons par exemple les verbes يحتاج et يتطلب qui partagent communément les objets : الصبر, المثابرة, الشجاعة.Ainsi, nous regroupons tous les objets communs dont les verbes ont été jugés proches sémantiquement par le seuil de similarité le plus élevé parmi l'ensemble des couples de verbes. Nous complétons alors le corpus initial en attachant à chaque mot les objets communs.

La dernière étape d'ExpLSA est l'application de l'approche LSA sur le corpus enrichi.

3.5. L'enrichissement appliqué à la classification de textes

La classification de textes consiste à regrouper des contextes (dans notre cas des documents) dans différentes classes qui correspondent à des catégories thématiques (par exemple, les thèmes *« politique »*, *« sport »*, *« technologies »*, etc). Un contexte contient des descripteurs qui peuvent être insuffisants pour la réalisation d'une classification automatique. L'approche ExpLSA propose une solution à ce manque d'information en enrichissant le contexte.

Prenons par exemple, les phrases suivantes :

P1 : **دافع المحامي عن المتهم**

P2 : **نطق القاضي بالحكم**

On constate que les phrases P1 et P2 n'ont aucun mot en commun (sans considérer les mots outils) ce qui classerait ces phrases dans deux catégories différentes en nous appuyant sur des méthodes statistiques. Après expansion avec la méthode ExpLSA, il est possible d'enrichir ces phrases de la manière suivante (pour faciliter la lecture de cet exemple, la lemmatisation n'a pas été ici répercutée) :

P3 : **دافع (محامي قاضي حكم) عن (متهم حكم قضية محكمة)**

P4 : **نطق (قاضي محامي حكم) بـ(حكم متهم قضية محكمة)**

Dans ce cas, les phrases P3 et P4 possèdent six mots communs signifiant une proximité thématique.

On montre par cet exemple que deux phrases proches sémantiquement peuvent s'avérer difficiles à classer sans utiliser de connaissances sémantiques. Avec notre enrichissement, l'information apportée peut remédier à cette problématique.

3.6. Comparaison

Selon les travaux de [Béchet, et al., 2008] , L'approche ExpLSA combiné avec des algorithmes de classification appliqué sur un corpus de 2 828 articles et est constitué de 914 540 mots (5,3 Mo), a donné les meilleurs résultats par rapport aux autres approches combinés avec les mêmes algorithmes. Les résultats obtenus ont montrés aussi que l'approche ExpLSA combiné avec l'algorithme KPPV a donné les meilleurs résultats.

3.7. Conclusion

Dans ce chapitre, nous avons présenté les concepts de base de la représentation textuelle et les différentes approches utilisées pour représenter un texte.

Nous avons présenté aussi les deux célèbres méthodes LSA et ExpLSA pour la représentation sémantique des textes, LSA est une méthode statistique utilisée entre autres pour regrouper des contextes afin d'établir une classification de textes. Néanmoins, cette méthode donne des résultats parfois décevants. Ceux-ci s'expliquent notamment par l'absence de connaissances linguistiques. L'approche ExpLSA consiste à effectuer une expansion des contextes avant d'appliquer LSA.

Selon [Béchet, et al., 2008], ExpLSA a montré sa performance pour la classification des textes de grandes tailles.

CHAPITRE 4
Classification d'opinions

4.1. Introduction

La classification des documents est une technologie de base de fouille de texte, le but est le suivi de thèmes et la construction de résumés en cherchant des proximité ou des cohérences intra-groupes se situant au niveau de vocabulaire et en particularité des proximité sémantiques, c'est-à-dire regroupement des textes qui « parlent de la même chose ».

La classification d'opinions est une tâche importante en opinion-mining, il s'agit de regrouper les opinions des utilisateurs envers un objet selon leurs orientations sémantiques en trois catégories : *positive*, *négative*, et *neutre*. Cette tâche a attiré l'attention de plusieurs chercheurs en TAL, à titre d'exemple pour détecter les points faibles et les points forts d'un objet commenté par les utilisateurs. Nous trouvons que les travaux de recherche actuels dans ce domaine, se basent sur cette tâche à cause de son importance. [Hu et al., 2005], [Maurel et al., 2007].

Les techniques de la classification automatiques sont répertoriées selon deux principales approches, la première est dite classification supervisée ou catégorisation et est basée sur l'apprentissage supervisé, le deuxième est dite classification non supervisée ou clustering ou encore apprentissage non supervisée pour la classification [Kelaiaia, 2008].

Dans ce chapitre, nous nous sommes intéressés à la classification d'opinions i.e. : segments de textes subjectifs porteurs d'opinions, en fonction des valeurs de jugements qu'ils contiennent. Dans un premier lieu, nous présentons, la définition de quelques concepts essentiels, les deux techniques de classification : supervisée et non supervisée, les critères d'agrégations utilisées pour comparer les classes deux à deux pour sélectionner les classes les plus similaires, les approches utilisées pour chaque technique. Enfin, les méthodes de classification existante actuellement pour la classification d'opinions : la méthode symbolique, statistique, et hydrique, que nous allons détailler dans le présent chapitre.

4.2. Définitions

- **Classification**

La classification est une méthode d'analyse de données qui vise à regrouper en classes homogènes un ensemble d'observations.

- **Partition**

La partition est un ensemble de parties non vides, deux à deux disjointes et dont la réunion est égale à l'ensemble partitionné

On appelle donc, partition de I, l'ensemble P, P = { $C_i, i \in I$ }, tels que:

I: ensemble d'indices.

C_i: partie de I (ou une classe).

Possédant les propriétés suivantes:

1. $\forall i \in I, C_i \neq \phi$
2. $\forall i \in I, \forall j \in I, i \neq j \rightarrow C_i \cap C_j = \phi$
3. $\bigcup_{i \in I} C_i = I$

- **Recouvrement**

Un recouvrement d'un ensemble I est un ensemble P de sous-ensembles non vides de I tel que l'union de ces sous-ensembles soit égale à I. Autrement dit P est un recouvrement de I si et seulement si tout élément x de I se trouve dans au moins l'un des éléments de P. Une partition est un recouvrement particulier.

On appelle donc, recouvrement de I, l'ensemble de P, P= { $C_i, i \in I$ }

I: ensemble d'indices.

C_i: partie de I (ou une classe)

Possédant les propriétés suivantes:

1. $\forall i \in I, C_i \neq \phi$
2. $\bigcup i \in I, C_i = I$

- **Hiérarchie**

On appelle hiérarchie de parties de I, toute sous-ensemble, H de P(I) tel que :

1. $\phi \in H$
2. $\forall x \in I, \{ x \} \in H$
3. $I \in H$
4. $\forall H_1 \in H, \forall H_2 \in H \Rightarrow H_1 \cap H_2 = \phi$ ou $H_1 \subset H_2$ ou $H_2 \subset H_1$

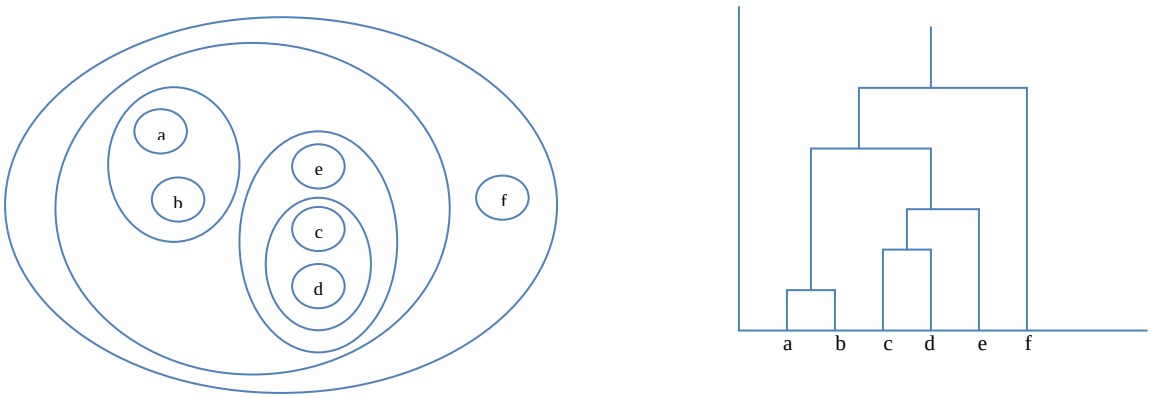

Figure 4.1 : Une hiérarchie de partie de I

- **Hiérarchie indicée de parties**

On appelle hiérarchie indicée de parties de I, la donnée (H,h) où H: Hiérarchie de parties.

h: $H \to \Re^+$ tel que

$$\forall x \in I, h(\{x\}) = 0; \forall H_1 \in H, \forall H_2 \in H, H_1 \subseteq H_2 \Rightarrow h(H_1) \leq h(H_2)$$

h(H) s'appelle niveau hiérarchique de H.

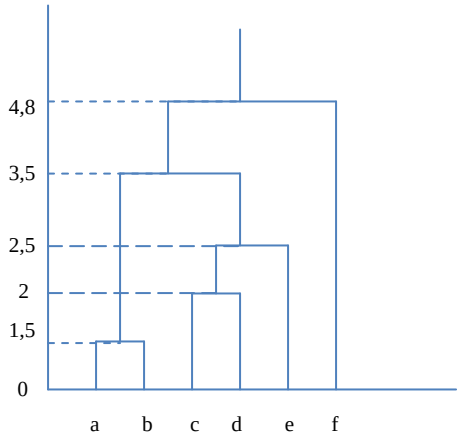

Figure 4.2 : Exemple d'hiérarchie indicée de parties

Une hiérarchie indicée permet de construire des familles de partition totalement ordonnées. [Boubou, 2007].

4.3. Techniques de classification de textes

Comme on a déjà vue, il existe deux techniques de classification de textes : classification supervisée et classification non supervisé.

4. 3.1. Classification supervisée

La classification supervisée suppose qu'il existe déjà une classification de documents. C'est le cas par exemple d'une bibliothèque ou d'un moteur de recherche. Le but est alors de classer automatiquement un nouveau document.

Comme les documents sont nombreux ou que leur nombre augmente sans cesse, il serait difficile de programmer à l'avance des règles de décision pour déterminer la classe d'un nouveau document. Même si cela était possible, ces règles devraient être régulièrement modifiées par l'utilisateur pour qu'elles reflètent la réalité actuelle.

Soit $D = \{d_1, d_2, \dots, d_i, \dots, d_m\}$, un ensemble de documents représentés chacun par une description $\overrightarrow{d_1}$, $\overrightarrow{d_2}$,…, $\overrightarrow{d_m}$, et $C = \{C_1, C_2, \dots, C_k, \dots, C_c\}$, un ensemble de classes, la classification supervisée suppose connues deux fonctions. La première fait correspondre à tout individu d_i, une classe C_k, Elle est définie au moyen des couples (d_i,C_k) donnés comme exemples au système. La deuxième fait correspondre à tout individu d_i, sa description $\overrightarrow{d_i}$. La classification supervisée consiste alors à déterminer une procédure de classification : $C^f : \overrightarrow{d_i} \rightarrow C_k$, qui à partir de la description de l'élément détermine sa classe avec le plus faible taux d'erreurs. La performance de la classification dépend notamment de l'efficacité de la description. De plus, si l'on veut obtenir un système d'apprentissage, la procédure de classification doit permettre de classer efficacement tout nouvel exemple (pouvoir prédictif). [Site1, 2008].

Un exemple de classification supervisée concerne la médecine : étant donnés les résultat d'analyse d'un patient, et la connaissance de l'état d'autres patients pour lesquels les mêmes analyses ont été menées, il est possible d'évaluer le risque de maladie de ce nouveau patient en fonction de la similarité de ses analyses avec celles des autres patients.[Candilier, 2006].

4.3.1.1. Algorithmes de classification supervisée

La plupart des algorithmes d'apprentissage supervisés tentent donc de trouver un *modèle* (une fonction mathématique), qui explique le lien entre des données d'entrée et les classes de sortie.

Dans le cas de la classification de documents, on fournit donc à la machine des exemples sous la forme (Document, Classe). Cette méthode de raisonnement est appelée *inductive* car on induit de la connaissance (le modèle) à partir des données d'entrée (les Documents) et des sorties (leurs Catégories). Grâce à ce modèle, on peut alors déduire les classes de nouvelles données : le modèle est utilisé pour prédire. Le modèle est bon s'il permet de bien prédire. [Site1, 2008].

Il existe de nombreuses méthodes d'apprentissage supervisé :

- K plus proches voisins (et ses variantes : Category-based Search et Cluster-based Search) ;
- Arbres de décisions ;
- Naïve Bayes (ou encore Simple Bayes) ;
- Réseaux de neurones ;
- Machines à support de vecteurs (ou SVM) ;
- Programmation génétique.

(Voir l'explication de ces algorithmes en annexe 6)

4.3.2. Classification non supervisée

La classification non-supervisée est utilisée lorsque que l'on possède des documents qui ne sont pas classés et dont on ne connaît pas de classification. A la fin du processus de classification non-supervisée, les documents doivent appartenir à l'une des classes générées par la classification. On distingue deux catégories de classifications non-supervisées : hiérarchiques et non-hiérarchiques.

Dans la **classification hiérarchique(CH)**, les sous-ensembles créés sont emboîtés de manière hiérarchique les uns dans les autres. On distingue la CH *descendante (ou divisive)* qui part de l'ensemble de tous les individus et les fractionne en un certain nombre de sous-ensembles, chaque sous-ensemble étant alors fractionné en un certain nombre de sous-ensembles, et ainsi de suite. Et la CH *ascendante (ou agglomérative)* qui part des individus seuls que l'on regroupe en sous-ensembles, qui sont à leur tour regroupés, et ainsi de suite. Pour déterminer quelles classes on va fusionner, on utilise le critère d'agrégation.

Dans la **classification non-hiérarchique**, les individus ne sont pas structurés de manière hiérarchique. Si chaque individu ne fait partie que d'un sous-ensemble, on parle de *partition*. Si chaque individu peut appartenir à plusieurs groupes, avec la probabilité P_i d'appartenir au groupe i, alors on parle de *recouvrement*.

4.3.2.1. Quelques algorithmes de classification non supervisée

- Algorithme **CURE** (**C**lustering **U**sing **RE**presentatives)
- Algorithme **BIRCH** (**B**alanced **I**terative **R**educing and **C**lustering using **H**ierarchies)
- Algorithme de **ROCK** (**RO**bust **C**lustering using lin**K**s)
- Algorithme **TSVQ** (**T**ree **S**tructured **V**ector **Q**uantization)

(Voir l'explication de ces algorithmes en annexe 7)

4.3.3. Critères d'agrégation

Le critère d'agrégation permet de comparer les classes deux à deux pour sélectionner les classes les plus similaires suivant un certain critère.

Les critères les plus classiques sont le plus proche voisin, le diamètre maximum, la distance moyenne et la distance entre les centres de gravités. [Site1, 2008]

De nombreux critères d'agrégation ont été proposées les plus connus sont:

- **Le critère du saut minimal**

La distance entre deux classes C_1 et C_2 est définie par la plus courte distance séparant un individu de C_1 et un individu de C_2:

$$D(C1, C2) = \min(\{d(x, y)\ \}), x \in C_1, y \in C_2)$$

- **Le critère de saut maximal**

La distance entre deux classes C_1 et C_2 est définie par la plus grande distance séparant un individu de C_1 et un individu de C_2.

$$D(C1, C2) = \max(\{d(x, y)\ \}), x \in C_1, y \in C_2)$$

- **Le critère de la moyenne**

Ce critère consiste à calculer la distance moyenne entre tous les éléments de C1 et tous les éléments de C2.

$$D(C1, C2) = \frac{1}{n_{C_1} \times n_{C_2}} \sum_{x \in C_1} \sum_{y \in C_2} d(x, y)$$

Avec:
- nc_1 : Le cardinal de C1
- nc_2 : Le cardinal de C2

- **Le Critère de Ward**

Ce critère ne s'applique que si on est muni d'un espace euclidien. La dissimilarité ρ entre deux individus doit être égale à la moitié du carré de la distance euclidienne d. Le critère de Ward consiste à choisir à chaque étape le regroupement de classes tel que l'augmentation de l'inertie intra-classe soit minimale:

$$D(C1,C2) = \frac{n_{C1} \times n_{C2}}{n_{C1} + n_{C2}} d^2(g_{C1}, g_{C2})$$

Avec :
- g_{C1} : Le centre de gravité de C_1.
- g_{C2} : Le centre de gravité de C_2.

- **Le critère des centres de gravité**

La distance entre deux classes C1 et C2 est définie par la distance entre leurs centres de gravité.

$$D(C_1, C_2) = d(g_{C1}, g_{C2})$$

La difficulté du choix du critère d'agrégation réside dans le fait que ces critères peuvent déboucher sur des résultats différents. Selon les parts des références le critère le plus couramment utilisé est celui du Ward.

4.4. Classification d'opinions

La classification d'opinions est une technologie fondamentale en opinion-mining et en analyse des sentiments [Pang et al., 2008]. Selon [Hu et al., 2006], la classification des sentiments ou des opinions, classifie une critique ou un texte en positive ou négative. Beaucoup de chercheurs ont travaillés sur ce problème et ont utilisés entre autres :

- Des modèles inspirés de la linguistique cognitive ;
- Un lexique créé manuellement, et plusieurs fonctions de calcul de score, pour classifier les messages envoyés par les utilisateurs (cas de forum, blog, sites de e-commerces…etc.);
- Des techniques d'apprentissage non supervisées, basées sur l'information mutuelle entre les phrases d'un texte.

La majorité des travaux existant, focalisent sur la classification des critiques postés par les utilisateurs sur le Web. D'après [Hu et al., 2006], la classification d'opinions se diffère d'autres travaux, en deux aspects principaux

Il ne classifie pas toute la critique, mais la classification de chaque phrase dans la critique, qui contient une caractéristique d'un produit. Dans une critique, quelques phrases peuvent exprimer des opinions positives sur certaines caractéristiques et d'autres peuvent exprimer des opinions négatives sur d'autres caractéristiques ;

Peu de travaux existant, ont ressués d'identifier des caractéristiques des produits commentés par les utilisateurs. Cette information est très utile en classification d'opinions.

Pour aboutir cette tâche, les auteurs ont proposés de suivre les étapes suivantes :

(1) Extraction des informations à partir des textes

Il y a deux techniques principales :

- Les approches symboliques qui dépendent de la description syntaxique des termes ;
- Les techniques statistiques, utilisent les méthodes d'apprentissage automatiques.

(2) Identification des synonymes

La tâche d'extraction des caractéristiques des produits peut extraire un nombre important des noms de caractéristiques, et beaucoup d'entre eux sont synonymes, par exemples les mots en anglais « picture » et « image », dans des critiques sur une « caméra » ont le même sens et doivent être regroupés ensemble. Pour construire un résumé basé sur la description des caractéristiques des produits, il est crucial d'identifier et grouper les synonymes ensemble. Les méthodes existantes en extraction des synonymes et similarité lexicale se basent sur l'utilisation des dictionnaires, et le calcul de l'information mutuelle des paires des mots [Turney, 2001].

Notons que le regroupement des synonymes est une technique utilisée en représentation conceptuelle des textes (voir chapitre 3, section 3.3.2).

(3) Résumé de textes

Le résumé basé sur les caractéristiques des produits, employé par les auteurs, est lié à au résumé classique de textes. Cependant le résumé d'opinions est un résumé structuré au lieu de petites phrases. La majorité des travaux en résumé de textes, sont basés sur l'extraction des informations pertinentes pour construire un résumé. Les travaux ont été faits sur un ou plusieurs documents i.e. Construire un résumé d'un seul document ou plusieurs. Le résume est basé sur les critiques positives et les critiques négatives qui dépendent des caractéristiques des produits.

Dans [Généreux et al, 2009], les auteurs ont montré un système permettant de produire automatiquement des résumés de textes porteurs d'opinions, des blogs et des critiques de jeux notamment. Le système repose sur des techniques classiques en résumé : le calcul de différents traits reposant sur des heuristiques découlant de diverses expériences, et ils ont proposé de construire des résumés en identifiant les opinions, leurs orientations positives et négatives et si possible leurs intensités, et éventuellement calculer l'orientation générale de l'opinion sur un sujet donné.

4.4.1. Travaux de recherches

Nous présentons dans cette section, quelques travaux de recherche en classification d'opinions :

Dans [Généreux et al, 2009], et dans les expériences montrées, les auteurs ont classifiés les opinions en deux grandes classes, positives et négatives. Pour faire cette classification,

l'approche a été basée sur un classifieur binaire, reposant sur un apprentissage à partir de documents représentatifs préalablement annotés. Le classifieur essaie de repérer automatiquement les éléments pertinents pour une prise de décision (texte véhiculant une opinion positive *vs* texte véhiculant une opinion négative). Ils ont appuyés sur une machine à vecteur support (SVM) binaire entraînée sur des données typiques pour la tâche. Le but de cette étape est de outre la reconnaissance des opinions en elles-mêmes et de permettre ensuite le regroupement des phrases extraites en fonction de l'opinion exprimée et de tenir compte de la proportion d'opinions allant dans un sens ou dans l'autre, afin d'améliorer le rendu et la lisibilité du résumé.

Dans [Pang et al., 2008], les points suivants, ont été étudiés comme concepts fondamentaux en classification d'opinions :

- Polarité de sentiment et degré de positivité ;
- Détection de la subjectivité et identification d'opinions ;
- Analyse des sentiments (opinions) identifiés.

Dans [Liu, 2009], l'auteur a définie deux types de classification d'opinions :

- La tâche de classification de phrase comme opinioné (opinionated) ou non opinioné (not opinionated) est appelée *classification de subjectivité*.
- La phrase opinée, est ensuite classifiée en positive ou négative, cette tâche est appelée *classification de sentiments au niveau de phrase*.

Selon l'auteur, la majorité des techniques de classification de sentiments au niveau de documents, sont basées sur un apprentissage supervisé. Cependant, il y a des méthodes non supervisée.

(1) Classification basée sur un apprentissage supervisé

La classification des sentiments peut êtres évidement formalisée comme un problème de classification supervisée avec deux étiquettes : positive et négative. Apprentissage et test des données dans les travaux de recherche existants sont surtout des critiques de produits qui peuvent êtres positives ou négatives. Dans les forums, les sites de e-commerce…etc., un critique avec 4 à 5 est considéré comme positif (thumbs-up) et un critique de 1 à 2 étoiles est considéré comme négatif (thumbs-down).

La classification des sentiments est différente de la classification thématique des textes qui classifie les textes en des classes thématiques prédéfinies : politique, sciences, sport...etc.,

Les méthodes supervisées existantes peuvent êtres appelées en classification des sentiments : Naïve Bayes, Machines à support de vecteurs…etc.

(2) Classification basée sur un apprentissage non supervisée

Il n'est pas difficile d'imaginer que les mots et les phrases d'opinions sont les indicateurs dominants en classification d'opinions, ainsi, un apprentissage supervisé basé sur ces mots et phrases peut être assez naturel.

La méthode employée dans [Turney, 2002], utilise une classification basée sur l'utilisation phrases exprimant des opinions. L'algorithme a été basé sur l'extraction des phrases qui contiennent des adjectives et des adverbes. La raison de faire ça est que la recherche a démontré que les adjectifs et les adverbes sont des bons indicateurs pour la subjectivité et les opinions. [Généreux et al, 2009], [Hu, et al., 2005], [Maurel et al., 2007]. Cependant, des adjectifs indépendants peuvent indiquer la subjectivité, ce qui pose le problème du contexte insuffisant pour déterminer l'orientation de l'opinion.

Selon [Liu, 2006], la classification d'opinions est utile, mais ne détermine pas ce que les propriétaires d'opinions veulent et ce que ne veulent pas.

- Un sentiment négatif sur un objet ne signifie pas que le propriétaire d'opinion ne veut pas tout l'objet.

- Un sentiment positif sur un objet ne signifie pas que le propriétaire d'opinion veut tout l'objet.

4.4.2. Méthodes de classification

Dans [Maurel et al, 2007], les auteurs ont employés trois méthodes pour classifier les opinions extraites en positives et négatives :

- **Méthode symbolique**

Consiste à analyser syntaxiquement le texte et le découper en phrases, puis vérifier chaque phrase si elle contient des relations de sentiment en employant une grammaire spéciale, les relations syntaxiques de base ont été employés, comme les modifieurs des noms (Exemple : une *belle* maison), et les modifieurs des verbes (Exemple : lire *attentivement*), d'autres relations plus complexes ont été aussi employés, comme le sujet d'un verbe (Exemple : *Pierre* fait des courses), les relations de sentiment (Exemple : j'*aime* beaucoup Grenoble)…etc. Pour les sentiments positifs et négatifs, ont employés les adjectifs comme *magnifique*, *affreux*…etc., et les verbes comme *aimer*, *regretter*…etc. Pour les sentiments moyens, ont coordonnés des sentiments positifs et des sentiments négatifs (Exemple : un livre *passionnant* mais *inabouti*). D'autres mots clés ont étés aussi employés, comme *pourtant*, *malgré* (Exemple : *Malgré* un début superbe…). L'utilisation de l'inversion de la polarité dans le cas d'une négation (Exemple : Un restaurant *pas cher*).

L'opinion globale du texte est calculée à base des sentiments positifs, négatifs et moyens retenus pour chaque phrase qui sont mise en relation pour donner un sentiment global du texte entier.

- **Méthode statistique**

Basée sur des techniques de l'apprentissage automatique, et une classification au niveau des textes entiers. Le fonctionnement de cette approche est basé sur l'extraction des phrases qui contiennent des sentiments à l'aide de la méthode symbolique et sur l'entrainement des modèles des corpus sur les extraits des textes, et enfin la classification des textes.

D'après les auteurs, cette technique n'est plus efficace car difficilement reproductible sur d'autres corpus.

- **Méthode hybride**

Basée sur la comparaison des deux méthodes précédentes, et le résultat global est calculé d'après les indices de confiances attribués.

D'après les auteurs, et après tester ces méthodes sur trois corpus, la combinaison des méthodes symboliques et statistiques a donnée des résultats plus précis que chacune des méthodes employées séparément.

4.5. Critères pour une bonne classification

L'objectif principal des techniques de classification est de trouver une partition où les objets d'une classe devraient être semblables (entre eux), les objets de différentes classes devraient être différents. Une bonne classification devrait accomplir différents critères [Boubou, 2007]:

- Validité interne:

1. Chaque classe d'une partition doit être homogène: Les objets qui appartiennent à la même classe doivent être semblables.

2. Les classes doivent êtres isolés entre eux: les objets de différentes classes doivent être différents.

3. La classification doit s'adapter aux données: La classification doit pouvoir expliquer la variation des données.

- Interprétabilité:

Les classes doivent avoir une interprétation substantive: Il est possible de donner des noms aux classes. Dans le meilleur des cas, ces noms doivent correspondre aux types déduits d'une certaine théorie.

- Stabilité:

Les classes doivent être stables: Les petites modifications dans les données et dans les méthodes ne doivent pas changer les résultats.

- Validité externe:

Les classes doivent êtres valides (validité externe): Les classes doivent se corréler avec les variables externes qui sont connues pour être corrélées avec la classification et qui ne sont pas employées pour grouper.

- Validité relative:

La classification doit être meilleure que d'autres classifications.

- D'autres critères:

Parfois la taille et le nombre de clases sont employés en tant que critères additionnels: Le nombre de classes doit être aussi petit que possible. La taille des classes ne doit pas être trop petite.

4.6. Conclusion

Dans ce chapitre, nous avons présenté, plusieurs méthodes pour le problème de classification des documents. Elles diffèrent par les mesures de proximité qu'ils utilisent, la nature des données qu'ils traitent et les objectifs finals de la classification. Chacune de ces méthodes possèdes ses points forts et ses points faibles. Les méthodes hiérarchiques ascendantes sont utilisées en cas des données de petite taille car la complexité est très élevée. Si au contraire, des problèmes de temps d'exécution se posent, alors c'est les méthodes de k-means qui sont utilisées. Si l'objectif est de fournir des classes de forme quelconque, alors ce sont les méthodes basées sur la densité ou sur des grilles qui sont utilisées. C'est donc, le choix d'une méthode approprié dépend fortement de l'application, la nature des données et les ressources disponibles. Une analyse attentive des données aide à bien choisir le meilleur algorithme. Il n'existe pas un algorithme qui peut répondre à toutes les demandes.

Dans la section 4.4, contrairement à la classification thématique des documents, et quelque soit la méthode utilisée, la classification d'opinions classifie les opinions identifiés en trois grandes catégories : *positive*, *négative* et (probablement) *neutre*.

CHAPITRE 5
Notre système IOJAR

5.1. Introduction

L'information textuelle dans le monde peut être largement classifiée en deux catégories, faits et opinions.

1. Les faits sont des rapports objectifs sur des entités et des événements dans le monde.

2. Les opinions sont des rapports subjectifs qui reflètent les sentiments ou les perceptions des personnes sur des entités et des événements.

Une grande partie de la recherche existante en traitement de l'information (presque exclusivement) a été concentrée sur l'exploitation de l'information factuelle, par exemple, la recherche documentaire, la recherche sur le Web…etc. Peu de recherche a été faite en opinion-mining. Cependant, les opinions sont si importantes, toute fois qu'on veut faire une décision, on doit entendre d'autres opinions. Cela vaut non seulement pour les individus mais également pour les organismes [Breck, 2008].

Avant le Web, quand un individu doit prendre une décision, il demande typiquement les avis de ses amis et de sa famille. Quand une organisation doit trouver des opinions du général public sur l'un de ses produits ou sur l'un de ses services, elle recrute des groupes spécialisés pour cette raison.

Avec le Web, en particulier avec la croissance explosive du contenu écrit par l'utilisateur sur le Web, le monde a changé, dans des forums d'Internet, des groupes de discussion, et des blogs, l'utilisateur peut écrire son avis sur un produit ou un service.

Maintenant, si quelqu'un veut acheter un produit, il n'est plus nécessaire de demander l'avis de ses amis et de sa famille, il suffit de consulter un site ou un forum sur le Web pour voir les avis des utilisateurs existants du produit. Pour une compagnie, elle n'a plus besoin d'employer des conseillers externes pour trouver les avis des consommateurs sur ses produits et ceux de ses concurrents.

La recherche en opinion-mining, a commencée par identification des mots de roulement d'opinion (ou sentiment), par exemple, *grand*, *merveilleux*, *mauvais*, et *pauvre*...etc. Beaucoup de chercheurs ont utilisé de tels mots pour l'identification des orientations sémantiques (c.-à-d., positif ou négatif).

Dans le cadre de recherche d'information, et du point de vue du traitement automatique, l'identification d'opinions présente une tâche difficile à maitriser et cela dû à la complexité de savoir comment représenter une opinion ?, de quoi se compose une opinion ?, et comment extraire ses différents constituants à partir du texte ?. Notons qu'aucune approche n'a été proposée jusqu'à nos jour pour identifier une opinion, et tous les travaux existants dans ce domaine ont été limités sur l'extraction de l'orientation sémantique des expressions subjectives et la classification de ces orientations en deux catégories : positive et négative. [Généreux et al, 2009], [Hu, et al., 2005], [Maurel et al., 2007].

Nous présentons dans ce chapitre, notre approche proposée pour l'identification d'opinions, le modèle conceptuel utilisé pour représenter une opinion, la représentation XML utilisée pour stocker les opinions identifiés qui sera utilisé dans la phase de développement de notre système **IOJAR** (**I**dentification d'**O**pinions dans les **J**ournaux **AR**abes), et l'architecture proposée pour notre système, en détaillant chaque processus à part.

5.2. Identification d'une opinion

Dans ce paragraphe, nous présentons un modèle conceptuel pour représenter une opinion, basé sur 4 éléments : *prédicat*, *source*, *sujet* et *contenu*. Nous présentons aussi une représentation basée sur l'utilisation des fichiers XML pour stocker les opinions qui seront utilisées dans des étapes ultérieures de notre travail. L'utilité d'élaborer un modèle, est pour faciliter la représentation des opinions extraites, et de séparer les différents constituants d'une opinion, en vue de faciliter l'identification de l'orientation sémantique et l'intensité globale de l'opinion à partir des différentes orientations sémantiques et différentes intensités des éléments qui composent l'opinion. Notons que la présence d'une opinion dans le texte est fortement liée à la présence des expressions subjectives qui reflètent le jugement personnel de celui qui opine sur un sujet ou un objet. L'extraction de ces expressions nécessite une analyse syntaxique et un découpage des textes en utilisant les relations syntaxiques de base comme les verbes et les adjectifs. [Hu et al., 2005].

5.2.1. Modèle conceptuel

D'après [Rosá, 2008], l'opinion est conçue comme un objet conformé par plusieurs éléments, dont le principal est le *prédicat* (verbal ou nominal). Les autres éléments représentent les arguments de la prédication : la *source* (personne, document, publication…etc.) à laquelle on peut attribuer l'opinion, le *contenu* de l'opinion et le *sujet* sur lequel porte l'opinion. Schématiquement on peut représenter le modèle conceptuel d'opinion comme suit :

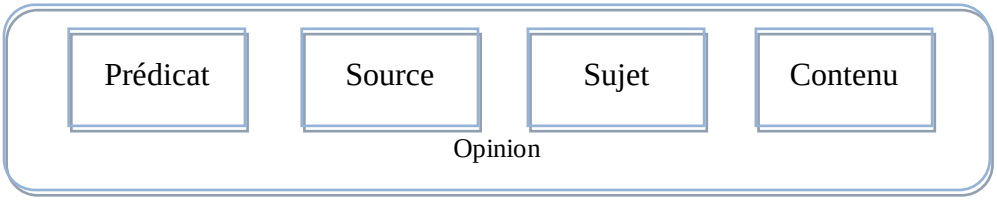

Figure 5.1 : Modèle conceptuel pour représenter une opinion

La définition des deux propriétés, *intensité* et *polarité*, inspirée des travaux de Mathieu [Mathieu, 2000], permet d'établir des relations entre les opinions d'un texte selon leurs différents degrés d'intensité (une opinion pourra être plus faible ou plus forte qu'une autre) et leur valeur de polarité (une opinion pourra être contraire ou pas à une autre). La valeur de ces propriétés dépend des éléments qui composent l'opinion, qui défini, donc, ces mêmes propriétés pour chaque élément de l'opinion. A partir des valeurs de polarité et d'intensité des éléments, les valeurs finales de polarité et d'intensité sont calculées, en appliquant les règles linguistiques proposées par [Ding et al., 2007].

En ce qui concerne la polarité, l'opinion peut être *neutre*, *positive* ou *négative*, par rapport au sujet sur lequel on opine. Il faut regarder à l'intérieur de tous les éléments présents pour calculer la valeur de polarité de l'opinion toute entière. Le prédicat (أكد, a réaffirmé) a une polarité positive et (رفض, a refusé) a une polarité négative, à partir de ces valeurs, on peut dire que l'opinion est positive pour le premier cas et négative pour le deuxième. Les prédicats (أجاب, a répondu), (صرح, a déclaré) et (قال, a dit) ont une polarité neutre.

Le tableau suivant illustre les polarités de quelques prédicats utilisés :

Prédicat	**Traduction**	**Polarité**
قال	a dit	Neutre
قرر	a décidé	+
أعلن	a énoncé	Neutre
ذكر	a rappelé	Neutre
صرح	a avoué	Neutre
أكد	a affirmé	+
رفض	a refusé	-
أوضح	a précisé	Neutre
كشف	a révélé	Neutre
نفى	a refusé	-

Tableau 5.1 : Polarités de quelques prédicats

Pourtant, pour une opinion dont, le *sujet* introduit par exemple par (ضد, *contre*) a une polarité négative, l'opinion sera donc négative.

D'autre part, pour un énoncé à l'intérieur de l'élément *contenu* on trouve le mot (مشجعة, *(encourageante)*, ايجابية (*positive),...etc.*), qui donne au contenu et à l'opinion complète une polarité positive ou le mot (ضعيفة (faible),سلبية (négative)…etc.), qui donne au contenu et à l'opinion complète une polarité négative. Exemples :

ذكر وزير التربية الوطنية أن نتائج البكالوريا لهذه السنة كانت ايجابية

أعلن الديوان الوطني لإنتاج الحبوب أن المحصول كان ضعيفا هذه السنة

On peut penser aussi à des éléments qui peuvent modifier la polarité de la *source*, par exemple, certains adjectifs comme (متحمسا, enthousiasmé),(مؤكدا, *affirmé*) dans la phrase, (... قال الرئيس مؤكدا, *le président a affirmé que* ...). Le tableau suivant illustre les polarités de quelques adjectifs utilisés :

adjectif	**Traduction**	**Polarité**
مؤكدا	affirmant	+
متحمسا	enthousiasmé	+
نافيا	refusant	-
مشيرا	indiquant	Neutre
معتبرا	considérant	Neutre

Tableau 5.2 : Polarités de quelques adjectifs

Les principaux travaux de recherche considèrent que l'orientation sémantique d'une opinion est exprimée par l'intermédiaire des adjectifs [Turney, 2002], [Taboada et al, 2006], [Voll et al,. 2002], [Hatzivassiloglou et al,. 1997], [Kamps et al, 2004]. Dans [Stavrianou, et al. 2008], il existe trois méthodes, pour identifier l'orientation sémantique:

- Manuelle → Effort important + coût ;
- Basé sur corpus → Dépendant du domaine ;
- Basée sur dictionnaire →Manque d'information contextuelle ;

Pour l'intensité, nous avons défini trois degrés : *faible*, *neutre* et *forte*. L'intensité de l'opinion dépend de tous les éléments qui la composent. Les prédicats (قال, dire), (أجاب, répondre) ont une intensité neutre, tandis que (أكد ,réaffirmer) a une intensité forte. (أشار, indiquer) ont une intensité faible. À l'intérieur de l'élément contenu il peut y avoir des éléments qui affectent l'intensité, par exemple on peut trouver (مشجعة جدا, *très encourageantes*). Dans ce cas, en plus d'une polarité positive, le contenu, et aussi l'opinion, auraient une intensité forte.

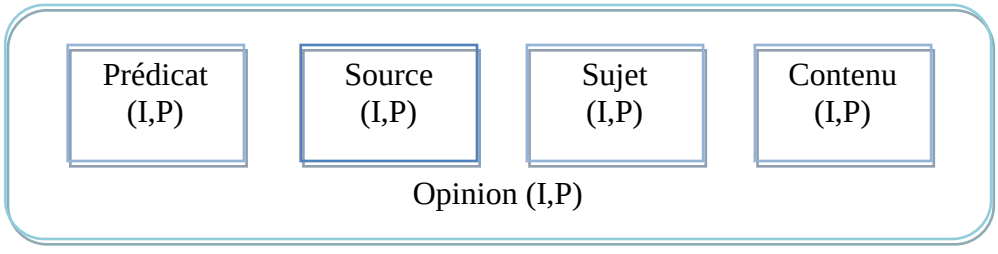

Figure 5.2 : Modèle conceptuel modifié pour représenter une opinion

Le tableau suivant illustre les éléments du modèle utilisé (Figure 5.1), en montrant quelques exemples extraits du corpus de travail. Dans les exemples, le *prédicat* est souligné, la *source* est marquée en gras, le *sujet* en caractère gras et souligné et le *contenu* est marqué en gris.

Le premier et le troisième exemple contiennent tous éléments qui intègrent l'opinion. Dans notre étude, nous avons remarqué que, le sujet et le *contenu* ne sont pas toujours présents.

	Texte	Traduction
1	أكد **مدير فرع الغاز المميع بمؤسسة نفطال**, أن الطلب على قارورات الغاز خلال الفترة التي امتدت من عيد الأضحى إلى غاية الأسبوع الماضي ارتفعت بـ 200 ألف قارورة. (El-Khabar, N° 5510 du 27/12/2008, page 6)	Le directeur de la direction générale de Naftal a affirmé que la demande pour les bouteilles de gaz pendant la période s'étendant de Aïd El-adha jusqu'à la semaine dernière ont augmenté de 200 mille bouteilles
2	قال **رئيس الوزراء البريطاني السابق توني بلير** في حوار ساخن مع أسبوعية بريطانية, ترجم في ما بعد إلى كثير من اللغات و تناقلته الصحافة الدولية بكثير من المفاجأة أو الصدمة ما يلي : "انه يقرأ القرآن يوميا و هو مهتم كثير بالإسلام". (Echourouk, N° 2488 du 25/12/2008, Page 24)	Le Premier ministre britannique, Tony Blair, dans une altercation avec un hebdomadaire britannique, par la suite traduit en plusieurs langues et transmis à la presse internationale avec beaucoup plus d'une surprise ou un choc comme suit: «Il lisait le Coran tous les jours et est plus intéressé à l'Islam».
3	أعلن **الدكتور علي زغدود الترشح للرئاسيات المقبلة**, ودعا المواطنين إلى مساندته بمنحه التوقيعات عن طريق لجان خاصة لمساندته. (El-Khabar,N° 5510, du 27/12/2008, page 5)	Dr. Ali Zaghdod a annoncé son candidature pour les prochaines élections présidentielles, et appelé les citoyens à contribuer à l'octroi de signatures par des commissions spéciales pour l'appuyer.
4	قال **محمد بوخطة, المكلف بالإعلام بمجلس ثانويات الجزائر**, أن القانون الجديد و ما جاء به من تغييرات لا يلبي طموحات العمال. (El-Khabar,N° 5510, du 27/12/2008, page 12)	Mohammad Boukhatta, en charge des écoles secondaires, a dis que la nouvelle loi et ce qui était permis par les changements ne répondaient pas aux aspirations des travailleurs…
5	اتهمت **النقابات المستقلة الخمس لمستخدمي قطاع الصحة**, أمس, الوزير الأول, أحمد أويحيى, بتجاهل مطالبهم المهنية و الاجتماعية, و نسيان معالجة ملف موظفي قطاع الصحة . (Echourouk, N° 2488 du 25/12/2008, Page 6)	Les cinq syndicats indépendants a accusé les utilisateurs du secteur de la santé, hier, le Premier ministre, Ahmed Ouyahia, en ignorant leurs demandes de professionnels et sociaux, et oublier le personnel de traitement des dossiers du secteur de la santé.
6	وفي ما يخص **الوضع في فلسطين** أعلن **المتحدث** علي ضرورة عقد جلسة طارئة	En ce qui concerne la situation en Palestine, le porte-parole a déclaré sur la nécessité de tenir une réunion d'urgence.

Tableau 5.3 : Exemples extraits du corpus

5.2.2. Représentation XML d'une opinion

Pour notre système d'identification nous avons défini, à partir du modèle conceptuel, un ensemble d'étiquettes pour indiquer la présence d'une opinion dans un texte sous la forme d'une annotation de type XML. Les éléments qui composent l'opinion sont représentés comme des éléments de XML et les propriétés qui caractérisent les éléments sont représentées comme des attributs de XML.

Pour un texte donné, le but du traitement est d'y incorporer les étiquettes pour marquer les opinions et les éléments qui les composent, ainsi que d´établir les valeurs pour les attributs intensité et polarité. Prenons l'exemple N°1 du tableau précédent, l'annotation XML correspondante à l'énoncé suivant :

أكد مدير فرع الغاز المميع بمؤسسة نفطال, أن الطلب على قارورات الغاز خلال الفترة التي امتدت من عيد الأضحى إلى غاية الأسبوع الماضي ارتفعت بـ 200 ألف قارورة

Est la suivante :

```
- <Opinion id="1" polarité="+" intensité="forte">
    <Prédicat chaine="أكد" />
    <Sujet chaine="" />
    <Source chaine="مدير فرع الغاز المميع بمؤسسة نفطال" />
    <Contenu chaine="الطلب على قارورات الغاز خلال الفترة التي امتدت من عيد الأضحى إلى غاية الأسبوع الماضي ارتفعت بـ 200 ألف قارورة" />
  </Opinion>
```

Figure 5.3 : Représentation XML d'une opinion

5.2.3. Extraction des éléments d'opinion

Dans ce paragraphe nous présentons notre approche pour l'identification des éléments et l'établissement des valeurs d'attributs. Dans cette étape, nous nous sommes consacrés à toutes les tâches de reconnaître le *prédicat*, le *sujet*, la *source* et le *contenu*.

Pendant notre travail, nous avons remarqué que tous les segments de textes porteurs d'opinions commencent généralement par un prédicat (exemple : أعلن, ذكر, صرح, أكد, قال… etc.).

Nous avons remarqué aussi que si la *source* est présente dans une opinion, il suit directement le prédicat, et cela dû à la syntaxe de la langue arabe (le sujet suit le verbe).

Exemples :

أعلن وزير الطاقة و المناجم…

ذكرت وكالة الأنباء الجزائرية...

Nous remarquons que le sujet est généralement un être humain ou un établissement, représenté par un mot ou une suite de mots.

D'après [Choi et al., 2005], L'identification des sources d'opinions sera une critique pour les systèmes Questions-Réponses orienté-opinion (par exemple, les systèmes qui répondent à des questions de la forme « comment [X] se sent environ [Y] ? »), et pour les systèmes de résumé orienté-opinion qui doivent distinguer les opinions d'une source de ceux des autres. La méthode CRF (Conditional Random Field) hybridée avec la méthode d'extraction des modèles (Extraction Patterns) a été employée dans [Choi et al., 2005] pour l'identification de la source d'opinion. Cette hybridation a donnée 79,3% de précision.

La méthode CRF et d'autres méthodes comme AutoSlog proposée par [Riloff, 1996] nécessitent trop de calculs et une analyse profonde des textes à traiter. Ces méthodes forment des sujets de travail indépendants, d'où, nous se limitons à la recherche symbolique des sources.

L'identification du contenu est généralement liée à la présence de certains marqueurs linguistiques comme (بأن, على ضرورة, أن...etc.).

Exemples :

أعلن وزير الطاقة و المناجم على ضرورة زيادة الإنتاج

...وأعتبر المتحدث أن القضايا التي كانت خلافية بين الإدارة و اللجان المتساوية الأعضاء حسمت لصالح الأولى.

...وأعتبر المتحدث أن القيادات النقابية لا أمل فيها.

Pendant notre travail, nous avons rencontré une difficulté d'identifier le *sujet* dans le segment porteur de l'opinion. On a remarqué que le sujet suit certains marqueurs linguistiques comme : ...في شأن, ...في ما يتعلق ب, ...في ما يخص, comme il peut être au début de segment porteur d'opinion, comme : ... : الوضع في العراق.

A partir de ce que nous avons vu, on peut construire une liste de règles pour identifier les éléments d'opinion:

Etant donné un texte T :

1. Segmenter le texte T, en utilisant un lexique de prédicats. Soit E, l'ensemble de segments.

2. Extraire les segments qui ont de caractère subjectif à partir E, en utilisant les relations syntaxiques de base (prédicats+adjectifs+adverbes). Soit E`, l'ensemble contenant les expressions à caractère subjectif.

3. Pour chaque segment de E`, identifier les éléments d'opinions en utilisant des marqueurs linguistiques, comme nous avons a précédemment mentionné.

En considérant que le texte arabe s'écrit de droite à gauche, vérifier le contexte gauche et droit de chaque élément dans les cas suivants :

- [Contenu] [Marqueur linguistique] <Source> [Prédicat]
- [Contenu] <Marqueur linguistique> <Source> [Prédicat] <Sujet>
- [Contenu] [Marqueur linguistique] <Source> [Prédicat] [Sujet] <Marqueur linguistique>

4. Identifier les attributs, c.-à-d. la polarité et l'intensité pour chaque élément, en utilisant toujours les relations syntaxiques de base comme les modifieurs des noms (les adjectifs) et les modifieurs des verbes (les adverbes).

5. Calculer la polarité et l'intensité globale de chaque opinion, à partir des polarités et intensités des éléments qui la compose.

5.3. Notre système d'identification

Dans cette section, nous allons présenter les différentes étapes que nous avons suivi pour concevoir notre système d'identification IOJAR, en expliquant le détail de chaque phase, en commençant par la constitution de corpus jusqu'à la classification des opinons identifiées.

5.3.1. Constitution du corpus

C'est à partir des journaux arabes que nous avons construit notre corpus de textes en langue arabe. Les textes, qui sont des articles de presse, ont été rapatriés sans restriction quant à leur nature ou leur volume. La raison en est que nous estimons que plus le corpus est étendu et varié, plus il sera représentatif et plus important sera le nombre de marqueurs d'opinions qu'il contiendra.

5.3.2. Architecture générale

La mise en œuvre fonctionnelle de notre système IOJAR repose sur plusieurs modules pour identifier et classifier les opinions figurantes dans les textes.

Les modules de notre système sont représentés en série, Nous décrivons dans ici toutes les tâches de chaque module dont IOJAR est composé et les différentes interactions entre les constituants de chaque module.

L'architecture générale que nous avons proposé pour notre système IOJAR est représentée par la figure suivante :

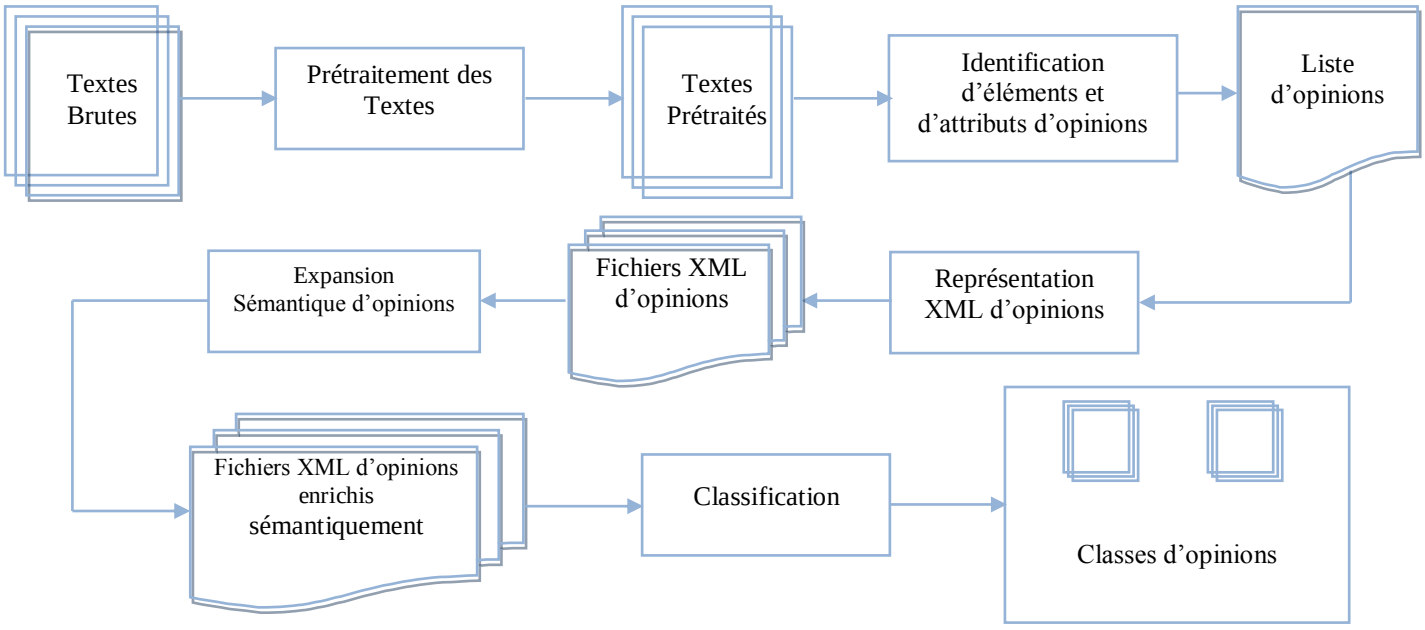

Figure 5.4 : Architecture générale de notre système IOJAR

Les modules suivants entre dans la composition de notre système :

- Prétraitement des textes
- Identification d'éléments et d'attributs d'opinions
- Expansion sémantique d'opinions
- Représentation XML d'opinions
- Classification

Dans cette section, nous allons détailler chaque module à part.

5.3.2.1. Prétraitement des textes

Comme on a mentionné dans le chapitre II, le prétraitement des textes est indispensable pour pouvoir extraire les informations pertinentes. Le prétraitement des textes nécessite des ressources externes comme le lexique d'abréviations et des exceptions pour la phase de normalisation, un anti-dictionnaire pour la suppression des mots vides, un dictionnaire de suffixes et préfixes pour la correction des fautes d'orthographe et les incohérences…etc. Nous détaillons dans ce paragraphe les tâches du module de prétraitement représenté par la figure suivante:

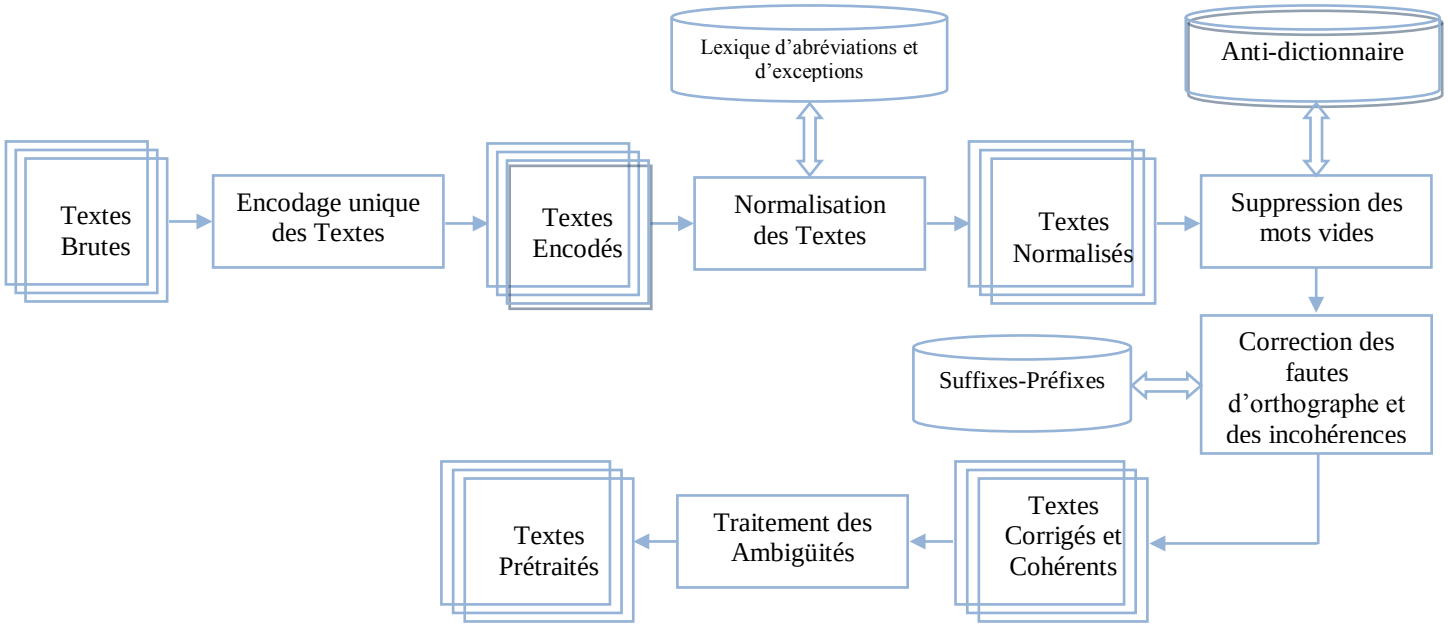

Figure 5.5 : Processus de prétraitement proposé

5.3.2.1.1. Encodage uniques des textes

L'encodage unique des textes en format standard, permet de représenter les textes sans aucune déformation au niveau de caractère lors de lecture.

Tous les textes de notre corpus sont représentés avec un encodage UTF-8[1].

5.3.2.1.2. Normalisation des textes

La normalisation, consiste à transformer une copie du document original dans un format standard plus facilement manipulable. Avant la lemmatisation, le document est normalisé comme suit [Bouzidia, 2004]:

• Suppression des caractères spéciaux et les chiffres

• Remplacement de آ , إ et أ avec ا

• Remplacement de la lettre finale ي avec ى

• Remplacement de la lettre finale ة avec ه

[1] UTF-8 (**U**CS **T**ransformation **F**ormat **8** bits) est un format de codage de caractères défini pour les caractères Unicode (UCS). (Pour plus de détail, consulter : http://fr.wikipedia.org/wiki/UTF-8).

Cette étape est nécessaire à cause des variations qui peuvent exister lors de l'écriture d'un même mot arabe. L'extraction se fait à partir du document original ce qui permet de préserver l'intégralité de l'information.

5.3.2.1.3. Suppression des mots vides

Consiste à éliminer tous les mots non significatifs. Pour chaque mot reconnu, nous le comparons avec un des éléments dans l'anti-dictionnaire (voir annexe 5) qui contient tous les mots non-significatifs. Si un mot en fait partie, il ne sera pas pris en considération.

5.3.2.1.4. Correction des fautes d'orthographes et des incohérences

La correction des fautes d'orthographes et des incohérences nécessitent un traitement particulier et une analyse profonde des textes. Dans notre travail nous ne s'intéressons pas beaucoup à cette phase, parce que les textes de notre corpus journalistique sont corrigés avant leur publication d'où les fautes d'orthographes et les incohérences sont rarement présentes dans les textes.

5.3.2.1.5. Traitement des ambigüités

Comme on a mentionné dans le chapitre 2, La désambiguïsation consiste à attribuer à chaque unité lexicale une étiquette unique (classe syntaxique et information morphologique de base) en contexte.

Le traitement des ambiguïtés nécessite une étude approfondie, nous se limitons dans notre travail à quelques techniques telles que la segmentation des textes et la détection des racines des mots que nous avons déjà représentés dans le chapitre 2, section 2.3.4.

5.3.2.2. Identification d'opinions

Après avoir prétraités les textes, nous arrivons à l'étape fondamentale de notre travail, c'est l'identification d'opinions en utilisant les règles mentionnées précédemment. Les opinions extraites sont stockés dans des fichiers XML, que nous allons les utiliser dans la phase de classification.

Nous avons commencé par l'extraction des segments porteurs d'opinions en effectuant une segmentation basée sur les marqueurs d'opinions (prédicats) qui indiquent la présence d'opinions dans les textes.

Ensuite, nous analysons les segments pour extraire les éléments et les attributs des opinions en appliquant toujours les règles mentionnées précédemment.

Enfin, nous regroupons les éléments d'opinions et leurs attributs sous forme de balises qui donne une représentation en format XML.

Remarquons, qu'un texte peut porter plusieurs opinions, d'où un fichier XML pour porter plusieurs représentations différentes.

La figure suivante illustre le processus d'extraction d'opinions :

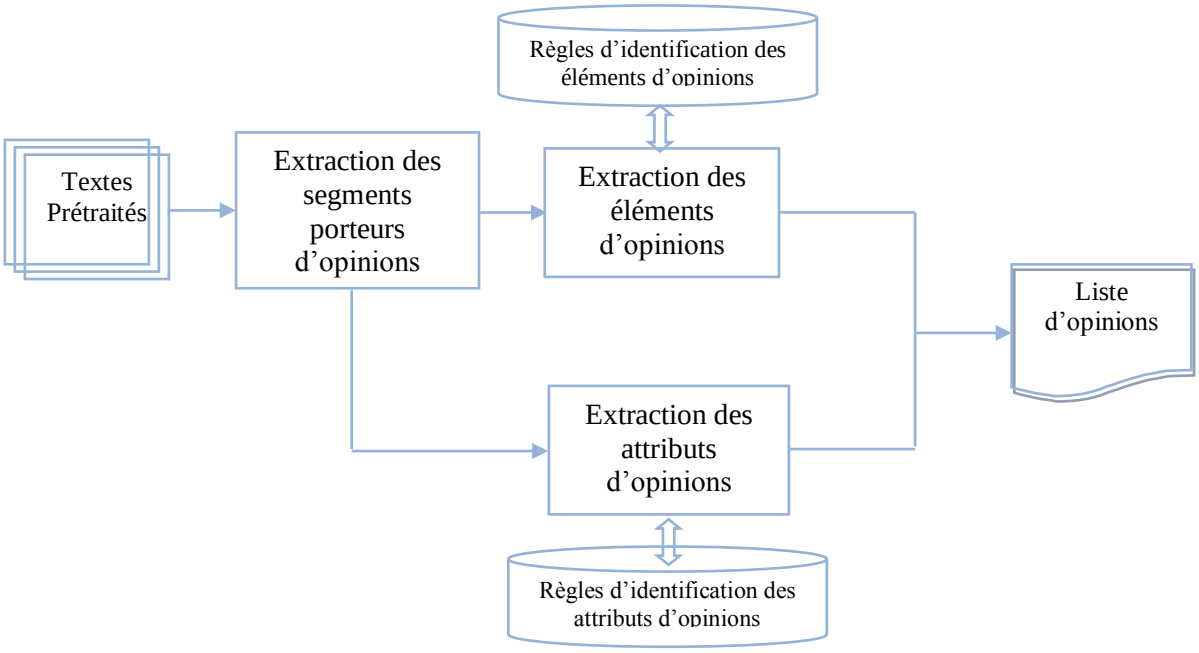

Figure 5.6 : Processus d'identification d'opinions

5.3.2.2.1. Extraction de segments porteurs d'opinions

Il s'agit de segmenter les textes en utilisant des marqueurs d'opinions. La présence d'un marqueur indique la présence d'une opinion dans le texte.

5.3.2.2.2. Extraction des éléments et d'attributs d'opinions

L'extraction des éléments d'une opinion : *prédicat*, *source*, *sujet* et *contenu*, est basée sur l'utilisation de règles mentionnées dans le paragraphe 5.2.3 de ce chapitre.

5.3.2.3. Expansion sémantique des textes d'opinions

Dans le chapitre 3, nous avons présenté une approche de représentation textuelle très moderne appelée ExpLSA (Expansion Latent Semantic Analysis), cette méthode consiste à enrichir sémantiquement les textes, avant d'appliquer la méthode LSA.

Cette phase nécessite une lemmatisation des mots pour faciliter l'expansion sémantique des textes.

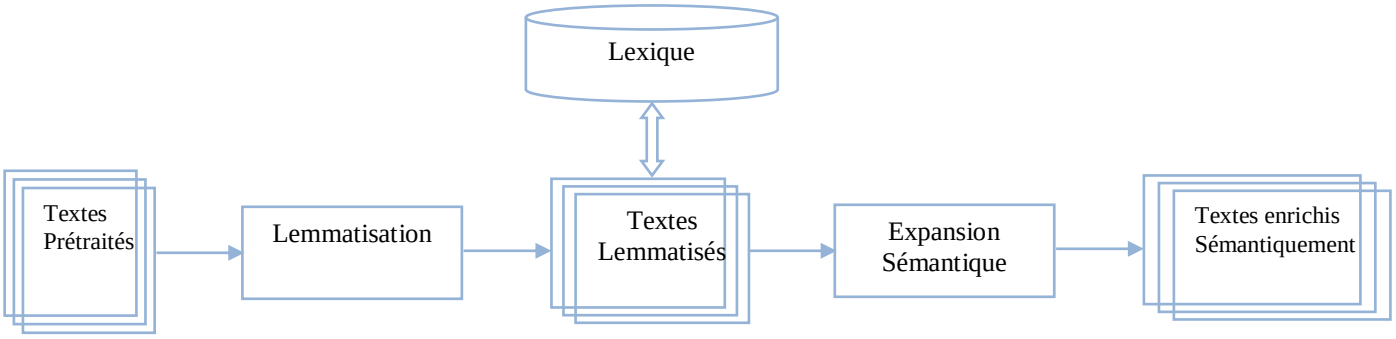

Figure 5.7 : Expansion sémantique des textes d'opinions

5.3.2.4. Classification d'opinions

Contrairement à la classification classique des textes, qui classifie des textes dans des classes de thèmes prédéfinies : politiques, scientifiques…etc. La classification d'opinions a été définie comme suit : Etant donné un ensemble de documents évaluatifs D, le système de classification doit déterminer si un document d ∈ D exprime une opinion (ou un sentiment) positive, négative ou neutre sur un objet [Hu et al., 2006]. Par exemple, étant donné un ensemble d'opinions sur un film, le système le classifie en opinions *positives*, *négatives* et *neutres*.

Dans la classification thématique des textes, les marqueurs linguistiques jouent un rôle très important pour déterminer le thème, tandis qu'en classification d'opinions ces marqueurs ne jouent aucun rôle, cependant, les mots qui expriment les polarités des opinions sont importants. Comme nous avons déjà mentionné dans le chapitre 1 de notre thème, et après l'étude comparative des trois méthodes existantes de classification d'opinions [Sigrid et al, 2007], nous trouvons que la méthode symbolique est plus adéquate que la méthode statistique, parce que cette dernière nécessite de développer de corpus pour chaque domaine, qui n'est pas possible pour notre cas.

La méthode symbolique, que nous avons adapté pour la classification d'opinons, se base sur l'analyse syntaxique des textes [Sigrid et al, 2007]. Contrairement aux méthodes de classification existantes, qui classifient les opinions en deux ou en trois grandes classes : positive et négative et probablement neutre, nous avons introduit l'attribut « intensité » dans la classification, l'attribut « intensité » combiné avec l'attribut « polarité », a nous aidé de proposer 5 classes au lieu de 3, et nous avons adapté les cinq classes:

(1) Positive forte

(2) Positivé faible

(3) Négative forte

(4) Négative faible

(5) Neutre

5.4. Conclusion

Dans ce chapitre, nous avons décrit notre système IOJAR, dont l'objectif est de concevoir un système capable d'identifier les opinions présentes dans des textes et plus précisément dans des textes journalistiques arabes. Les règles d'identification, que nous avons proposé sont extraites après une longue lecture d'un ensemble important de journaux arabes, et sont expérimentées ensuite manuellement sur un ensemble de textes journalistiques arabes dans notre corpus. Cette conception sera mise en fonction dans le chapitre qui suit.

CHAPITRE 6

Implémentation

6.1. Introduction

Nous présentons dans ce chapitre l'implémentation de notre système IOJAR, nous commençons tout d'abord par la présentation de l'environnement de développement, en détaillant les différents outils utilisés, puis nous expliquons le déroulement de l'application, et enfin nous interprétons et commentons les résultats obtenus.

6.2. Environnement de développement

Nous présentons dans cette section, le langage de programmation Java utilisé, et l'environnement de développement Eclipse.

6.2.1. Java

Java est un langage de programmation récent (les premières versions datent le 1995) développé par Sun MicroSystems. Il est fortement inspiré des langages C et C++.

Comme C++, Java fait partie de la grande famille des langages orientés objets. Il répond donc aux trois principes fondamentaux de l'approche orienté objet (POO) : l'encapsulation, le polymorphisme et l'héritage.

Java a rapidement intéressé les développeurs pour quatre raisons principales :

- C'est un langage orienté objet dérivé du C, mais plus simple à utiliser et plus « pur » que le C++. On entend par le « pur » le fait qu'en Java, on ne peut faire que la programmation orienté objet contrairement au C++ qui reste un langage hybride, c'est-à-dire autorisent plusieurs styles de programmation. C++ est hybride pour assurer une compatibilité avec le C ;

- Il est doté, en standard, de bibliothèques de classe très riches comprenant la gestion des interfaces graphiques (fenêtres, boites de dialogue, contrôles, menus, graphisme), la programmation multi-threads (multitâches), la gestion des exceptions, les accès aux fichiers et au réseau…L'utilisation de ces bibliothèques facilitent grandement la tâche du programmeur lors de la construction d'applications complexes ;

- Il est doté, en standard, d'un mécanisme de gestions des erreurs (les exceptions) très utile et très performant. Ce mécanisme, inexistant en C, existe en C++ sous forme d'une extension au langage beaucoup moins simple à utiliser qu'en Java ;

- Il est multi plates-formes : les programmes tournent sans modification sur tous les environnements où Java existe (Windows, Unix et Mac).

6.2.2. Eclipse IDE (Integrated Development Environment)

Eclipse est un environnement de développement intégré, libre, extensible, universel et polyvalent, permettant potentiellement de créer des projets de développement mettant en œuvre n'importe quel langage de programmation. Eclipse IDE est principalement écrit en langage java (à l'aide de la bibliothèque graphique SWT d'IBM), et ce langage grâce à des bibliothèques spécifiques, est également utilisé pour écrire des extensions.

La spécifié d'Eclipse vient du fait de son architecture totalement développé autour de la notion de plugin : toutes les fonctionnalités de cet atelier logiciel sont développé entant que plugin.

La base de cet environnement de développement intégré est *Eclipse Platform* composé de :

- *Platform Runtime* démarrant la plateforme et gérant les plugins.

- *SWT*, la bibliothèque graphique de base de l'IDE.

- *JFace*, une bibliothèque graphique de haut niveau basé sur SWT.

- *Eclipse Workbench*, la dernière couche graphique permettant de manipuler des composants, tels que des vues, des éditeurs et des perspectives.

Ces composants peuvent êtres réutilisés pour développer des clients lourds indépendants d'Eclipse grâce au projet Eclipse RCP (Rich Client Platform).

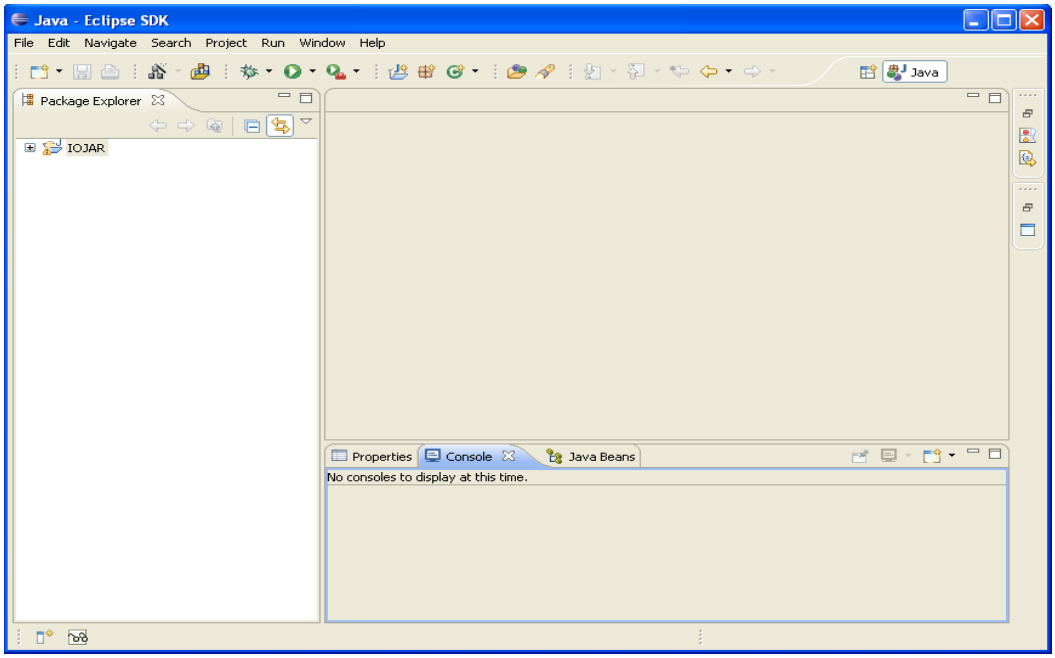

Figure 6.1 : Environnement Eclipse

6.3. Description de IOJAR

Notre système IOJAR développé en Java à l'aide de l'environnement Eclipse, il est muni d'une interface graphique conçue à l'aide de la bibliothèque SWT, disponible sur le lien http://www.eclipse.org/swt, et à l'aide de la bibliothèque JFace qui est basée sur SWT elle-même.

Nous avons développé plusieurs classes Java pour l'implémentation de notre systèmes, organisées sous forme de packages Java :

- Un package pour le prétraitement des textes ;

- Un package pour l'expansion sémantique des textes prétraités ;

- Un package pour l'extraction des expressions subjectives ;

- Un package pour l'identification d'opinions (éléments + attributs);

- Un package pour la classification.

6.4. Déroulement

Nous présentons dans cette section les différentes étapes de déroulement de notre système IOJAR, dès la sélection de textes jusqu'à la classification des opinions, en passant bien sûre par les étapes intermédiaires : prétraitement, extraction des expressions subjectives, et identifications des éléments et d'attributs d'opinions.

Figure 6.2 : Fenêtre de présentation de notre application

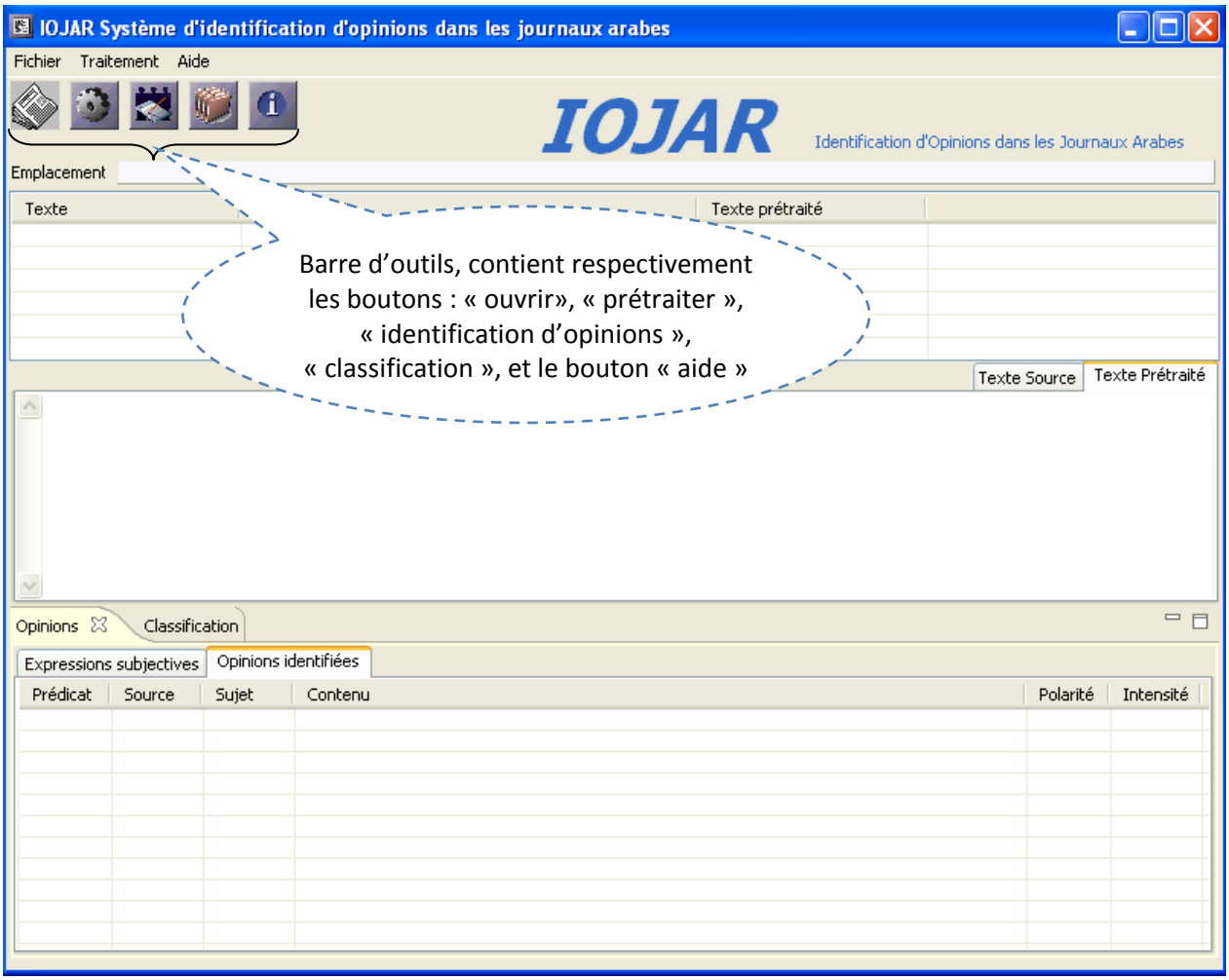

Figure 6.3 : Interface principale de notre application

6.4.1. Sélection des textes

Les textes qu'on a utilisés sont élaborés à l'avance et sont enregistrés en format « texte ». Comme on a déjà vu, ces textes sont collectés à partir d'un ensemble de journaux arabes couvrant plusieurs domaines : politique, social, économique…etc.

Nous système offre à l'utilisateur, la possibilité de sélectionner un répertoire contenant un ensemble de fichiers en filtrant ceux qui portent l'extension TXT, parce que ce format de fichiers est le plus simple à manipuler par rapport à d'autres formats comme DOC, RTF,…etc., qui nécessitent des éditeurs spécialisés pour visualiser leurs contenus.

La figure suivante, nous montre cette possibilité :

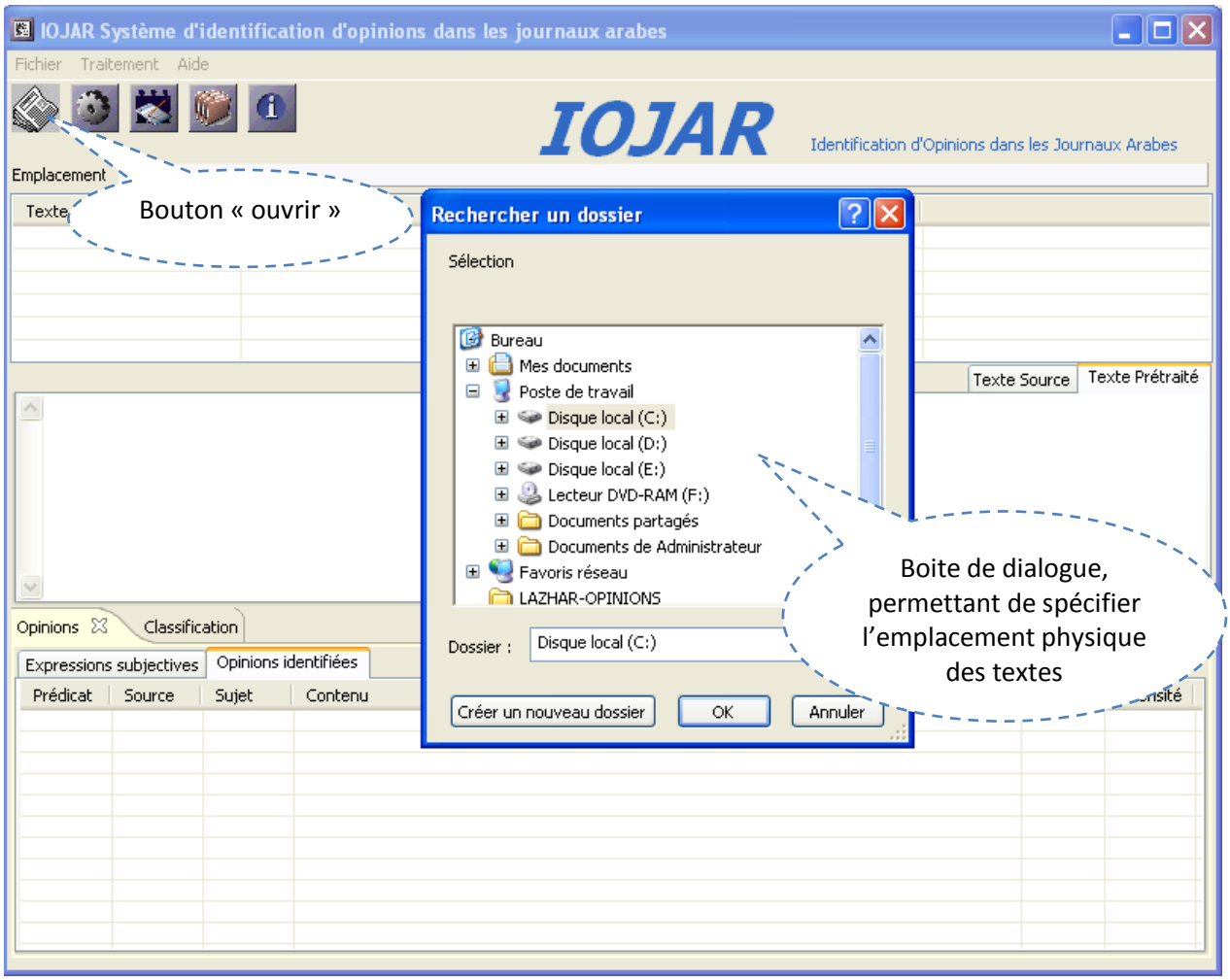

Figure 6.4 : Sélection de textes

Après sélection de textes avec l'utilisateur et confirmation avec le bouton «OK», la liste des textes s'affiche dans le tableau en haut de la fenêtre, avec la possibilité de limiter les textes sur lesquels on va travailler, en cochant ou décochant les cases correspondantes, le contenu du texte s'affiche automatiquement dès qu'on sélectionne une ligne sur le tableau comme il est indiqué dans la figure suivante .

Pour visualiser correctement les caractères arabes, nous avons développé une méthode Java, permettant de convertir l'encodage actuel du texte en encodage UTF-8, en effectuant la conversion caractère par caractère sans aucun changement morphologique.

Figure 6.5 : Visualisation des textes sources

6.4.2. Prétraitement des textes

Le module de prétraitent que nous avons indiqué dans le chapitre 5, est un pré-requis pour notre travail, et comme ce module nécessite beaucoup de ressources externes (lexiques d'abréviation et d'exceptions, anti-dictionnaire, liste des suffixes et des préfixes…etc.) qui dépasse souvent le cadre de notre sujet, nous se limitons dans notre travail a l'encodage unique des textes en format UTF-8 et à leur normalisation, en supposant que les fautes d'orthographe sont manuellement corrigées.

Notre système offre la possibilité d'effectuer cette phase, et l'affichage des textes prétraités. La figure suivante, indique cette possibilité :

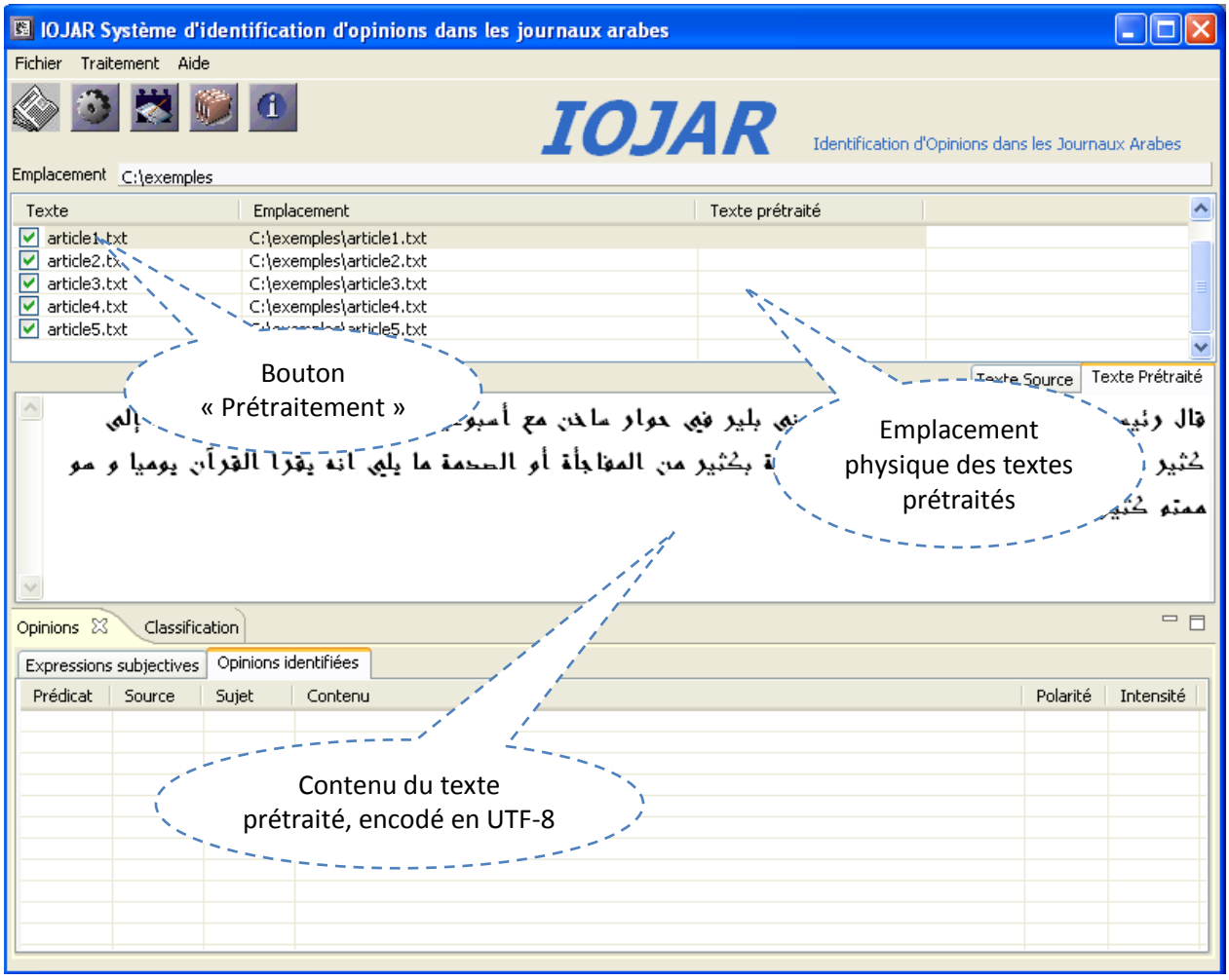

Figure 6.6 : Visualisation des textes prétraités

6.4.3. Identification des opinions

Dans cette étape, nous arrivons à l'étape d'extraction des éléments et d'attributs d'opinions présentes dans les textes.

Pour simplifier cette tâche, nous avons la diviser en sous-tâches :

- La première consiste à extraire les segments porteurs d'opinions i.e. les expressions subjectives, ces segments commencent généralement par des prédicats comme : قال (a dit), قرر (a décidé), رفض (a refusé)…etc.

- La deuxième consiste à identifier les éléments de chaque opinion (Prédicat, Source, Sujet et Contenu) représentée par une expression subjective.

- La troisième tâche pour identifier les attributs des opinions identifiée (polarité et intensité).

L'identification des éléments et d'attributs est basée sur les règles indiquées dans le chapitre 5 (section 5.2.3).

6.4.3.1. Extraction des segments porteurs d'opinions

Comme nous avons indiqué, les segments porteurs d'opinions commencent généralement par un prédicat, par exemple :

أشار المتحدث … (Le parlant a indiqué…)

Comme il peut commencer par un marqueur indiquant la présence du sujet par exemple :

…في ما يخص نتائج الانتخابات الرئاسية, قال المتحدث… (…et dans ce qui concerne les élections présidentielle, le parlant a dit…). La figure 6.7, illustre cette tâche :

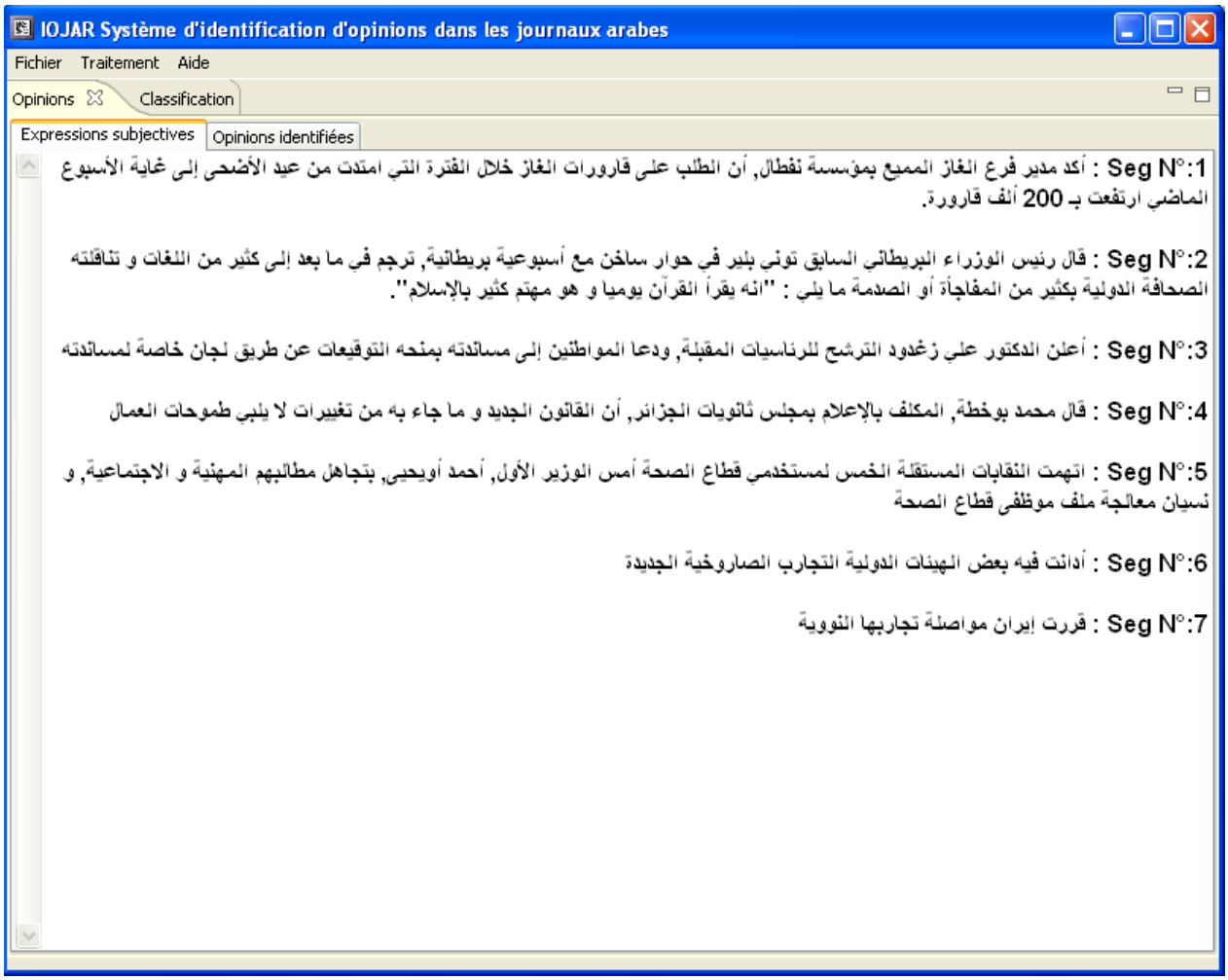

Figure 6.7 : Liste des segments porteurs d'opinions

6.4.3.2. Identification des éléments d'opinions

En cliquant sur l'anglet « Liste », nous visualisons la liste des opinions présentes dans les textes.

La figure suivante, explique cette tâche :

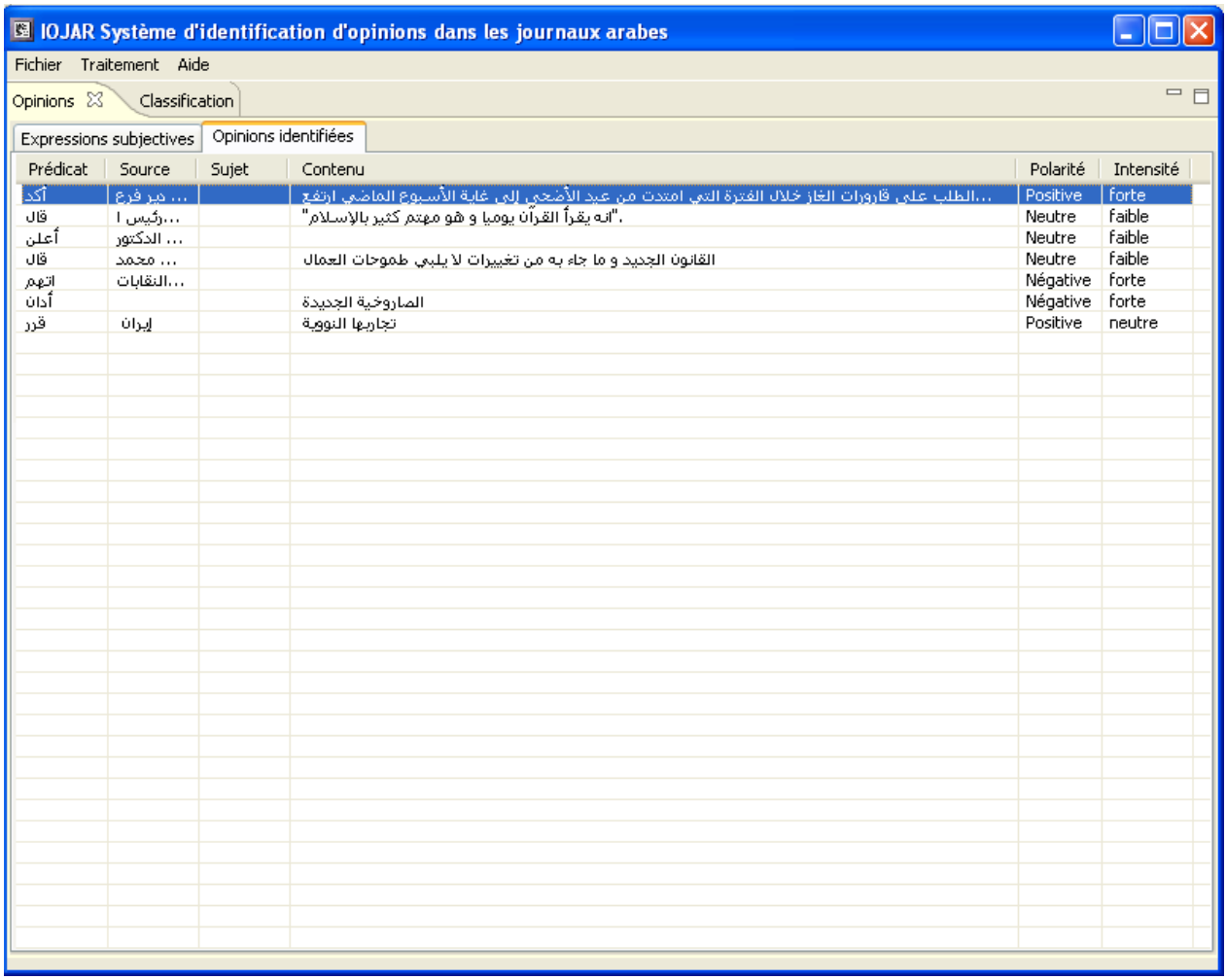

Figure 6.8 : Liste des opinions identifiées

6.4.3.3. Représentation XML d'opinions identifiées

Un moyen simple pour stocker les opinions extraites, est l'utilisation du langage XML. Les éléments qui composent l'opinion sont représentés comme des éléments de XML et les propriétés qui caractérisent les éléments sont représentées comme des attributs de XML.

Pour visualiser le contenu de fichier XML, il suffit d'utiliser un navigateur Web. Voir figure 6.9.

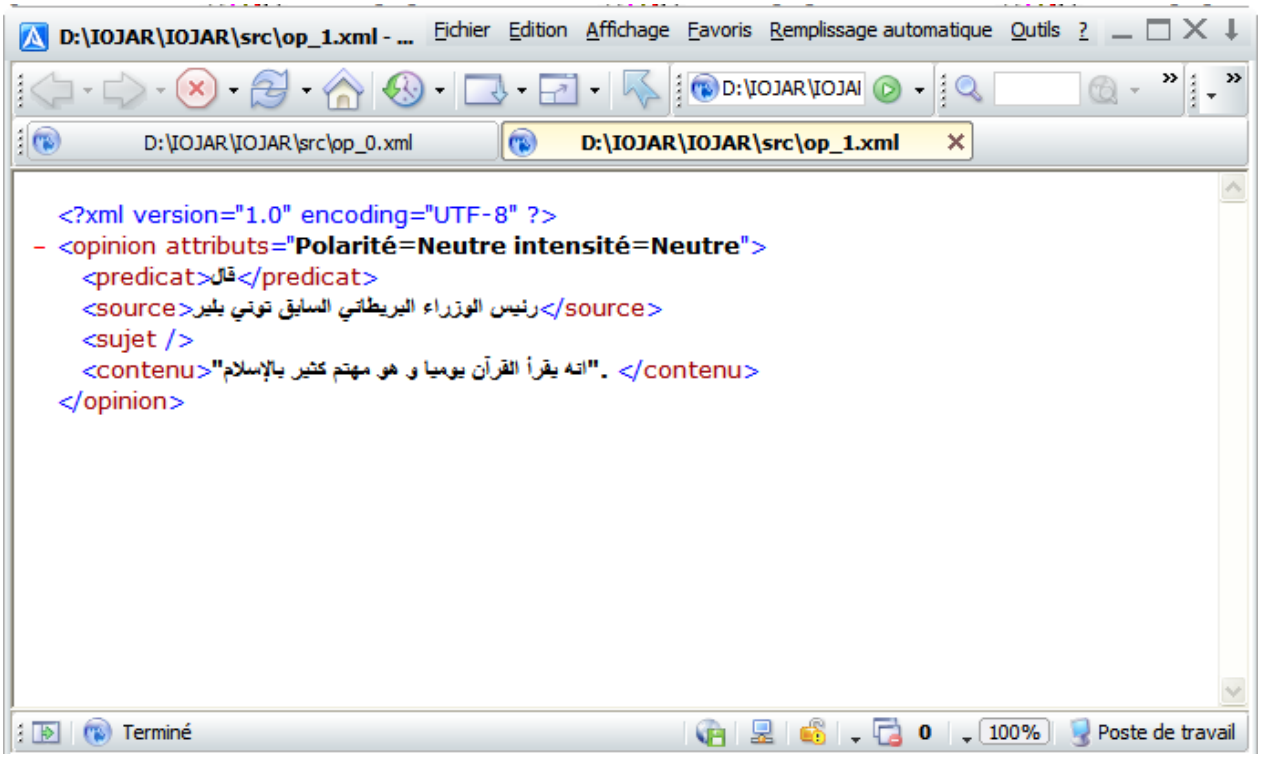

Figure 6.9 : Représentation XML d'une opinion

6.4.4. Expansion sémantique de textes d'opinions

Cette opération consiste à enrichir les textes d'opinions, en ajoutant pour chaque mot utile (en excluant les chiffres, dates, caractères spéciaux…etc.), les mots de même contexte qui sont sémantiquement proches, le tableau suivant illustre l'expansion de quelques mots :

Mot	Expansion
رئيس	وزارة وزير رئيس
حوار	خطاب حديث حوار
جريدة	صحافة يومية أسبوعية جريدة

Tableau 6.1 : Expansion sémantique de quelques mots

6.4.5. Classification des opinions

Nous arrivons à la dernière étape dans notre application, la classification d'opinions identifiés selon leurs orientations sémantique en trois catégories : positives, négatives et neutres.

C'est-à-dire regroupement des expressions subjectives porteuses de ces opinions en trois classes possibles. La figure suivante nous montre cette tâche.

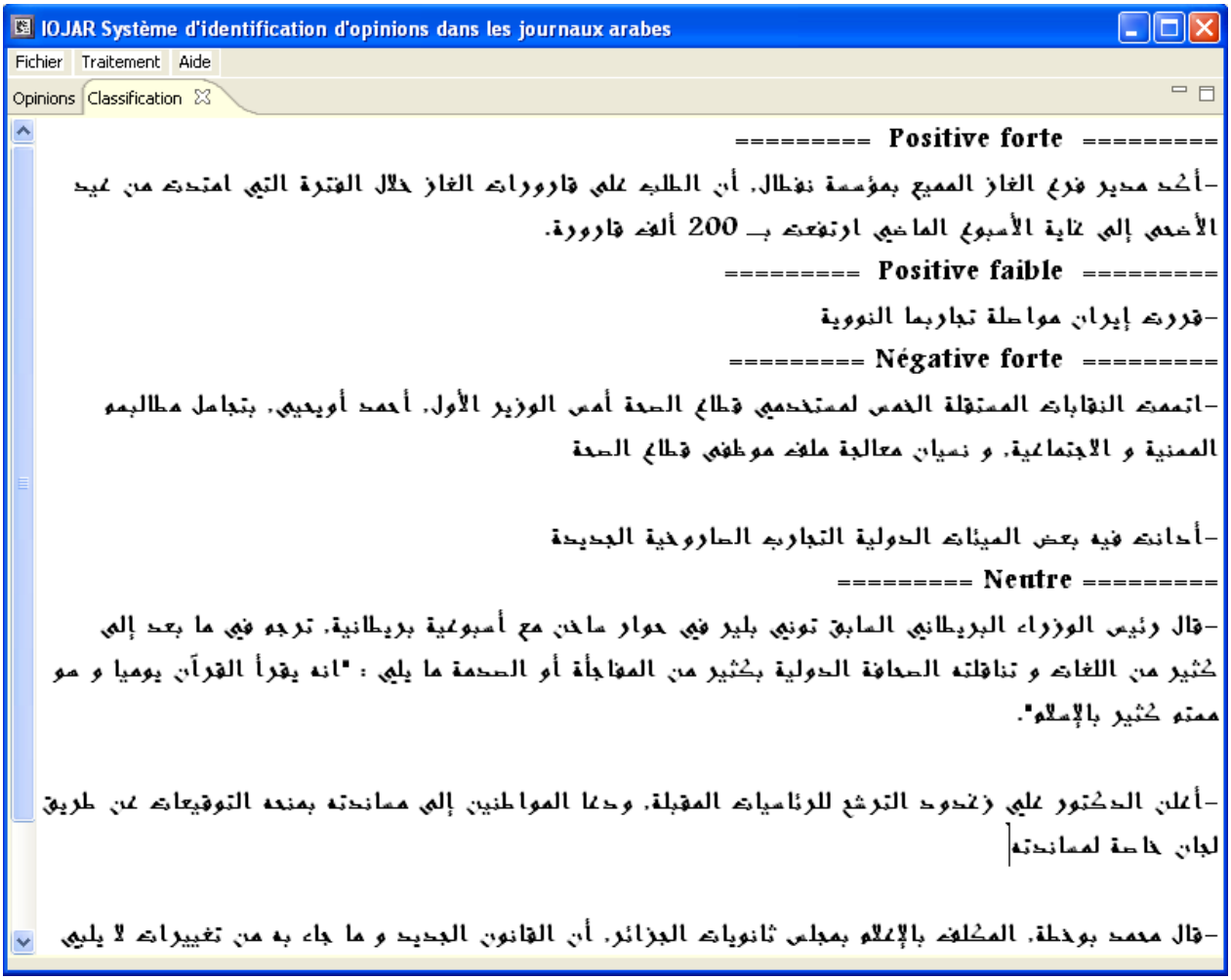

Figure 6.10 : Classification des opinions

6.5. Interprétation des résultats

La mise en test de notre approche proposée pour l'identification d'opinons, basée sur l'exploration symbolique des textes, a donnée des résultats non décevants malgré les difficultés rencontrés lors de l'extraction des éléments d'opinons comme la sujet d'opinion qui nécessite un travail particulier. Dans notre jeu de test, nous avons obtenu les résultats suivants :

- Dans plus de 95% des cas, la source d'opinion est identifiée s'il est lié un prédicat verbal comme : قال, أشار, …etc. Nous trouvons des cas comme : …كما أشار سابقا, … قال ذلك في إحدى تصريحاته السابقة…,etc. où notre système identifie ce genre d'expressions

subjectives sans identifier la source d'opinion parce qu'elle se trouve dans une autre expression subjective, dont on cannait pas son emplacement exacte dans le texte.

- Difficulté d'identifier la source d'opinion, s'il est lié un prédicat nominal, comme : ... حسب التقرير, ... جاء على لسان,...etc. Ce genre d'expressions présente des cas particuliers qui nécessitent une analyse un peu spéciale.

- Nous n'avons pas arrivé à identifier le sujet d'opinion : pendant notre étude, nous avons remarqué que l'identification de cet élément nécessite les techniques de résumé automatique de textes, qui n'est pas notre sujet de recherche.

- Nous avons arrivé à extraire tous les segments représentant les contenus d'opinions identifiés, parce qu'on a lié la présence du contenu à la présence des connecteurs linguistiques mentionnés précédemment.

- L'orientation sémantique d'une opinion est calculée en fonction de polarités de ces constituants (prédicat, source, sujet, contenu). Nous avons arrivé identifié les polarités de ces éléments avec un taux de 100%, parce que ces polarités sont souvent exprimées avec simples adjectives et adverbes. Les règles linguistiques de [Ding et al., 2007], nous ont aidées à déterminer les polarités des relations syntaxique de base (adjectifs, adverbes) par rapport à d'autres déjà identifiées.

- L'approche symbolique de classification d'opinions qu'on a utilisée pour classifier les opinions identifiée est inspirée de la méthode de [Maurel et al, 2007]. Cette dernière est basé sur l'analyse des sentiments, on a l'adapté pour la classification d'opinions, et a donnée les résultats souhaités.

6.6. Conclusion

Dans ce chapitre, nous avons présenté l'implémentation de notre système d'identification d'opinions dans les journaux arabes, ce système a pour rôle d'extraire les opinions présentes dans les textes et de les classifier en catégories.

Les résultats de notre système ont été liés au choix des techniques utilisées, ainsi aux règles d'identification des éléments et d'attributs d'opinions que nous avons proposé dans notre approche. La bonne préparation des textes (prétraitement), la division de processus d'identification en plusieurs modules, l'utilisation de la méthode d'expansion sémantique des textes (ExpLSA), sont des facteurs indispensables, qui ont participé dans l'amélioration de la performance de notre système.

Conclusion et perspectives

L'identification d'opinions a pour but de détecter les opinions inclues dans des textes. Ces opinions sont cachées derrière des mots, des phrases et des documents. Une expression d'opinion est l'unité la plus petite à partir de laquelle les opinions sont identifiées. Les mots exprimant des sentiments, le propriétaire d'opinion (holder), et les informations contextuelles sont les indices dans l'identification des expressions d'opinions et dans la détermination de leurs tendances. Donc, la démarche d'identification est basée, en premier lieu, sur l'extraction des mots de sentiments, et ensuite, l'identification des polarités (tendances) des expressions porteuses d'opinions. Les opinions identifiées sont utiles pour les individus, les organismes et même pour les gouvernements, qui sont appelés consommateurs d'opinions.

Notre travail a pour objectif d'identifier les opinions présentes dans des textes, et plus particulièrement dans des textes journalistiques arabes. Ces textes véhiculant une masse importante d'opinions dans un format brut, d'où la fouille de ces textes nécessite un temps et un coût élevé pour extraire ce qui est pertinent. Nous avons considérée qu'une opinion est représentée par un ensemble d'éléments: prédicat, source, sujet et contenu, et à chacun de ces élément, on a ajouté deux attributs : polarité et intensité, pour calculer l'orientation sémantique de chaque composant. L'orientation globale de l'opinion est ensuite calculée en fonction de tous ses éléments.

L'approche que nous avons proposée pour d'identification et la classification d'opinions est purement symbolique. Elle est basée sur l'extraction des expressions subjectives à partir des textes. Ces expressions expriment les points de vue de leurs propriétaires (opinions holders) et leurs orientations sémantiques : positives, négatives et neutres. Un lexique de prédicats, d'adjectifs et d'adverbes a été utilisé pour cette raison. Les opinions identifiées, sont ensuite classées selon leurs orientations sémantiques en cinq catégories : positive forte, positive faible, négative forte, négative faible et neutre.

Notre proposition a rencontré quelques difficultés dues à la complexité syntaxique de la langue arabe, qui a posé des problèmes au niveau d'analyse symbolique des textes sur laquelle se base notre approche, ce qui a nous laissé de réviser plusieurs fois les règles d'identification que nous avons proposé. Cette complexité syntaxique a influencé considérablement l'extraction de certaines classes syntaxiques, ce qui a nous demandé d'appliquer certaines règles linguistiques bien définies. Le prétraitement des textes d'opinions et l'application de l'approche ExpLSA (Expended Latent Semantic Analysis) pour l'enrichissement sémantique des textes ont démontré leur nécessité. La difficulté d'extraction du sujet d'opinion est fortement liée à l'absence d'application des techniques de résumé de textes, qui déborde le cadre de notre travail.

Notre approche qui a donné des bons résultats, a un coût d'entée élevé. Cette considération est liée au temps de configuration, de repérage et de la création de lexiques spécifiques. Dans nos futurs travaux, nous souhaitons adapter les techniques utilisée en fouille de textes, pour formaliser des algorithmes plus efficaces applicables sur des sources d'opinions hétérogènes pour l'identification et la classification d'opinions.

ANNEXE 1

Quelques propriétés linguistiques de la langue arabe

a) C'est une langue flexionnelle

Par exemple, du verbe كسر *kasara* signifiant « casser », ses dérivés, par doublement de la consonne « s », le verbe *ka**ss**ara* كسَّر « casser en mille morceaux », et par ajout du préfixe « in », le verbe ***in****kasara* انكسر « se casser ».

.

b) C'est une langue cliticisante (procliticisante et encliticisante)

Exemple : أكل → وليأكلوها (et pour la manger). Proclitique : ولِ , enclitique: لوها

c) C'est une langue à ordre des mots mixtes

Exemple:

- ضرب زيد الأولاد (Zayd a frappé les enfants).
- الأولاد ضربهم زيد (Les enfants, ont été frappé par Zayd).

d) C'est une langue pro-drop

Le verbe s'accorde néanmoins en personne, genre et nombre avec le pronom omis, comme le montre l'exemple suivant :

- أكلوا (Ils ont mangé).
- أكلن (Elles sont mangé).

e) C'est une langue parataxique

Dans l'exemple suivant, la proposition يأكلون (*Ils mangent*) remplit la fonction d'un coprédicat objet:

رأى زيد الأولاد يأكلون (Zayd a vu les enfants mangent).

ANNEXE 2

Liste de quelques prédicats

أضاف

أعلن

أفاد

أكد

أورد

أوضح

بين

تحدث

ذكر

شرح

صرح

فسر

قال

قرر

كشف

ANNEXE 3

Liste de quelques adverbes et adjectifs

مؤكدا

موضحا

مشيرا

مشددا

رافضا

منددا

مستنكرا

ايجابية

سلبية

مشجعة

مخيبة

مرضية

ANNEXE4

Liste des journaux

Journal	Site Web
الخبر	www.elkhabar.com
الشروق	www.echourouk-online.com
أخر ساعة	www.akher-saa.net
النصر	www.ennsasar.net
النهار	www.ennahar.com

ANNEXE 5

Liste de quelques mots non significatifs

!
"
#
$
%
&
'
(
)
*
+
,
.
/
{
|
}
~
[
\
]
^
_
`
:
;
<
=
>
?
@
0
1
2
3
4
5
6
7
8
9
A
B
C
D

E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z
a
b
c
d
e
f
g
i
j
k
l
m
n
o
p
q
r
s
t
u
v
w
x
y
z

ANNEXE 6

Algorithmes de classification supervisée

- **K plus proches voisins et ses améliorations**

Plus connus en anglais sous le nom K-nearest neighbor (K-NN), ou encore Memory Based Reasoning.

Cette méthode diffère des traditionnelles méthodes d'apprentissage car aucun modèle n'est induit à partir des exemples. Les données restent telles quelles : elles sont simplement stockées en mémoire. Pour prédire la classe d'un nouveau cas (où ranger un nouveau document ?), l'algorithme cherche les K plus proches voisins de ce nouveau cas et prédit (s'il faut choisir) la réponse la plus fréquente de ces K plus proches voisins. La méthode utilise donc deux paramètres : le nombre K et la fonction de similarité pour comparer le nouveau cas aux cas déjà classés.

- **Arbres de décisions**

Les arbres de décision sont les plus populaires des méthodes d'apprentissage. Ils sont également populaires pour la classification de documents.

Comme toute méthode d'apprentissage supervisée, les arbres de décision utilisent des exemples. Si l'on doit classer des documents dans des catégories, il faut construire un arbre de décision par catégorie. Pour déterminer à quelle(s) catégorie(s) appartient un nouveau document, on utilise l'arbre de décision de chaque catégorie auquel on soumet le document à classer. Chaque arbre répond Oui ou Non (il prend une décision).

Concrètement, chaque nœud d'un arbre de décision contient un test (un IF...THEN) et les feuilles ont les valeurs Oui ou Non. Chaque test regarde la valeur d'un attribut de chaque exemple. En effet, on suppose qu'un exemple est un ensemble d'attributs/valeurs. Pour des documents, chaque attribut peut être un mot, et la valeur sera par exemple 0 ou 1 selon que ce mot appartient ou non au document.[Site2, 2008].

Pour construire l'arbre de décision, il faut trouver quel attribut tester à chaque nœud. C'est un processus récursif. Pour déterminer quel attribut tester à chaque étape, on utilise un calcul statistique qui détermine dans quelle mesure cet attribut sépare bien les exemples Oui/Non. On crée alors un nœud contenant ce test, et on crée autant de descendants que de valeurs possibles pour ce test.

- **Naïve Bayes (ou Simple Bayes)**

Nommés d'après le théorème de Bayes, ces méthodes sont qualifiées de « Naïve » ou « Simple » car elles supposent l'indépendance des variables. L'idée est d'utiliser des conditions de probabilité observées dans les données. On calcule la probabilité de chaque classe parmi les exemples. Ce sont les « prior probabilities ». Par exemple, si la classe

« informatique » revient 2 fois sur les 5 documents donnés en exemple, sa « prior probability » sera de 2/5. En plus des "prior probas", l'algorithme calcule les fréquences d'apparition de chaque variable d'entrée avec celles de sortie. Pour classer des documents, les variables d'entrée sont les mots présents dans l'ensemble des documents. A chaque mot on calcule le nombre de fois qu'il apparaît dans les documents classés dans une classe donnée. On calcule cette fréquence pour chaque classe.

Une variante des Naïve Bayes sont les réseaux Bayésiens : dans ce modèle, on ne suppose plus que les variables sont toutes indépendantes, et on autorise certaines à être liées. Cela alourdit considérablement les calculs et les résultats n'augmentent pas de façon significative. [Site2, 2008].

- **Réseaux de neurones**

Les réseaux de neurones sont utilisés pour leur capacité à apprendre à partir d'exemples bruités comme les caméras ou les micros (reconnaissance de forme ou de son). Mais ils sont aussi utilisables pour des problèmes où les méthodes symboliques (arbres de décisions) sont souvent utilisées. Leur performance est alors équivalente.

Les réseaux de neurone sont appropriés lorsque le temps d'apprentissage n'est pas essentiel : ce temps est en effet souvent très supérieur à d'autres méthodes comme les arbres de décision. Par contre, la classification d'un nouveau cas (par exemple un document) est très rapide.

Enfin, les réseaux de neurones sont appropriés si la compréhension de la fonction apprise par le réseau n'est pas essentielle. Avec un arbre de décision, l'opérateur humain peut toujours visualiser l'arbre et « comprendre » comment la machine décide. Avec un réseau de neurone, des techniques de visualisation existent, mais elles demandent généralement plus d'expertise que l'analyse d'un arbre de décision (qui peut être visualisé sous forme de règles). [Site2, 2008].

- **Machines à support de vecteurs (ou SVM)**

Cette technique - initiée par Vapnik - tente de séparer linéairement les exemples positifs des exemples négatifs dans l'ensemble des exemples. Chaque exemple doit être représenté par un vecteur de dimension n. La méthode cherche alors l'hyperplan qui sépare les exemples positifs des exemples négatifs, en garantissant que la marge entre le plus proche des positifs et des négatifs soit maximale. Intuitivement, cela garantit un bon niveau de généralisation car de nouveaux exemples pourront ne pas être trop similaires à ceux utilisés pour trouver l'hyperplan mais être tout de même situés franchement d'un côté ou l'autre de la frontière. L'efficacité des SVM est supérieure à celle de toutes les autres méthodes sur la classification de textes. Son efficacité est aussi très bonne pour la reconnaissance de formes. Un autre intérêt est la sélection de Vecteurs Supports qui représentent les vecteurs discriminant grâce auxquels est déterminé l'hyperplan. Les exemples utilisés lors de la recherche de l'hyperplan ne sont alors plus utiles et seuls ces vecteurs supports sont utilisés pour classer un nouveau cas. Cela en fait une méthode très rapide. [Site2, 2008].

- **Programmation génétique**

C'est une méthode générale qui peut être utilisée après n'importe qu'elle méthode précédente, par exemple avec les arbres de décisions. En entrée, un algorithme génétique reçoit une population de classifieurs non optimaux. Le but du programme génétique est de produire un classifieur plus optimal que chacun de ceux de la population d'origine. D'une façon simple, cela consiste à extraire les meilleures parties de chaque classifieur d'origine et de les mettre ensemble pour produire un nouveau classifieur. Cela suppose de pouvoir comparer l'efficacité d'un classifieur. Un résultat important de la méthode est qu'après chaque itération on obtient un classifieur meilleur qu'avant. On peut donc arrêter les itérations à tout moment, même si le résultat n'est pas l'optimum. [Site2, 2008].

ANNEXE 7

Algorithmes de classification non supervisés

- Algorithme **CURE** (**C**lustering **U**sing **RE**presentatives)

Il a été proposé par *Guha et al*, cet algorithme utilise un échantillon représentatif de l'échantillon total pour réduire la complexité temporelle des calculs. Cet échantillon sera divisé en sous-ensembles qui sont regroupés en sous-classes. Les sous-classes seront agrégées hiérarchiquement en utilisant la distance entre deux sous classes C_1 et C_2, la plus petite distance entre un représentant de C_1 et un représentant de C_2 jusqu'à obtenir k classes demandées.

- Algorithme **BIRCH** (**B**alanced **I**terative **R**educing and **C**lustering using **H**ierarchies)

Il a été développé par *Zhang et al* en 1996. C'est un algorithme qui travaille efficacement sur de gros jeux de données. L'idée principale du **BIRCH** est d'effectuer une classification sur un résumé compact des données, au lieu des données originales. C'est pourquoi il peut traiter un grand volume de données en utilisant une mémoire limitée. Il est incrémental, i.e. il a besoin d'un seul balayage du jeu de données. Il essaie de minimiser le coût d'entrée/sortie en organisant les données traitées en une structure d'arbre équilibré avec une taille limitée.

- Algorithme de **ROCK** (**RO**bust **C**lustering using lin**K**s)

Il a été développé par *Ghura et al*. En 1999. C'est un algorithme de classification hiérarchique. Il utilise le nouveau concept appelé liens (*links en anglais*) pour mesurer la similarité entre deux individus au lieu des métriques pour les données numériques ou le coefficient de *Jaccard*.

Deux individus sont voisins si leur similarité dépasse un seuil θ : $sim(p,q) > \theta$

La similarité entre un pair d'individus peut être mesurée en se basant sur les distances métriques, le coefficient de *Jaccard* ou n'importe quelle fonction de similarité non métrique.
Le nombre de liens entre une paire d'individus est le nombre de leurs voisins communs. En général, les individus se trouvent dans une classe ont le nombreux voisins, par conséquent, de nombreux liens. En se basant sur le concept liens (*links*) entre deux individus, on définit le lien ou l'interconnexion entre deux classes par le nombre de liens croisés (*cross links*) entre eux. Ce nombre est calculé par la somme de liens entre tous les points situés dans deux classes.

- Algorithme **TSVQ** (**T**ree **S**tructured **V**ector **Q**uantization)

Il a été proposé par *Gesho et Gray* en 1992. Cet algorithme utilise l'algorithme K-moyennes (k=2) pour le partitionnement. Utilise la somme des distances par rapport au centroïde (au lieu de la distance moyenne).

ANNEXE 8

Extraits de notre corpus

كشف المدير العام للوظيف العمومي بأن مصالحه ألغت نتائج جميع مسابقات قطاع التربية العام الماضي، بسبب تزوير كبير تورط فيه إطارات من القطاع. وطالب الوزير بن بوزيد بـ''تنظيم قطاعه بدل اتهام مديرية الوظيف العمومي بالبيروقراطية، وإذا كانت الشفافية ومحاربة الغش والتزوير بيروقراطية فنحن إذن بيروقراطيون"

كشفت مصادر أمنية مسؤولة، بأن المدير العام للأمن الوطني ''حذر'' أعوانه من ''الدخول في مواجهات مع المواطنين أثناء التدخل في حالات أعمال الشغب والاحتجاجات''، وألزمهم بالتعامل بمرونة مع الأوضاع تفاديا لأي انزلاقات

أشارت مصادر مصرفية لـ''الخبر'' أن كافة البنوك المعتمدة في الجزائر ستشرع ابتداء من الفاتح جانفي 2010 في اعتماد نظام آلي لتسيير وتحويل أجور عمال جميع المؤسسات العمومية والخاصة، وأنها ستقوم برفع تكلفة العمليات بناء على التدابير الجديدة، وستقع الزيادة على عاتق المؤسسات دون اقتطاعها من أرصدة العمال والأجراء

أكد المفوض العام لجمعية هيئة البنوك والمؤسسات المالية، السيد عبد الرحمان بن خالفة في تصريح لـ''الخبر''، بأن عملية التسيير الإلكتروني لتحويل أجور العمال المقررة مؤخرا من طرف وزارة المالية ستسمح بتحديث تبادل المعلومات بين المؤسسات والبنوك، إلى جانب ضمان سرعة التحويل وانعدام الأخطاء في تحويل الأجور. وأوضح السيد عبد الرحمان بن خالفة بأن مليون عامل أجير سيستفيدون من هذه العملية في مرحلتها الأولى، وذلك خلال السنة الأولى، لتتوسع رقعة الاستفادة لتمس جميع العمال على المدى الطويل

ذكرت مصادر مطلعة لـ''الخبر'' أن الاجتماع الأخير الذي جمع وزيري النقل، والطاقة والمناجم بالرؤساء المديرين العامين لموانئ أرزيو وسكيكدة وبجاية، والرئيس المدير العام لشركة ''أش تي أش'' تمخض عن قرار يقضي بتحويل كل القاطرات البحرية لمؤسسة ميناء أرزيو بعمالها، و50 بالمائة من قاطرات ميناء سكيكدة إلى شركة تسيير واستغلال موانئ المحروقات

صنفت الهيئة البريطانية المتخصصة ''ليغاتوم بروبرتي'' الجزائر في المراتب العشر الأخيرة في مجال الابتكار الصناعي والمقاولاتية والحريات النقابية والعمالية، فيما تصدرت دول أوروبا الشمالية وعلى رأسها فلندا ....

أعلن الأستاذ والباحث فريد شربال بأن الحكومة أخطأت في قرار خوصصة الجامعة، حيث تنبأ بفشل العملية، بالنظر إلى ''الأزمة'' الكبيرة التي تعيشها هاته الأخيرة

اقتحمت قوات الاحتلال الإسرائيلي، فجر أمس، بلدة ديرالغصون، شمالي طولكرم، وأجرت عملية تفتيش في عدد من منازل الفلسطينيين. في الوقت ذاتها طالب رئيس الحكومة الإسرائيلي السلطة الفلسطينية بمباشرة حوار غير مشروط بعد تلقي الضوء الأخضر من كلينتون

أعلن أمس، عبد الله عبد الله، وزير الخارجية السابق في حكومة حامد كرزاي، عن انسحابه من المنافسة، ستة أيام قبل إجراء الدور الثاني من الانتخابات . وبرر انسحابه بـ ''المعاملة السيئة للحكومة واللجنة المستقلة للانتخابات

قالت والدة ألكس وينز، الشخص الذي طعن مروة الشربيني بالسكين في محكمة بدريسدن الألمانية، أن السبب في الفعل ناتج عن الحصص التلفزيونية التي تدعو إلى كراهية الإسلام والمسلمين

اعتبر سفير روسيا بطهران، ألكسندر سادوفنيكوف، أن الوثيقة المعروضة على إيران ليست ''فخا''، إنما هي ''لصالح إيران''، في تصريح أدلى به أمس لوكالة إيرنا الإيرانية، حيث قال: ''نظن أن المشروع قيد التوقيع على اتفاق تقني في صنع الوقود لمفاعل البحث،

أكد المتدخلون في الندوة التي نظمتها مؤسسة الأمير عبد القادر، أمس، بمقر جريدة ''المجاهد''، حول العلاقة بين الأمير عبد القادر والباي أحمد بقسنطينة، على وطنية الرجلين، بينما طالبوا بعدم الحكم على أي طرف ودراسة التاريخ وفق نسقه العام .

دعا المشاركون في ندوة ثورة الزعاطشة ببسكرة، أمس، رئيس الجمهورية عبد العزيز بوتفليقة من أجل العمل على استعادة رؤوس قادة هذه الثورة ممثلة في الشيخ بوزيان وابنه والشيخ موسى الدرقاوي الموجودة بالمتحف الأنثروبولوجي بباريس، ودفنها في موقع المعركة ببلدية ليشانة. كما طالبوا أيضا بإنجاز فيلم يجسد وقائع هذه الثورة التي كانت مفتاحا للثورات الشعبية في الجزائر ضد المستعمر الفرنسي

أكدت السيدة دانييل جيولي غوفري، مساعدة عمدة باريس المكلفة بالذاكرة وبالعالم المناضل، أنها على استعداد للنضال من أجل تحقيق المقترح الذي تقدم به المؤرخ الفرنسي، أوليفيي لاكور غرانميزون، لإقامة نصب تذكاري تخليدا لأرواح شهداء مجزرة 8 ماي 45 بالجزائر

أحتضن متحف المجاهد ودار القيادة وإقامة الذكور الجامعية بمعسكر، نهار أمس، الإحتفالات بالذكرى ال176 للمبايعة الثانية للأمير عبد القادر، التي شملت عدة أنشطة

تحدّث جيمس موير، الرئيس المدير العام لـ''سيات''، الذي قام بزيارة للجزائر منذ أيام، عن مستقبل تواجد العلامة الإسبانية في الجزائر. وأعرب عن نيته في ضرورة تبني استراتيجية جديدة لـ''سيات'' في الجزائر لرفع المبيعات وتأكيد الصورة الجيدة التي تتمتع بها العلامة في السوق الجزائرية

يظهر أن كثيرا من فتياتنا خاصة المتعلمات واللواتي استطعن تأمين حياتهن بتوفر منصب عمل قار ورفيع، إلى جانب سكن فردي وسيارة، اكتسبن حرية تصرف في الحياة جعلتهن يفضلن الاستغناء عن الزوج، مما تسبب في ظهور حالات متعددة من ''العنوسة الاختيارية'' التي باتت رقعتها تتوسع يوما بعد يوم، وتتغذى من الاستقلالية التي اكتسبتها الفتاة الموظفة سواء إزاء أسرتها أو قدرتها المادية

قال الرائد لخضر بورقعة، في كلمة ألقاها في حفل تدشين نصب تذكاري بمقر جبهة القوى الاشتراكية تخليدا لشهداء الثورة وتكريما لشهداء الديمقراطية وحقوق الإنسان: إن الجزائر تعيش حالة هستيرية غير مسبوقة من الفضائح والجرائم المالية بتواطؤ مفضوح من مسؤولين في السلطة. ودعا بورقعة، الشباب والطلبة ''إلى تشكيل قوى مدنية جديدة للتحرك لوقف هذه الجرائم والقذارة السياسية والاقتصادية، وتغيير الأوضاع ومحاسبة المسؤولين وتحرير الأفكار والذهنيات''

أكد الدكتور حشماوي محلل اقتصادي وأستاذ بكلية العلوم الاقتصادية، بأن قانون المالية لسنة ,2010 تضمن مجموعة من الإجراءات الجديدة التي لها صلة بتنظيم قطاع التجارة الخارجية، التي كانت تعرف نوعا من الفوضى فيما يتعلق بمسالة الاستيراد. مضيفا بأن هذا القانون أبقى على كل التعديلات التي جاء بها قانون المالية التكميلي لـ ,2009 ودعم هذه الإجراءات بإجراءات جديدة .بالرغم من انه أثار حفيظة بعض المتعاملين الاقتصاديين الخواص الذي يرون مصالحهم الخاصة على عكس الدولة التي تنظر لمصالح الجميع. وقال أيضا بأن إجراءات قانون المالية لـ 2010 جعل المعاملات التجارية شفافة أكثر فأكثر، وهذا هو الهدف الرئيسي الذي جاء به القانون.

Références bibliographiques

[Abbès, 2004]

Ramzi Abbès ; « La conception et la réalisation d'un concordancier électronique pour l'arabe », Thèse de Doctorat, soutenue, le 13 décembre, 2004. Université Lumière, Lyon2.

[Abbès et al., 2008]

Ramzi Abbès et Joseph Dishy; « Extraction automatique de fréquences lexicales en arabe et analyse d'un corpus journalistique avec le logiciel AraConc et la base de connaissances DIINAR.1 », Université Lumière Lyon 2, ICAR-CNRS ; JADT 2008 : 9es Journées internationales d'Analyse statistique des Données Textuelles. 2008.

[Audibert, 2006]

Laurent Audibert ; « Etude des critères de désambigüisation sémantique automatique : résultats sur les cooccurrences », TALN 2003, Batz-sur-Mer, 11-14 Juin 2003.

[Alrahabi, 2004]

Motasem Alrahabi, Ghassan Mourad, Brahim Djoua, « Filtrage sémantique de textes en arabe en vue d'un prototype de résumé automatique », LaLICC (Langage, Logique, Informatique, Cognition et Communication) UMR 8139, Université Paris – Sorbonne, CNRS 96, Bd Raspail 75006 Paris – France. Le traitement automatique de l'arabe, JEP-TALN 2004, Fès, 19-22 avril 2004.

[Barque, 2004]

Lucie Barque ; « Opérations sémantiques sur une base de données Sens-Texte» DEA de linguistique théorique descriptive et formelle Option linguistique informatique Septembre 2003, Université Paris 2007, UFR Linguistique.

[Béchet, et al., 2008]

Nicolas Béchet, Mathieu Roche, Jacques Chauché ; « ExpLSA et classification des textes », JADT 2008 : 9es Journées internationales d'Analyse statistique des Données Textuelles, 2008.

[Belguith, 2006]

Lamia Hadrich Belguith, Nouha Chaâben, « Analyse et désambigüisation morphologiques de textes arabes non voyellés », TALN 2006, Leuven, 10-13 Avril 2006.

[Boubou, 2007]

Mounzer Boubou, « Contribution aux de classification non supervisée via des approches prétropoogiques et d'agrégation d'opinion », Thèse de Doctorat, soutenue, le 29 novembre, 2007.

[Breck, 2008]

Eric John Breck; « empirical methods for fine-grained opinion extraction from text », Thèse de Doctorat en Philosophie, soutenue en Août 2008. Faculty of the Graduate School of Cornell University, NY, USA, 2008.

[Brill, 1994]

Brill E. « Some advances in transformation-based part of speech tagging ». In AAAI, Vol. 1, pages 722-727. 1994.

[Boulaknadel, 2008]

Siham Boulaknadel, « Utilisation des syntagmes nominaux dans un système de recherche d'information en langue arabe », LINA FRE CNRS 2729- Université de Nante 2 rue de la Houssinière, BP 92208 44322 Nantes cedex 03, France. 2008.

[Candilier, 2006]

Laurent Candillier, « Contextualisation, visualisation et évaluation en apprentissage non supervisé », Thèse de Doctorat soutenue le, 15 septembre 2006.

[Chauché, 1984]

Chauché J. « Un outil multidimensionnel de l'analyse du discours ». In Proceedings of Coling, Standford University, California, pages 11-15. 1984.

[Choi, et al., 2005]

Yejin Choi,Claire Cardie, Ellen Riloff et Siddharth Patwardhan; « Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns », In Proceedings of HLT/EMNLP, 2005.

[Cherabit et al., 2005]

Noureddine CHERABIT, Amar DJERADI, Rachida DJERADI. «Modélisation Du Dialogue Homme Machine En Langue Arabe », SETIT 2005, 3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications March 27-31, 2005 – TUNISIA.

[Crestan et al., 2003]

Éric Creston , Marc El-Bèze et Claude de Loupy, « Peut-on trouver la taille de contexte optimale en désambiguïsation sémantique? », TALN 2003, Batz-sur-Mer, 11-14 Juin 2003.

[Ding et al., 2007]

Xiaowen Ding and Bing Liu; « The Utility of Linguistic Rules in Opinion Mining », Department of Computer Science University of Illinois at Chicago, 2007.

[Douzidia, 2004]

Fouad Soufiane Douzidia;« Résumé automatique de texte arabe » ; Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de M.Sc en informatique, Université de Montréal, Faculté des études supérieures, Septembre 2004.

[El-Kassas, 2005]

Dina El-Kassas ;« une étude contrastive de l'arabe et du français dans une perspective de génération multilingue », Thèse de Doctorat, soutenue, le 16 Décembre 2005.

[Esuli, 2006]

Andrea Esuli; « Opinion Mining »; Istituto di Scienza e Tecnologie dell'Informazion Consiglio Nazionale delle Ricerche, Pisa, Italy, Language and Intelligence Reading Group, June 14, 2006, Pisa, Italy.

[Esuli, 2008]

Andrea Esuli; « Automatic Generation of Lexical Resources for Opinion Mining: Models, Algorithms and Applications », Istituto di Scienza e Tecnologie dell'Informazione Consiglio Nazionale delle Ricerche via G. Moruzzi, 1 56124, Pisa – ITALY, 2008.

[Ferrari, 2008]

Stépahane Ferrai, Yann Mathet et Thierry Charnois, « Analyse d'opinions : discours évaluatif et classification de documents», Atelier FODOP'08. Pages 23-36.

[Ganapathibhotla et al, 2008]

Ganapathibhotla Murthy et Bing Liu ; « Mining Opinions in Comparative Sentences », COLING, 2008.

[Généreux et al, 2009]

Michel Généreux et Aurélien Bossard ; « Résumé automatique de textes d'opinions », TALN 2009, Senlis, 24–26 juin 2009.

[Harb, et al., 2008]

Ali Harb, Michel Plantié, Gerard Dray, Mathieu Roche, François Trousset, Pascal Poncelert ; « Web Opinion Mining : how to extract opinions from blogs » ;CSTST'08: International Conference on Soft Computing as Transdisciplinary Science and Technology, 2008.

[Hatzivassiloglou et al,. 1997]

Hatzivassiloglou V., McKeown K., « Predicting the semantic orientation of adjectives », In Proceedings of 35th Meeting of the Association for Computational Linguistics, 1997.

[Hu et al., 2005]

Minqing Hu, Bing Liu; « Opinion Extraction and Summarization on the Web », Department of Computer Science University of Illinois at Chicago, 2006.

[Jaillet, 2004]

Simon Jaillet, « Catégorisation automatique de documents », LIRM, UMR 5506, 161 rue Ada, 34392 Montpellier Cedex 5 – France, 2004.

[Kamps et al, 2004]

Kamps J., Marx M., Mokken R. J., de Rijke M., « Using WordNet to Measure Semantic Orientation of Adjectives», In Proceedings of LREC 2004, the 4th International Conference on Language Resources and Evaluation, vol. IV, p. 174-181, 2004.

[Kanejiya et al., 2003]

Kanejiya D., Kumar A. and Prasad S. (2003). « Automatic evaluation of students answers using syntactically enhanced LSA ». In Proceedings of the Human Language Technology Conference (HLT- AACL 2003) Workshop on Building Educational Applications using NLP.2003.

[Kelaiaia, 2008]

Kelaiaia Abdessalem, « Classification non supervisée de textes arabes : appliquée à la recherche documentaire », Thèse de Magister. Département d'informatique, Université du 08 Mai 1945, Guelma, 2008.

[Larkey et al., 2002]

Larkey L. S., Ballesteros L. and Connell M., « Improving Stemming for Arabic Information Retrieval: Light Stemming and Cooccurrence Analysis », In Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002), Tampere, Finland, August 2002.

[Lehmam et al., 2004]

Abderrafih Lehmam et Philipe Bouvet, « Un résumeur automatique de textes multilingues intégré dans une plate-forme de veille ; application à la langue arabe», Société Pertinence Mining-Paris, France, JEP-TALN, 2004, Traitement Automatique de l'Arabe, Fès, 20 Avril 2004.

[Liu, 2006]

Bing Liu; « Mining and Summarizing Opinions on the Web », Department of Computer Science, University of Illinois at Chicago, 2006.

[Liu, 2009]

Bing Liu; « HandBook of natural language processing », Second edition, Department of Computer Science, University of Illinois at Chicago, 2009.

[Mathieu, 2000]

Mathieu Y ; « Les verbes de sentiment. De l'analyse linguistique au traitement Automatique » ; CNRS Éditions, 2000.

[Maurel et al., 2007]

Sigrid Maurel, Paolo Curtoni; « Classification d'opinions par méthodes symbolique, statistique et hybride », AFIA 2007 - DEFT'07, Grenoble, 3 juillet 2007.

[Memmi, 2000]

Daniel Memmi, « Le modèle vectoriel pour le traitement de documents », 2000.

[Nasraoui, 2007]

Olfa Nasraoui, « Book Review : Web Data Mining-Exploring Hyperlinks Contents, and Usage Data», Knowledge Discovery & Web Mining Lab Computer Engineering and Computer Science department Speed School of Engineering University of Louisvill Louisville KY 40292. Page 23-25, 2007.

[Mouileh, 2008]

Mouileh Zoubir, « AraSeg : Un segmenteur semi-automatique des textes arabes», ICAR- Université Lumière, Lyon 2. JADT 2008 : 9es Journées internationales d'Analyse statistique des Données Textuelles.

[Pang, et al., 2008]

Bo Pang et Lillian Lee, « Opinion Mining and Sentiment Analysis», 2008. Yahoo! Research, 701 First Avenue, Sunnyvale, CA 94089, USA, et Computer Science Department, Cornell University, Ithaca, NY 14853,USA, 2008.

[Popescu, 2005]

Ana-Maria Popescu et Oren Etzioni, « Extracting Product Features and Opinions from Reviews, Department of Computer Science and Engineering University of Washington, 2005.

[Rehder et al., 1998]

Rehder B., Schreiner M., Wolfe M., Laham D., Landauer T. and Kintsch W. « Using latent semantic analysis to assess knowledge: some technical considerations ». In Discourse Processes, volume 25, pages 337-354. 1998.

[Riloff, 1996]

Ellen Riloff et William Phillips;« An Introduction to the Sundance and AutoSlog Systems »; School of Computing, University of Utah, Salt Lake City, UT 84112 USA, November 8, 2004.

[Robert, 2001]

Le petit Robert : Dictionnaire alphabétique et analogique de la langue française, Maison d'édition VUEF, 2001.

[Roche et al., 2006]

Roche M. and Chauché J. (2006). « LSA : Les limites d'une approche statistique ». In Proceedings of atelier FDC'06 (Fouille de Données Complexes), conférence EGC'2006, pages 95-106. 2006.

[Rosá, 2008]

Aila Rosá ; « Identification des marques d'opinions dans des textes », Facultad de Ingeniería - Universidad de la República, J. Herrera y Reissig 565, Montevideo, Uruguay, Recital 2008, Avignon, 9-13 juin 2008.

[Stavrianou, et al. 2008]

Anna Stavrianou, Jean et Hugues Chauchat ; « Opinion Mining and Agreement Identification in Forums », FODOP, 27 Mai 2008.

[Taboada et al, 2006]

Taboada M., Anthony C., Voll K., « Creating semantic orientation dictionaries », 2006.

[TLFI, 2009]

« Trésor de la langue française informatisée », 2009.

[Turney, 2002]

Turney; « Thumbs up or thumbs down ? Semantic orientation applied to unsupervised classification of reviews », In Proceedings of 40th Meeting of the Association for Computational Linguisticsp. 417-424, 2002.

[Van, 2007]

Tim Van de Cruys et Begõna Villada Moirón, « Semantics-based Multiword Expression Extraction», Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, pages 25–32, Prague, June 2007.

[Vernier, 2007]

Mathieu Vernier, Yann Mathet, François Rioult, Thierry Charnois, Stéphane Ferrari et Dominique Legallois ; « Classification de textes d'opinions : une approche mixte n-grammes et sémantique », DEFT'07, Grenoble, France, 2007.

[Vinot et al., 2003]

Romain Vinot , Natalia Grabar, Mathieu ,« Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'Internet» Valette. TALN 2003, Batz-sur-Mer, 11–14 juin 2003.

[Voll et al,. 2002]

Voll K., Taboada M., « Not All Words are Created Equal : Extracting Semantic Orientation as a Function of Adjective Relevance », 2007.

[Voorhees, 1994]

Voorhees E.M. « Query Expansion using Lexical-Semantic Relations ». In Proceedings of ACM SIGIR'94, Dublin.1994.

[Wiemer-Hastings et al., 2001]

Wiemer-Hastings P. and Zipitria I. (2001). Rules for syntax, vectors for semantics. In Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society.

[Wilson, 2003]

Theresa Wilson et Janyce Wiebe, « Annotating Opinions in the World Press», Intelligent Systems Program University of Pittsburgh Pittsburgh, PA 15260, USA, 2003.

Références Webographiques

[Site1, 2008] ; http://sis.univ-tln.fr/~tollari/ARTICLES/DEA2003

[Site2, 2008] ; http://www.dbmsmag.com/9807m05.html

[Site3, 2009]; http://www.toupie.org/Dictionnaire/Opinion.html

[Site4, 2009] ; http://www.dicoplus.org/definition/opinion

[Site5, 2009]; http://www.maphilio.net/objectif-subjectif-cours.html