REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

N° d'ordre : .....

UNIVERSITE DE M'SILA

FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

Département de L'INFORMATIQUE

MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique
Filière : Informatique
Spécialité : Système D'Information Avancé
Par: Gherabi Sara

SUJET

CLASSIFICATION AUTOMATIQUE DES TEXTES ARABE (ARABIC OPINION POLARITY)

Soutenu publiquement le : 19 / 06 /2014 devant le jury composé de :

……	**Université de M'sila**	**Président**
Brahimi Belkacem	**Université de M'sila**	**Rapporteur**
……	**Université de M'sila**	**Examinateur**
……	**Université de M'sila**	**Examinateur**

Promotion : 2013 /2014

" بِسْمِ اللَّهِ الرَّحْمَـــــن

الرَّحِيـــمْ "

Dédicace

À mes parents ;

Mon père, le plus beau père…

Ma mère, la plus belle mère…

Qu'Allah les garde en bonne santé !

Aux deux perles de ma vie : Mes deux sœurs,

Amel et souhila,

Que Dieu les protège !

À toutes les personnes qui m'aiment qu'ils trouvent…

Remerciements

Grâce à Allah tout d'abord de me donner la puissance, la santé et l'aide d'accomplir cette recherche. Sans l'aide de Dieu, je ne pouvais pas accomplir ce travail.

Je souhaite ensuite exprimer à monsieur **BELKACEM BRAHIMI,** qui a dirigé mes travaux, ma plus profonde gratitude pour sa disponibilité, ses conseils clairvoyants, son soutien sans faille et la confiance qu'il a bien voulu m'accorder.

Je tiens également à remercier vivement de me faire l'honneur de participer au jury.

Je souhaite aussi témoigner de ma sympathie et de ma gratitude à tous ceux qui ont toujours été agréable avec moi

Mes plus affectueux remerciements vont évidemment à toute ma famille, et tout d'abord à mes parents qui m'ont toujours soutenu et encouragé dans tout ce que j'ai entrepris.

Mes remerciements vont aussi à toute l'administration de la Département de l'informatique, pour leur gentillesse.

Je remercie tendrement mes amies pour leurs conseils et leur soutien moral, plus particulièrement : **Nabila, Nassima, Samia**.

Que toutes les personnes qui ont attribué de près ou de loin à l'élaboration de ce travail.

Merci à tous et à toutes.

Tables des matières

Bibliographie

Liste des tableaux

Liste des figures

Liste des abréviations

DM	Data Mining.
RI	Recherche d'Information.
KPPV	K plus Proche Voisin.
SVM	Support Vecteur Machine.
SQL	Structured Query Language.
TF	Term Frequency.
IDF	Inverse Document Frequency.
Doc	Document.
GNU	General Public License.
NB	Naïve Bayes.
PATB	Penn Arabe TreeBank.
OCA	Opinion Corpora Arabe.
EVOCA	English Version opinion corpora arabe.
ANER	Arabic Named Entity Recognition.
Prepare _data ;	Préparation des données.
Nbr	Nombre.
Nbr_tp	Nombre des termes positive.
Nbr_tn	Nombre des termes négatif.
SFU	Simon Fraser University.
JSE	Java Standard Edition.
JEE	Java Enterprise Edition.
JME	Java Micro Edition.
JDK	Java Development Kit.
JRE	Java Runtime Environment.
VP	Vrai positif.
VN	Vrai négatif.
FP	Faux positive.
FN	Faux négatif.
Temp_ESP	Temps des Expérimentations Sans Prétraitement.
Temp_EAP	Temps des Expérimentations Avec Prétraitement.
TALN	Traitement Automatique du Langage Naturel.

INTRODUCTION GENERALE

1. Contexte

Depuis l'émergence du Web 2.0 et des sites communautaires, une quantité croissante de textes non structurés prolifère sur la toile. Ces textes, généralement produits par les internautes, sont très souvent porteurs de sentiments et d'opinions sur des produits, des films, des musiques, etc. Ces données textuelles représentent potentiellement des sources d'information très riches permettant a priori de découvrir les attentes, désirs, besoins des utilisateurs ou encore de mesurer la popularité de certains produits, personnalités, décisions politiques, etc.

La classification automatique de textes est un domaine où la fouille de textes et les techniques statistiques produisent des résultats à partir des calculs de fréquence d'occurrence de termes extraits. Le domaine de la fouille d'opinion peut-être divisé en trois sous-domaines :

- L'identification des textes d'opinion, qui peut consister à identifier dans une collection textuelle les textes porteurs d'opinion, ou encore à localiser les passages porteurs d'opinion dans un texte. Plus précisément, on parle ici de classer les textes ou les parties de texte selon qu'il sont objectifs ou subjectifs ;
- le résumé d'opinion, qui consiste à rendre l'information rapidement et facilement accessible en mettant en avant les opinions exprimées et les cibles de ces opinions présentes dans un texte. Ce résumé peut être textuel (extraction des phrases ou expressions Différentes Approches Pour la Classification d'Opinion contenant les opinions), chiffré (pourcentage, note), graphique (histogramme) ou encore imagé (thermomètre, étoiles, pouce levé ou baissé…)
- la classification d'opinion, qui a pour but d'attribuer une étiquette au texte selon l'opinion qu'il exprime. On considère généralement les classes positive et négative, ou encore positive, négative et neutre. Nous nous intéresserons ici uniquement à la classification d'opinion. Nous cherchons à déterminer les goûts cinématographiques d'utilisateurs à partir de l'analyse de leurs commentaires.

Les données étudiées sont des textes d'opinion rédigés en arabe, chacun d'eux étant associé à un utilisateur et à un film mais également associé à une note attribuée par l'auteur au moment de la rédaction. Les textes de ce corpus présentent plusieurs particularités : ils sont en général très courts.

2. Problématique

La quantité d'information accessible de nos jours sur Internet est phénoménale, et sa catégorisation reste l'une des tâches les plus importantes et très difficile au même temps. Beaucoup de travail est actuellement concentré sur l'anglais à juste titre puisque, c'est la langue dominante du web. Néanmoins, un besoin se fait sentir pour les autres langues car le web est chaque jour plus multilingue. Le besoin est beaucoup plus pressant pour la langue arabe. Nos recherches sont sur la classification des textes arabes plus en détail la classification des textes d'opinion arabes selon leur polarité (positive, négative ou neutre),

3. Objectifs

L'objectif global de notre travail, est d'offrir une méthode automatisée pour la classification de chaque commentaire selon l'opinion qu'il exprime une opinion positive ou une opinion négative. Pour ce faire, différents axes d'étude ont été envisagés :

Le premier axe combine traitements linguistiques sur la langue arabe et représentation des textes. Nous avons appliqué des prétraitements linguistiques au corpus dans le but d'améliorer la représentation des textes.

Le deuxième axe nous distinguons parmi les types d'approches dans la classification automatique: L'approche linguistique, est la méthode qui nous avons choisis comme une approche de classification. Le principe de ce type de méthodes consiste à construire, à l'aide d'outils de traitement automatique des langues, des lexiques de mots porteurs d'opinion et à classer les textes selon la présence ou l'absence de ces mots.

4. Organisation du mémoire

Ce mémoire va être organisé de la façon suivante : Un premier chapitre préliminaire pour définir l'ensemble des concepts de texte mining. Un état de l'art va être étalé au cours des chapitres 2, 3 des techniques employées dans les différentes phases du processus de prétraitement et classification automatique de textes, le quatrième chapitre indique l'approche proposée pour réaliser notre système, et le dernier chapitre présente l'ensemble des expérimentations et des performances des approches proposées.

- Dans **le premier chapitre,** nous allons résumer la technologie du data mining et du text mining et nous abordons le lien entre le elle et la classification de textes.
- Dans **le deuxième chapitre**, nous allons exposer les différentes opérations de prétraitement nécessaires avant de commencer à coder un texte. La définition et le choix des descripteurs ou termes, qui vont servir à représenter les documents, c'est un choix primordial et important dans la classification de textes. La réduction de dimensionnalité qui va servir à diminuer la taille du

vocabulaire avant d'appliquer les techniques de classification les plus complexes et enfin l'attribution des poids à ces termes. Tous ces points vont être étalés dans cette partie.

- Dans **le troisième chapitre**, nous présentons deux parties très importantes dans ce travail, une partie pour la classification des textes et la deuxième partie pour la fouille d'opinion et l'analyse de sentiment. Dans la première partie, nous définissons la classification et les différents jeux de mots utilisés : classification, catégorisation et clustering, ensuite le processus général de la classification de textes avec son schéma global, puis nous citons les problèmes spécifiques aux textes lors de l'apprentissage automatique. Tandis que la deuxième partie commence par les définitions du fouille d'opinion et l'analyse des sentiments, ensuite les applications sur ces taches et pour quoi elle import dans les gouvernements, puis nous expliquons la classification des opinions des personnes et les méthodes utilisées pour réaliser cette tâche, pour en finir avec la classification d'opinion en arabe.
- Dans **le quatrième chapitre**, nous présentons l'approche proposée pour la classification des textes de notre corpus. Tel que cette approche nommé par linguistique et consiste à construire un dictionnaire d'opinion manuellement avec l'aide de techniques simples de Traitement Automatique des Langues.
- Dans **le cinquième chapitre,** on trouve la partie de réalisation qui est consacré à l'implémentation et l'expérimentation de notre application (exemple de démonstration).
- Enfin, nous concluons ce mémoire en résumant les contributions que nous avons pu apporter, et en évoquant les suites de ce travail et les perspectives de recherche dans le domaine.

CHAPITRE 1
TEXT MINING

1. Introduction

Nous présentons deux parties principales dans ce chapitre. La première partie du chapitre aborde les définitions et le processus de la fouille de données, l'importance de l'utilisation, les différents types et les taches de cette fouille de données. La deuxième partie détaille cette technologie appliquée à la donnée textuelle, le Text Mining. Après l'exposition des définitions et les approches du Text Mining, nous expliquons la chaine de traitement pour le processus de fouille de données textuelle. Puis nous citons quelques applications pour cette technologie. A la fin, nous abordons le lien entre le Text Mining et la classification de textes.

2. Fouille de données (Data Mining)

2.1. Définitions

Le data mining, dans sa forme et compréhension actuelle, à la fois comme champ scientifique et industriel, est apparu au début des années 90. il est textuellement minage de données mais souvent traduit en français par fouille de données, se réfère d'extractions **automatique** d'informations **prédictives** à partir de grandes bases de données.

Le Data Mining est un processus qui consiste à comprendre des données de taille relativement importante, dans le but de découvrir de nouvelles vues, corrélations, tendances, modèles, règles ou relations cachées dans ces données en utilisant un ensemble de moyens matériels et logiciels à l'intersection de l'intelligence artificielle, les statistiques, l'apprentissage automatique et les systèmes de bases de données.[1]

- L'intelligence artificielle
- Les statistiques sont une branche des mathématiques appliquées. Dans le cadre de la théorie statistique, l'aléatoire et l'incertitude sont modélisés par la théorie des probabilités. Aujourd'hui, de nombreuses méthodes statistiques sont utilisées dans le domaine du data mining.
- Les bases de données sont nécessaires afin d'analyser de grandes quantités de données efficacement. L'analyse des données avec des algorithmes de data mining peut être soutenue par des bases de données et donc l'utilisation de la technologie de base de données dans le processus du data mining peut être utile.
- L'apprentissage automatique est un domaine de l'intelligence artificielle, qui permet le

développement des techniques permettant aux ordinateurs « d'apprendre » par l'analyse de l'ensemble des données.

2.2. Processus du Data mining

Les outils du Data Mining s'intègrent dans un processus itératif à six phases qui doit être appliqué à un ensemble de données dans le but d'extraire un modèle utile. Ce processus est définit par CRISP-DM (Cross Industry Standard Process for Data Mining) conçu fin 1996. **[2]** La **Figure 1.1** suivante montre les phases d'un processus CRISP-DM. La séquence des phases n'est pas rigide, un va et vient entre les différentes phases est toujours nécessaire, il dépend de l'issue de chaque phase.

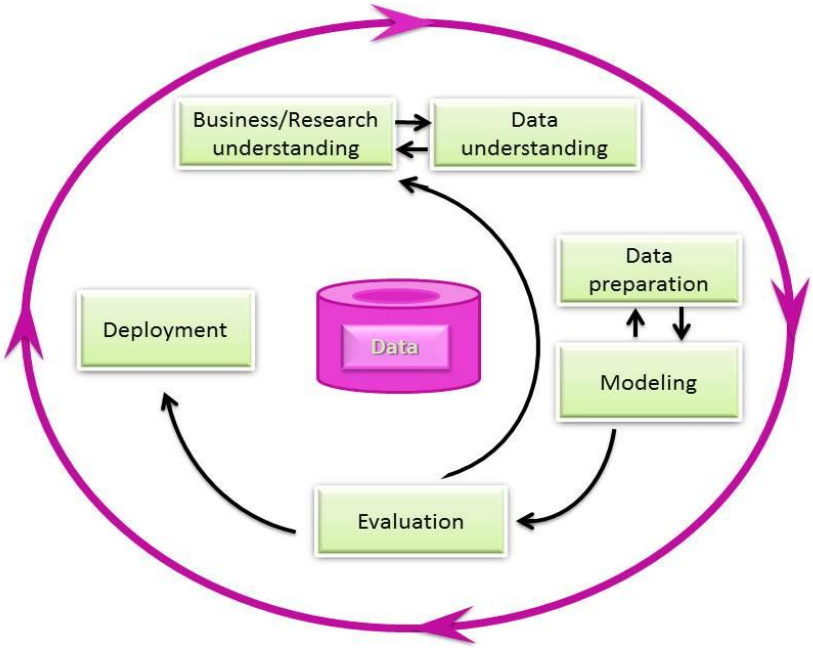

Figure 1.1 Les phases de Crisp DM. [2]

Les six phases sont les suivantes :

1. **Compréhension du problème** ***(****Business understanding phase****)*** *:*

La première phase dans le processus standard CRISP-DM peut aussi être appelée la phase de compréhension de la recherche.

- Déterminer les objectifs commerciaux
- Evaluer de la situation
- Déterminer les objectives du Data Mining
- Produire d'un plan du projet

2. **Compréhension des données (***Data understanding phase****)*** *:*

- Collecte des données initiales

- Description des données
- Exploration des données
- Vérification de la qualité des données

3. **Préparation des données (***Data preparation phase***)** *:*
   - Sélection des données
   - Nettoyage des données
   - Construction de nouvelles données
   - Intégration des données
   - Formatage des données
4. **Modélisation (***Modeling phase***)** *:*
   - Sélection des techniques de modélisation
   - Génération d'une conception de test
   - Création des modèles
   - Evaluation de modèles
5. **Evaluation (***Evaluation phase***)** *:*
   - Evaluation de résultats
   - Processus de révision
   - Détermination des étapes suivantes
6. **Déploiement (***Deployment phase***)** *:*
   - Planification du déploiement
   - Planification de surveillance et maintenance
   - Production de rapport final
   - Exécution d'une révision de projet final

2.3. Pour quoi faire la fouille de données

On dit souvent que le Data Mining est …compliqué. Pourquoi donc doit-on l'apprendre? Pourquoi ne pas utiliser seulement les bases de données relationnelles?

- Le data mining n'est-t-il pas coûteux? Il semble qu'il a besoin de beaucoup de talent, de programmation, de temps de calcul et d'espace de stockage.
- Aide au Simplification et automatisation de l'ensemble du processus statistique, allant de la/les source(s) de données jusqu'au modèle de l'application.
- Ce processus n'est pas seulement un exercice académique malin, il a des applications réelles très profitables. Pratiquement toutes les grandes compagnies et plusieurs gouvernements développent le data mining comme une tâche de leur planification et analyse.

- Est la solution idéale pour résoudre les problèmes suivants :
  - Intérêt économique : du produit aux clients.
  - Technologie de l'information : faible coût de stockage de données, saisie automatique de transaction (code bar, click, données de localisation GPS, internet)
  - Augmentation de la puissance de calculs des ordinateurs (loi de Moore)
  - Extraire de la connaissance à partir de grandes bases de données devient possible.

2.4. Les types de données

Les données sont des valeurs des champs des enregistrements des tables de l'entrepôt (base de données) possède différent types :

- **Données discrètes:**
  - les données binaires ou logiques : 0 ou 1 ; oui ou non ; vrai ou faux. ce sont des données telles que le sexe, être bon client, ...
  - les données énumératives : ce sont des données discrètes pour lesquelles il n'existe pas d'ordre défini a priori. Par exemple : la catégorie socioprofessionnelle, la couleur, ...
  - les données énumératives ordonnées : les réponses à une enquête d'opinion (1: très satisfait ; 2 : satisfait ; ...), les données issues de la discrétisation de données continues (1 : solde moyen < 2000 ; 2 : 2000 ≤ solde moyen < 5000 ; ...)
- **Données continues:** ce sont des données entières ou réelles (âge, salaire, le revenu moyen, …) mais aussi les données pouvant prendre un grand nombre de valeurs ordonnées.
- **Dates:** sont souvent problématiques car mémorisées selon des formats différents selon les systèmes et les logiciels. Pour les applications en fouille de données, il est fréquent de les transformer en données continues ou en données énumératives ordonnées. On transforme une date de naissance en âge entier ou en une variable énumérative ordonnée correspondant à des tranches d'âge.
- **Données textuelles** (chaînes de caractères) : un texte peut, pour certaines applications, être résumé comme un n-uplet constitué du nombre d'occurrences dans le texte de mots clés d'un dictionnaire prédéfini.
- Pages/liens web, Multimédia,…
- **Données structurées :** graphe, enregistrement.

2.5. domaine d'application

- ✓ **Marketing direct:** population à cibler (âge, sexe, profession, habitation, région, …) pour un publipostage.
- ✓ **Gestion et analyse des marchés :** Ex. Grande distribution : profils des consommateurs, modèle d'achat, effet des périodes de solde ou de publicité, «panier de la ménagère»
- ✓ **Détection de fraudes:** Télécommunications, **...**
- ✓ **Gestion de stocks:** quand commander un produit, quelle quantité demander, …
- ✓ **Analyse financière:** maximiser l'investissement de portefeuilles d'actions.
- ✓ **Gestion et analyse de risque**: Assurances, Banques (crédit accordé ou non)
- ✓ **Compagnies aériennes**
- ✓ **Bioinformatique et Génome:** ADN mining, …
- ✓ **Médecine et pharmacie:**
  - Diagnostic : découvrir d'après les symptômes du patient sa maladie
  - Choix du médicament le plus approprié pour guérir une maladie donné
- ✓ **Internet :** spam, e-commerce, détection d'intrusion etc...
- ✓ **Web Mining (**moteur de recherche sur internet**)**, **Text Mining (**extraction d'information depuis des textes**)**, etc.

2.6. Les taches de la fouille de données

Il y a Beaucoup de problèmes intellectuels, économiques ou même commerciaux peuvent être exprimés en termes des six tâches résumées dans ce qui suit:

La classification.

L'Estimation.

La Prédiction.

Le groupement par similitude (Règles d'association).

L'analyse des clusters (Segmentation).

La description.

Les trois premières tâches sont des exemples de la fouille supervisée de données dont le but est d'utiliser les données disponibles pour créer un modèle décrivant une variable particulière prise comme but en termes de ces données. Les règles d'association et l'analyse des clusters sont des tâches non-supervisées où le but est d'établir un certain rapport entre toutes La description appartient à ces deux catégories de tâche, elle est vue comme une tâche supervisée et non-supervisée en même temps.

- **Classification**

La classification est la tâche la plus commune de la fouille de données qui semble être une tâche humaine primordiale. Afin de comprendre notre vie quotidienne, nous sommes constamment obligés à classer, catégoriser et évaluer. La classification consiste à étudier les caractéristiques d'un nouvel objet pour l'attribuer à une classe prédéfinie. Les objets à classifier sont généralement des enregistrements d'une base de données, la classification consiste à mettre à jours chaque enregistrement en déterminant la valeur d'un champ de classe.

- **Estimation**

L'estimation est similaire à la classification à part que la variable de sortie est numérique plu-tôt que catégorique. En fonction des autres champs de l'enregistrement l'estimation consiste à compléter une valeur manquante dans un champ particulier.

- **Le groupement par similitude**

Le groupement par similitude consiste à déterminer quels attributs "vont ensemble". La tâche la plus répandue dans le monde du business, est celle appelée l'analyse d'affinité ou l'analyse du panier du marché, elle permet de rechercher des associations pour11mesurer la relation entre deux ou plusieurs attributs. Les règles d'associations sont, généralement, de la forme "Si <antécédent>, alors <conséquent>".

- **L'analyse des clusters**

Le clustering (ou la segmentation) est le regroupement d'enregistrements ou des observations en classes d'objets similaires. Un cluster est une collection d'enregistrements similaires l'un à l'autre, et différents de ceux existants dans les autres clusters.

La différence entre le clustering et la classification est que dans le clustering il n'y a pas de variables sortantes. La tâche de clustering ne classifie pas, n'estime pas, ne prévoit pas la valeur d'une variable sortantes. Au lieu de cela, les algorithmes de clustering visent à segmenter la totalité de données en des sous groupes relativement homogènes. Ils maximisent l'homogénéité à l'intérieur de chaque groupe et la minimisent entre les différents groupes.

- **La description**

La description Parfois le but de la fouille est simplement de décrire se qui se passe sur une Base de Données compliquée en expliquant les relations existantes dans les données pour premier lieu comprendre le mieux possible les individus, les produit et les processus présents dans cette base. Une bonne description d'un comportement implique souvent une bonne explication de celui-ci. Dans la société Algériennes nous pouvons prendre comme exemple comment une simple description, "les femmes supportent le changement plus que les

hommes", peut provoquer beaucoup d'intérêt et promouvoir les études de la part des journalistes, sociologues, économistes et les spécialistes en politiques.

3. Fouille de textes (Text Mining)

3.1. Text Mining et Data Mining

Historiquement le Data Mining est à la base du Text Mining au sens où celui-ci est l'extension du même but et du même processus vers des données textuelles. Néanmoins, les deux technologies se distinguent dans la nature des données à traiter. Le Data Mining s'intéresse aux données numériques et factuelles qui sont bien structurées dans des bases de données, alors que le Text Mining s'intéresse aux données textuelles non structurées, généralement exprimées en langage naturel. [3]

3.2. Définition :

Le Text Mining fait appel à diverses méthodes d'analyse, comme la linguistique, la classification automatique ou la catégorisation. L'application de ces méthodes, nécessite en fonction du type d'indicateur que l'on souhaite mettre en place, une plus ou moins grande connaissance formalisée du domaine couvert par les documents à analyser.

« La fouille de textes est la découverte à l'aide d'outils informatiques de nouvelles informations en extrayant différentes données provenant de plusieurs documents textuels. Un élément fondamental de ce processus réside dans les relations identifiées entre les informations extraites afin d'identifier de nouveaux faits ou de nouvelles hypothèses à explorer. ». [4]

3.3. Approches du Text Mining

Deux approches, peuvent être envisagées pour faire du Text mining :

3.3.1. Approche statistique

Elle consiste à ne voir le document que via le prisme du nombre et des chiffres. Ainsi l'outil statistique de Text Mining produit des informations sur le nombre d'occurrence d'un terme, la fréquence d'apparition d'un terme dans un document ou un corpus

3.3.2. Approche sémantique

L'analyse sémantique est une technique d'interprétation automatique des textes écrits en langue naturelle, c'est à dire tels qu'on les trouve dans les documents rédigés par et pour les humains. Cela permet à l'ordinateur de « comprendre » ces textes pour y collecter de l'information, pour classer les documents, pour en faciliter la recherche, etc.

3.4. Chaîne de traitement pour le processus de fouille de données textuelle

Un texte est considéré comme une entité porteuse d'une information qu'il faut préparer,

représenter et organiser pour lui permettre d'utiliser des outils de fouille de données et valider les résultats de la fouille, comme illustrer dans la **Figure 1.2** [5] suivante :

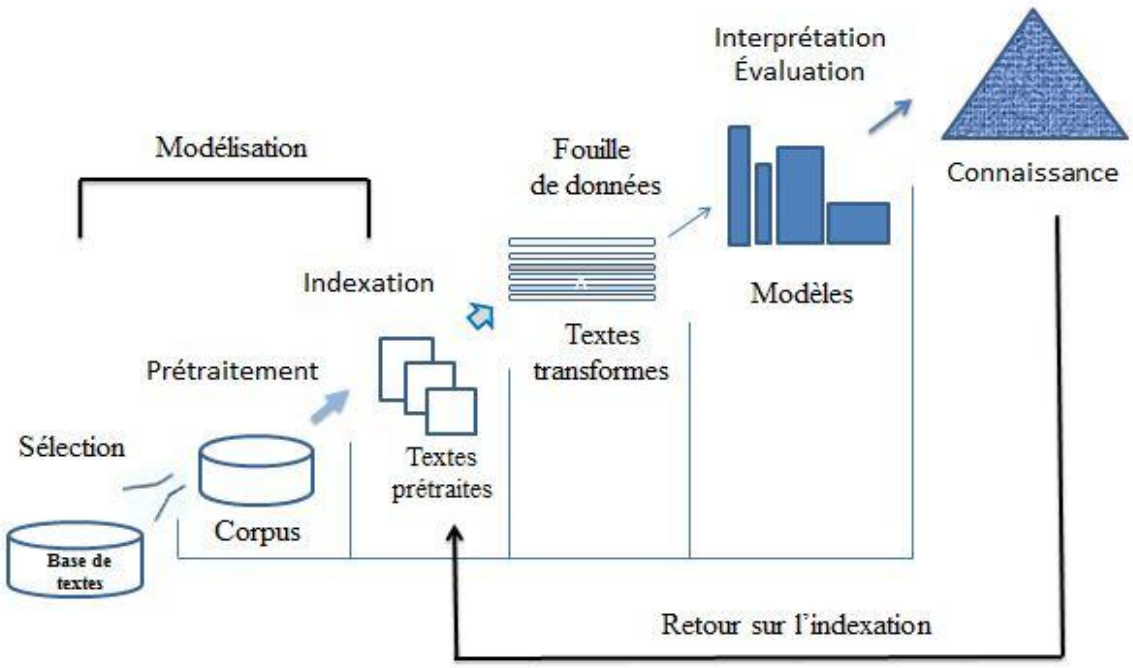

Figure 1.2 La chaîne de traitement pour le processus de fouille de textes. [5]

- **Le prétraitement:** est une tâche très importante et celle qui consomme le plus de temps, elle englobe les trois premières étapes du modèle CRISP-DM, de la compréhension du problème et des données à la préparation de ces dernières. Cette phase inclut tous les traitements, les processus et les méthodes nécessaires pour la préparation des données pour les opérations de bases de la découverte de connaissance du système Text Mining. L'étape du prétraitement en général converti les informations de leur source originale en un format intermédiaire.
- **La modélisation :** est le cœur du système Text Mining et inclut les opérations de bases de la fouille qui utilisent les algorithmes de Data Mining pour la découverte de connaissances.
- **L'évaluation :** pour décider de ce qu'il faut faire après, on termine le processus dans le cas où les résultats sont bien adaptés à l'application, sinon si le résultat est signifiant mais non satisfaisant, on réitère et le résultat généré sera utilisé comme une partie de l'entrée d'une ou de plusieurs étapes précoces.

Remarque : Le prétraitement et la modélisation sont les deux plus importantes phases dans n'importe quel système Text Mining, et généralement ils décrivent la série de processus au sein d'une vue généralisée d'une architecture d'un système Text Mining.

3.5. Tâches principales de la fouille de textes

Dans cette section, nous allons énumérer les trois principales tâches auxquelles s'attaque la fouille de textes. Chacune de ces tâches sera un cas particulier du schéma général de la **figure 1.3**, pour lequel nous préciserons :

- la nature des données et des résultats (en particulier, s'il s'agit de textes, quelle représentation est privilégiée)
- la nature des ressources utiles, à titre obligatoire ou facultatif
- la nature des méthodes utilisées pour la programmer, et si elle peut être abordée par apprentissage automatique
- les applications concrètes de cette tâche

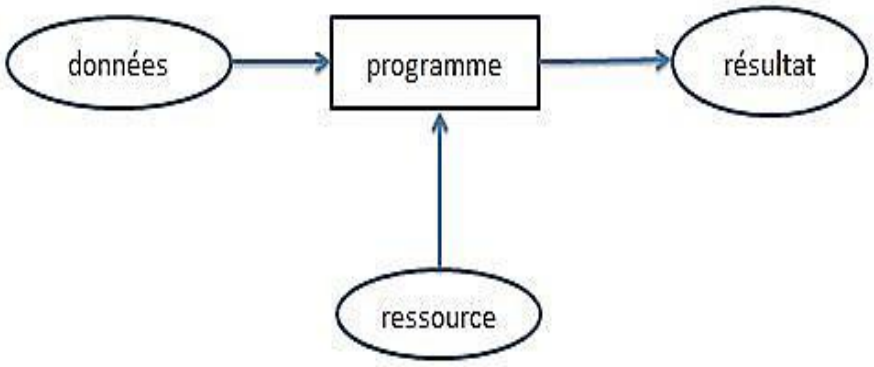

Figure 1.3: schéma général d'une tâche de fouille de textes.

La tâche la plus "naturelle" à envisager, étant donnée la section précédente, est:
La classification de textes. Elle consiste à ranger des textes ou des documents dans des "classes" prédéfinies : [6]

- **les données** sont donc des textes, la plupart du temps représentés sous la forme de vecteurs. Des variantes de ce type de représentation ont été étudiées spécialement pour cette tâche, par exemple pour donner plus d'importance aux mots présents dans des titres, ou privilégier certaines catégories grammaticales.
- **les ressources** nécessaires sont celles qui permettent la représentation du texte : antidictionnaire, lemmatiseur voire analyseur morphologique, compte d'occurrences, étiqueteur "part of speech" si on privilégie certaines catégories...

- cette tâche est presqu'exclusivement abordée par apprentissage automatique, à partir d'exemples de textes déjà classés. Parmi **les méthodes** on trouve, celles basées sur des comptes statistiques ("Naïve Bayes") ou sur les techniques SVM donnent les meilleurs résultats ou la méthode de K plus proche voisin KPPV.
- il y a de très nombreuses **applications** concrètes de cette tâche. L'une d'elle fonctionne déjà sur la plupart des gestionnaires de courrier électronique : il s'agit du programme qui suggère que certains des mails reçus sont probablement des "spams" non désirés. Les deux classes sont alors "spam" et "non spam" et l'ensemble des courriers déjà reçus constitue l'échantillon d'apprentissage à partir duquel le programme apprend à poser son diagnostic. De manière générale, la classification automatique de textes par "thème" peut rendre de grands services. On peut aussi utiliser des méthodes similaires pour retrouver l'auteur d'un texte (l'étiquette de la classe est alors un nom d'auteur) à partir d'exemples de textes attribués à coup sûr : des critiques littéraires s'en sont servi pour argumenter que certaines pièces signées par Molière avaient en fait été écrites par Corneille. Enfin, l'autre type d'application en plein développement de la classification est la reconnaissance automatique des opinions véhiculées par un texte : les classes, dans ce cas sont par exemple "favorable" et "défavorable". Certaines sociétés qui reçoivent des courriers électroniques de consommateurs à propos de leurs produits s'en servent pour analyser leur contenu. Dans ce cas, la représentation des textes a intérêt à privilégier les adjectifs et les verbes, qui sont les principaux moyens d'exprimer une opinion.

La recherche d'information(ou RI) est l'autre "tâche" générale d'ors et déjà omniprésente dans nos usages quotidiens des ordinateurs. Nous la sollicitons chaque fois que nous recherchons des documents répondant à une "requête". [7]

- la **donnée** fournie par l'utilisateur est donc une requête. Celle-ci peut prendre des formes diverses, suivant le niveau d'expertise de cet utilisateur et la structure de la base de documents à interroger : simple liste de mots clés, langage de requête structuré (combinaisons de critères booléens, expressions rationnelles, requêtes type SQL...), voire document "exemple" dont on cherche des exemplaires "proches" parmi un ensemble de textes.
- les **ressources** sollicitées sont tout d'abord le corpus de textes ou de documents que l'on cherche à interroger. Ce peut être une base d'articles, une encyclopédie, ce peut être Internet... Comme précédemment, il est éventuellement fait appel aux ressources nécessaires à la représentation de la requête par un vecteur.

- on distingue trois familles de **méthodes** pour aborder la RI :
    - les méthodes *booléennes* fonctionnent à l'aide d'un simple index qui donne, pour chaque unité lexicale figurant dans la requête, la liste des textes où cette unité est présente. Les requêtes acceptées sont alors généralement des combinaisons de critères booléens (avec les opérateurs NON, ET, OU). Des calculs simples permettent d'obtenir la liste des textes où tous ces critères sont satisfaits en même temps.
    - les méthodes *vectorielles*, comme leur nom l'indique, codent toutes les informations (la requête et les documents de la base) sous la forme de vecteurs. La représentation TF-IDF est née dans ce contexte, et y est particulièrement efficace. La RI se ramène alors à trouver les vecteurs les plus "proches" d'une vecteur donné (celui représentant la requête).
    - les méthodes *statistiques* qui en fait reviennent à faire de la classification automatique en supposant que l'on connaît déjà, pour la requête, un ensemble de documents "pertinents" et de documents "non pertinents", et que l'on cherche à trouver tous les documents devant être classés comme pertinents.
- la recherche sur Internet est, bien sûr, **l'application** phare de cette tâche. Les moteurs de recherche mettent en œuvre des méthodes booléennes : leur index fait leur force ! Or ces méthodes ne permettent pas de classer en "plus ou moins pertinent" les documents obtenus (en l'occurrence les sites Web). C'est pourquoi ils doivent employer d'autres techniques (d'où l'importance du fameux "Page Rank" de Google) pour classer par ordre de pertinence ces sites.

Enfin, ***L'extraction d'information*** est la dernière tâche fondamentale que nous voulons présenter ici. Comme son nom l'indique, elle se fixe comme objectif d'*extraire* de textes des informations factuelles précises. Imaginons par exemple les textes de petites annonces de vente de voitures, rédigées librement. Les informations qu'elles contiennent peuvent se résumer à la valeur de quelques "champs" factuels : qui vend quel type de voiture, de quel kilométrage, à quel prix, etc. On appelle *wrapper* (terme anglais qui signifie "envelopper") un programme capable de remplir automatiquement les valeurs de ces champs à partir du texte initial de la petite annonce. Un wrapper est nécessairement spécialisé dans le traitement d'un certain type de textes : celui qui traite les petites annonces de vente de voiture ne saura pas quoi faire d'annonces de locations d'appartements, et inversement. [8]

- les **données** d'entrées sont des représentations de textes de même type, où la notion de *séquence* est préservée, elles peuvent aussi être des *documents structurés* (pages

HTLM ou XML) ; les sorties sont des *données structurées*, en général sous la forme d'une liste d'attributs (prédéfinis) remplie ;

- Les **ressources** linguistiques utiles à la réalisation de cette tâche dépendent de la méthode employée : toutes les techniques d'identification d'entités nommées (liste de valeurs possibles, mais aussi expressions régulières ou automates) sont intéressantes car, souvent, la plupart des données à extraire (noms propres ou valeurs numériques) sont des entités nommées.
- Les **méthodes** les plus efficaces font appel, pour chaque champ à remplir, à des automates ou à des expressions régulières qui repèrent les environnements possibles où peut apparaître l'information visée. Mais, depuis quelques années, l'apprentissage automatique de wrappers à partir d'exemples de textes d'où ont été extraites des données factuelles est un thème de recherche très actif.
- Un système d'extraction automatique fournit rapidement un "résumé structuré" d'un texte. Les données "attributs/valeur" qu'il fournit en sortie peuvent facilement alimenter une feuille de calcul ou une base de données relationnelle, ce qui intéresse tous ceux qui doivent manipuler de nombreux exemplaires de documents standardisés.

3.6. Text Mining et la classification de textes

L'objectif du Text Mining est de faire ressortir, dans une masse très importante de données textuelles, l'information utile afin qu'elle devient exploitable par l'informatique. Il intervient donc dans la :

- Recherche d'information (Information retrieval) : Interrogation de textes par concepts, mots-clés, sujets, phrases visant à obtenir des résultats triés par ordre de pertinence, (ex : Google)
- Construction de résumé (Summarization) : Abstraction et condensation d'un texte pour élaborer une version réduite conservant au maximum la sémantique.
- Extraction d'information (Information extraction) : Identification d'éléments sémantiques dans un texte (entités, propriétés, relations, patterns …)
- Interrogation en langage naturel (Question answering) : Interrogation de bases de données en langage naturel. Et notamment
- La classification automatique des documents.

4. Conclusion

Nous avons tenté, tout au long de cette première partie, de présenter les deux technologies Data Mining et plus en détail le Text Mining, qui en résumé est divisé en deux étapes principales, étape d'analyse qui permet de structurer le texte, et une étape d'interprétation de l'analyse, qui fait appel aux méthodes de fouille de données. Le chapitre suivant présente nos expérimentations dans laquelle nous présentons une étude comparative des étapes de prétraitement et les approches de représentation de texte arabe pour le processus de classification automatique des textes.

CHAPITRE 2
PRETRETEMENT ET REPRESENTATION DES TEXTES

1. Introduction

L'information textuelle, est actuellement stockée sous différents formats de fichiers, tels que HTML, XML, CSV, SGML, DOC, PDF, etc. Ces collections sont peu structurées,

Le manque de structure au sein de ces collections volumineuses rend difficile l'accès à l'information qu'elles contiennent, d'où la nécessité aujourd'hui, de chercher comment structurer automatiquement ces corpus pour les rendre utilisables d'une façon rapide et optimale pour y faciliter leurs traitements automatiques et notamment la classification.

Pour pouvoir y appliquer les différentes techniques et algorithmes d'apprentissage, une transformation de ces documents non ou peu structurés est indispensable. La transformation ou le codage de ces documents est une préparation à « l'informatisation ».

Plusieurs approches de représentation des documents textuels ont été proposées dans ce contexte, la plupart étant des méthodes vectorielles, et les principales méthodes de représentation de textes n'utilisent pas d'information grammaticale ni d'analyse syntaxique des mots : seule la présence ou l'absence de certains mots est porteuse d'informations.

Un état de l'art des différentes approches de prétraitement et représentation de textes est développé dans ce chapitre.

2. Caractéristiques de la langue arabe

La langue arabe (العربية, *al ʿarabīya*) est originaire de la péninsule Arabique, L'Arabe, langue sacrée du Coran, connaît une grande stabilité dans un créneau bien précis qui est celui de la littérature classique, des milieux de l'enseignement, la culture officielle et de la presse.

Elle est comme langue officielle de 22 pays, l'arabe est parlé par plus de 300 millions de personnes, et le plus rapide-langue croissante sur le web.

L'alphabet arabe comprend les 28 caractères suivants:

أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق

ك ل م ن ه و ي

En plus de hamza de l'arabe (ء) qui a considéré comme une lettre par certains linguistique arabe. Les lettres (ا و ي) sont des voyelles, le reste sont des consonnes.

L'arabe est écrit de droite à gauche. Les lettres arabes ont des modèles différent en apparaissant dans un mot selon la position de lettre (le début, le milieu ou la fin d'un mot) et si la lettre peut être reliée à elle est allumé les lettres voisines ou pas. par exemple, la lettre (س) a les modèles suivants (سـ) si elle apparaît que au début d'un mot (tel que le mot ساعة que signifie l'horloge); (ـسـ) si la lettre apparaît au milieu d'un mot (tel que le mot يسجل que signifie registrer); (ـس) si la lettre apparaît à la fin du mot (tel que les mot حبس que signifie emprisonnez). En fin, la lettre (س) peut apparaître comme (س) si elle apparaît à l'extrémité d'un mot mais démontée de la lettre située à son côté droite (tel que le mot درس que signifie étuder). Les diacritiques sont des signaux placés au-dessous ou au-dessus des lettres pour doubler la lettre dans la prononciation ou pour agir en tant que voyelle courte.

Les diacritiques arabes incluent: shada, dama, fatha, kasra, double dama de sukon, double fatha, double kasra. Les différents modèles et diacritiques de lettre font à analyse le texte arabe une tâche non triviale.

Les mots arabes ont deux genres, masculins (مذكر) et féminins (مؤنث) ; trois nombres, le singulier (مفرد), conjuguent (مثنى), et le pluriel (جمع); et trois cas grammaticaux, nominatifs (الرفع) accusatifs (النصب) et génitifs (الجر). Un nom a la forme nominative quand il est soumis (فاعل); accusatif quand c'est l'objet d'un verbe (مفعول); et le génitif quand c'est l'objet d'une préposition (مجرور بحرف جر) .Des mots sont classifiés dans trois parties du discours, noms (اسماء) (y compris des adjectifs (صفات) et adverbes (ظروف), verbes (افعال) et des particules principaux (ادوات).

A la fin du mot	Au milieu de mot	Au début du mot
أ, ؤ , ئ , ء	ـئـ	أ
ـب, ب	ـبـ	بـ
ـت, ت	ـتـ	تـ
ـه, ه	ـهـ	هـ
ـم, م	ـمـ	مـ
ـي, ي	ـيـ	يـ
ـغ, غ	ـغـ	غـ

Table 2.1 Etat de transcription des lettres arabes

شَهْدُ	**Miel (cire d'abeille)**
شَهِدَ	**Informer, affirmer, a été présent, a vu**
شَهُدَ	**A fait une déposition**
شُهَّدٌ	**Comme**رُكَّعٌ
شَهْدٌ	**Pluriel de شاهد: des témoins**
شَهْد	**Nom propre féminin, Plante**

Table 2.2 Les diacritique (Les différentes vocalisations du mot « شهد »)

3. Prétraitements

Le prétraitement des textes est une phase capitale du processus de classification, puisque dans les documents textuels de nombreux mots apportent peu (voir aucune) d'informations sur le document concerné.

Le prétraitement des textes inclut beaucoup d'étapes comprenant la réduction de dispositif en utilisant des techniques d'analyse morphologique, et peser de limite.

3.1. Encodage unique des textes :

L'encodage unique des textes en format standard permet de représenter les textes sans aucune déformation au niveau du caractère lors de la lecture. Tous les textes de notre corpus sont représentés avec le codage (UTF-8 : le supporté par le langage java).

3.2. Suppression des caractères inutiles: Cette étape consiste à supprimer:

3.2.1. Les signes de ponctuation :

Supprimer toute séquence de caractères de ponctuation délimitée par des lettres ou des espaces comme la virgule et le point-virgule …etc. Nous prenons en considération l'orientation de quelques caractères qui s'écrivent de droite à gauche dans les textes arabes comme le signe d'interrogation « ؟ », et la virgule « ، ».

3.2.2. Les nombres et les caractères latins :

Nous éliminons toutes les séquences de caractères situées entre deux espaces et contenant des chiffres en arabe «١….٩ » ou des chiffres latin ou romains « 1….9, I.…IX », et nous avons éliminé aussi les caractères latins « A...Z, a …z ».

3.2.3. Les abréviations et les lettres isolées :

Les abréviations de mot arabe, comme :ت pour تاريخ « date », م pour ميلادي, ص pour صفحة « page », ج pour جواب « réponse», س pour سؤال « question». Ou de coordination comme بِ, وَ, فَ, لِ, كَ (bi-, wa-, fa-, li-, ka-…) notés comme des formes isolées à côté des chiffres (ex : 362 ب 54) ou dans les formules mathématique comme (3س+ع= 0).

3.3. Suppression des mots fréquents ou élimination des ”Mots Outils”

Les mots qui apparaissent le plus souvent dans un corpus sont généralement les mots grammaticaux, mots vides (empty words) ou mots outils (stop words) : les articles, les prépositions, les mots de liaisons, les déterminants, les adverbes, les adjectifs indéfinis, les conjonctions, les pronoms et les verbes auxiliaires etc.., qui constituent une grande part des mots d’un texte, mais malheureusement sont faiblement informatifs, sur le sens d’un texte puisqu’ils sont présents sur l’ensemble des textes.

Pour l'arabe, la liste de stopwords inclut des pronoms (هو هي الذي التي هما ..) , les mots liaisons (من, مع, و, في...) adverbes (فوق, تحت, بين..), jours de semaine (السبت, الاحد, الاثنين...) , mois de l'année (جانفي, فيفري, مارس...) . La liste de Stopwords est enlevée parce qu'ils n'aident pas à déterminer la matière de document et ne réduisent pas des dispositifs.

Ces mots doivent être supprimés de la représentation des textes pour deux raisons :

- D'un point de vue linguistique, ces mots ne comportent que très peu d'informations. La présence ou l'absence de ces mots n'aident pas à deviner le sens d'un texte. Pour cette raison, ils sont communément appelés « mots vides ».
- d'un point de vue statistique, ces mots se retrouvent sur l'ensemble des textes sans aucune discrimination et ne sont d'aucune aide pour la classification.
- Leur élimination lors d'un prétraitement du document permet par la suite de gagner beaucoup de temps lors de la modélisation et l'analyse du document.

3.4. Suppression des mots rares

Nous pouvons également vouloir exclure " les mots rares " qui soient définis comme mots qui n’apparaissent qu’une ou deux fois sur un corpus, et qui se produisent dans un pourcentage très bas des documents traités. La suppression de ces mots n'est pas nécessairement justifiée ; certains mots peuvent être très rares, mais très informatifs. Mais il n’est pas possible de construire de statistiques fiables à partir d’une ou deux occurrences ;

Une des méthodes communément retenues pour supprimer ces mots consiste à ne considérer que les mots dont la fréquence totale est supérieure à un seuil fixé préalablement.

3.5. Le traitement morphologique

Cette étape est spécifique à la langue arabe. Consiste à effectuer un traitement au niveau de chacun des mots en fonction de leurs variations morphologiques : flexion, dérivation, composition afin de rassembler les mots de sens identiques. Donc, le but est de regrouper par exemple les termes «لعب» et «لاعب» ou les termes « سافر » et « مسافر » car ils ont la même signification.

Il s'agit de prétraitements relatifs à *la Normalisation Morphologique* de certains caractères arabe, *Stemming* et *Light Stemming*

3.5.1. la Normalisation Morphologique

- ***Hamza et Alif :*** cette normalisation consiste à convertir el « أ », « إ » et « آ » en «ا » car la plus part des textes arabes négligent l'ajout d'El hamza sur El Alif.
  Dans cette étape nous proposons aussi de supprimer totalement ***Alif El Tanwin*** « اً » si elle existe, voici quelques exemples de mots avant et après la normalisation d'El hamza (voir la **Table 2.3**) :

Pour « أ »	امر ← أمر
Pour « إ »	استعمال ← إستعمال
Pour ***el tanwin*** « اً »	استعمال ← استعمالاً
Pour el ***alif avec mada*** « آ »	امر ← آمر

Table 2.3 La normalisation El Hamza

- **Yâ'** ***et*** **el tâ marbouta** : *hamza* ajoute une confusion, si elle situe à la fin de mot, entre ي (lettre *yâ'* finale) et ى (ou *'alif maqsûra*) : Le mot نادي *nâdî* «club», peut être noté نادى, (lu comme *nâdâ*, « convier, convoquer »). De même, « ىء » écrit comme au lieu « ئ» on les converti, comme dans les moteurs de recherches (google par exemple) lorsqu'on écrit dans la barre de recherches le mot أولي les premiers pages qui s'affichent contiennent le mot اولى avec *alif maqsûra*.
  Le même principe pour el « ة (*tâ marbouta)* » telque la normalisation consiste à remplacer el « ة » par « ه *(lettre hâ)* » comme par exemple : جنة ← جنه
- **Le caractère ' ـــ ' (kashida) :** Les typographes font un usage fréquent du caractère ' ـــ ' (appelé kashida) (تطويل), qui permet l'allongement du trait au milieu des mots, pour une meilleure lisibilité, et limiter les espaces blancs sur une ligne justifiée ou même parfois pour des raisons purement calligraphiques. Ce caractère, ne faisant pas partie de l'alphabet arabe, est souvent une source de confusion pendant le traitement des textes arabes.

منزل	مـنـــــزل
حقوق الانسان	حقـــــــوق الانســـــــان
	حقـــــــــــوق الانســـــــــــان
	حقـــــــــــــوق الانســـــــــــــن

Table 2.4 Tatweel (kashida)

- **des signes de vocalisation :** les textes religieux de notre corpus contiennent quelques mots arabes vocalisés. Dans cette étapes on supprime tous les signes de vocalisation : «ً, ٌ, ٍ, ِ, ُ, َ, ّ »,

Double Constante	Aucune Voyelle	Nounation			Voyelle		
بّ /bb/	بْ /b/	بٍ /bin/	بٌ /bun/	بً /ban/	بِ /bi/	بُ /bu/	بَ /ba/

Table 2.5 Diacritiques

3.5.2. L'Analyse morphologique (Light Stemming et le Stemming)

a. Light Stemming :

Light steming (Enracinement lumière) dépend de la suppression des lettres additionnelles du mot seulement c-à-dire : supprimer le préfixe et le suffixe du mot sans transformer le terme à la racine de l'original.

Light steming pour supprimer les caractères additionnels d'un mot où conserve le sens du mot, parce que son idée est basée sur les rejets de la racine donnent significations des mots différentes, même si c'est de la même racine, il vise l'enracinement doux à la sténographie les étiquettes / les mots-clés tout en maintenant un sens, contrairement à l'enracinement (steming) ce qui pourrait affecter la signification des mots

Exemple :

Mot	→	Light stemming
الشجاعة	الشجاعة	شجاع
بلادي	بلادي	بلاد
Mot	→	Le stemming
الشجاعة	الشجاعة	شجع
بلادي	بلادي	بلد

Figure 2.1 Comparaison entre les deux stemmers

b. Le Stemming :

L'idée de steming (enracinement) dépend de la suppression des caractères additionnels d'un mot, puis convertir le terme à la racine de tout mot revient à son origine.

La richesse de la langue arabe augmente la taille des vecteurs de dispositif. Heureusement, l'arabe a son mécanisme de filtrage intégré – où les mots peuvent être tracés dans leur modèle de racine en utilisant le stemming. Les modèles de racine en arabe sont devisées dans trois, quatre, cinq ou six modèles de lettre. Plus de 80% des mots arabes peut être tracé dans des modèles de racine de 3-lettres.

En représentant un mot à son modèle de racine réduisez considérablement le nombre de mots. Cependant, les racines sont sémantiquement semaine dans le car que plusieurs mots peuvent être mises en correspondance avec la même racine et desserrer de ce fait le sens (passé, le présent, ou le futur), les multiples formes du verbe disparaissent (درس) sont réduites au pattern (درس) de racine de trois-lettre .En dépit de ceci, tracer un mot à sa racine réduit la dimension des vecteurs de document. [9]

L'extraction de la racine ou les algorithmes de stemming pour le texte arabe tombent dans deux groupes:

1. Les algorithmes qui enlèvent des préfixes, antéfixe, infixes, suffixes et des postfixes des mots et puis les tracent dans un ensemble de modèles prédéfinis de racine. Dans ce modèle, un mot peut devoir être balayé plusieurs fois avant qu'il puisse être tracé dans son modèle de racine.

Voici la liste des préfixes et les suffixes :

Affixes en arabe	Exemples
Préfixes de la longueur 3	ولل وال كال بال
Préfixes de la longueur 2	ال لل
Préfixes de la longueur 1	ل ب ف س و ي ت ن ا
Suffixes de la longueur 3	تمل همل تان تين كمل
Suffixes de la longueur 2	ون ات ان ين تن كم هن نا يا ها تم كن ني وا ما هم
Suffixes de la longueur 1	ة ه ي ك ت ا ن

Table 2.6 Affixe réglé en langue arabe

Exemple1 :

Mots	préfixe	infixe	suffixe	racine
الشجاعة	**الشجاعة**	**الشجاعة**	**الشجاعة**	**شجع**
الاستماع	**الاستماع**	**الاستماع**	**الاستماع**	**سمع**

Table 2.7 version du mot (الشجاعة) quand éliminer les affixes

Exemple2 :

Est-ce que vous allez vous l'approprier ? - ***أستمتلكونه؟*** Ce mot peut être segmenté ainsi :

Mots	postfixe	suffixe	infixe	préfixe	antéfixe	racine
أستمتلكونه	**ه**	**ون**	**متلك** Corps schématique	**ت**	**أس**	**ملك**
--	←					**--**

Table 2.8 version du mot (*أستمتلكونه*) quand éliminer les affixes

'Segmentation d'un mot arabe'

Les antéfixes sont des prépositions ou conjonctions (question, futur..) ;

- Les préfixes, infixes et suffixes expriment les traits grammaticaux et indiquent :
  - Cas du nom ;
  - Mode du verbe (actif, passif);
  - Modalités : nombre (singulier, duel, pluriel), genre (masculin, féminin), personne (1^{er}, 2^{eme} ou 3^{eme} type);
- Les postfixes sont des pronoms personnels.

2. Algorithmes qui utilisent un poids de lettre et un arrangement d'ordre – où des lettres dans un mot sont données des poids et sont assignées des rangs ou des ordres; et alors la racine est extraite en traitant ces poids et rangs assignés.

Pour l'extraction de racine. Al-shalabi [10], extrait la racine d'un mot en affectant des poids et des rangs aux lettres qui constituent un mot. Les poids sont des nombres réels dans la gamme 0 à 5. L'attribution des poids aux lettres a été déterminée par expérimentation étendu avec le texte arabe. L'alphabet arabe, selon Al-shalabi [10], était devisé dans six groupes comme montré dans cette table

Lettres arabes	Poids
ا ة	5
ي ى ء	3.5
ت ى و	3
أ إ م ن	2
ل س ه	1
Reste de l'alphabet arabe	0

Table 2.9 Assignements des lettres aux poids

Le range ou l'ordre des lettres dans un mot dépend de la longueur de ce mot et dessus si le mot contient le chiffre impair et pair des lettres. Tableau 2.10 montre l'attribution des rangs aux lettres. N est le nombre de lettres dans un mot.

Position de lettre de droite	Rang (si la longueur de mot est pair)	Rang (si la longueur de mot est impair)
1	N	N
2	N-1	N-1
3	N-2	N-2
[N/2]	N/2+1	[N/2]
[N/2]+1	N/2+1-0.5	[N/2]+1-1.5
[N/2]+2	N/2+2-0.5	[N/2]+2-1.5
[N/2]+3	N/2+3-0.5	[N/2]+3-1.5
N	N-0.5	N-1.5

Table 2.10 Ordre des lettres

Après la détermination du poids et du grade de chaque lettre dans un mot, poids de lettre sont multipliées par le range de lettre. Les trois lettres avec la plus petite valeur de produit constituent la racine (lue de droite à gauche).

Voire le tableau suivant :

Mot	المحافظة							
Lettres	ة	ظ	ف	ا	ح	م	ل	ا
Poids	5	0	0	5	0	2	1	5
Range	7.5	6.5	5.5	4.5	5	6	7	8
Produit	**37.5**	**0**	**0**	**22.5**	**0**	**12**	**7**	**40**
Racine	حفظ							

Table 2.11 Un exemple d'employer l'extracteur de la racine

3.6. Le traitement syntaxique

La syntaxe traite les combinaisons et l'ordre des mots dans la phrase. Le traitement ***syntaxique*** identifie et regroupe un ensemble de mots dont la sémantique dépend de leur association. Par exemple, les mots « casque bleu » ne signifient habituellement pas qu'on a affaire à un casque qui est bleu, mais plutôt à une organisation militaire dépendante de l'ONU. L'analyseur syntaxique a pour but d'identifier ce type de cas. La phase d'analyse syntaxique consiste aussi à éliminer des ambiguïtés comme par exemple les problèmes d'homographie. [11]

3.7. Le traitement sémantique

Le traitement ***sémantique*** consiste à extraire la signification des expressions et traiter la polysémie à savoir les différents sens possibles d'un même mot. Par exemple, cette phase permet de différencier le mot « ***base*** » « ***قاعدة***»qui peut correspondre à une « *base militaire قاعدة عسكرية* » ou à une « *base de données قاعدة معطيات* ». C'est une opération laborieuse, qui fait appel aux ontologies, et qui n'est pas aujourd'hui bien maîtrisée et dont l'intérêt en terme de meilleures performances, dans les systèmes de classification, n'est pas toujours démontré. [11]

Remarque :

Le Traitement appliqué à tous les mots arabes, à l'exception du nom de Majesté '**الله**'

4. Représentation des textes

La représentions des textes aussi est la phase la plus importante dans le processus de

catégorisation de textes parce que les algorithmes d'apprentissage ne sont pas capables de traiter directement les textes, plus précisément les données non-structurées comme les images, les sons, les vidéos. C'est pour cette raison, qu'on doit opter pour une façon efficace de représenter les instances à traiter (les textes), la quasi-totalité des systèmes de catégorisation de texte représentent les documents par la présence ou l'absence de termes dans le texte. Ces termes sont les unités minimales constitutives d'un texte, ils peuvent être plus ou moins complexes : chaînes de caractères, mots, racines de mot, groupes des mots (concept) ou une expression.

Cette étape consiste généralement à représenter chaque document par un vecteur, dont les composantes sont par exemple les termes contenus dans le texte, à ces termes on associe des poids pour rendre chaque vecteur exploitable par les algorithmes d'apprentissage, et enfin réduire la dimensionnalité.

4.1. Représentation en « sac de mots » « bag of words »

La représentation des textes la plus simple a été introduite dans le cadre du modèle vectoriel et porte le nom de "sac de mots" « *Bag-of-words* ». Les textes sont transformés simplement en vecteurs dont chaque composante représente un terme.

Dans un premier temps, les termes sont les mots qui constituent un texte. Les mots ont l'avantage de posséder un sens explicite.

On peut le considérer comme étant une suite de caractères appartenant à un dictionnaire, ou bien, de façon plus pratique, comme étant une séquence de caractères non délimiteurs encadrés par des caractères délimiteurs.

Dans les langues comme l'arabe, les mots sont séparés par des espaces ou des signes de ponctuations ; ces derniers, tout comme les chiffres, sont supprimés de la représentation.

Les composantes du vecteur sont une fonction de l'occurrence des mots dans le texte Cette représentation des textes exclut toute analyse grammaticale et toute notion de distance entre les mots : c'est pourquoi cette représentation est appelée « sac de mots »

Ces descripteurs ont un vrai privilège de posséder un sens explicite cependant la représentation en « bag of words » affiche plusieurs anomalies :

- Le problème des mots composés comme : Arc-en-ciel, peut-être et le problème des sigles comme : APN, FAF, IBM. [12]
- La présence parmi les descripteurs, des mots outils, qui constituent une grande part des mots d'un texte, mais qui portent peu d'informations utiles pour classer un texte. [12]

- La distinction des mots d'une même famille en raison de leur variation morphologique (ex : كتب, كتابه, اكتب, كاتب,…) est en général un handicap, car chaque variation a une fréquence très faible alors que les regrouper permettrait d'avoir des fréquences importantes et d'amoindrir le phénomène d'imprécision des fréquences. [12]
- Enfin cette représentation est un regroupement en vrac de tous les mots du document « sac de mots » sans prendre en compte les combinaisons et l'ordre des mots dans la phrase entraine une perte dans la sémantique du texte. [12]

معالج الصيانة هو عبارة عن برنامج يأتي مع الويندوز بكل اصداراته وهو بالأصل عبارة عن ثلاثة برامج مدمجة مع بعض وهي **:**تنظيف القرص **,** تفحص القرص وازالة التجزئة والهدف منه هو التخفيف على الجهاز لكي تزيد سرعته وينصح بأن يكون ذلك كل أسبوع **,** فهو ينظف الهار دسك **Hard disc** من كل الملفات المؤقتة وملفات الانترنت المؤقتة التي تثقل على الجهاز والتي لا عمل لها**.**

Mots	Occurrence	Mots	Occurrence	Mots	occurrence
أسبوع	1	تثقل	1	مع	2
اصداراته	1	ثلاثة	1	معالج	1
ازالة	1	جهاز	1	ملفات	2
أصل	1	ذلك	1	من	1
التي	2	سرعته	1	مؤقتة	2
أن	1	عبارة	2	هار دسك	1
انترنت	1	على	2	هدف	1
برامج	1	عمل	2	هو	3
برنامج	1	عن	2	هي	1
بعض	1	قرص	2	ويندوز	1
تزيد	1	كل	3	يأتي	1
تنظيف	1	كي	1	يكون	1
تفحص	1	لا	1	ينصح	1
تجزئة	1	لها	1	ينظف	1
تخفيف	1	مدمجة	1		

Table 2.12 La représentation de texte en « sac de mots »
Les chiffres et dates sont supprimés de la représentation

4.2. Représentation des textes par des phrases

Malgré la simplicité de l'utilisation de mots comme unité de représentation, Un certain nombre de chercheurs proposent d'utiliser les phrases comme unité de représentation au lieu des mots. Les phrases sont plus informatives que les mots seuls, car les phrases ont l'avantage

de conserver l'information relative à la position du mot dans la phrase. Et ont un degré plus petit d'ambiguïté que les mots constitutifs, et aussi que les phrases ont l'avantage de conserver l'information relative à la position du mot dans la phrase.

L'utilisation de "sac de phrases" entraîne évidemment un problème de taille (pour n mots il y existe potentiellement n^k combinaisons de longueur k).

Pour y remédier, on ne considère pas toutes les séquences possibles mais on tente d'effectuer une sélection des phrases, en privilégiant celles qui sont sémantiquement riche. Dans la phrase

''La personne rêveuse ambitieuse établit un futur lumineux'' **=** ***'' الشخص الحالم الطموح يبني مستقبله الزاهر ''***
Par exemple, on peut dire que des séquences comme
- ***'' Personne rêveuse ambitieuse'', '' شخص حالم طموح ''***
- ***'' Un futur lumineux, " مستقبل زاهر "***
- ***'' Personne rêveuse'', " شخص حالم "***
- **..... ... etc. sont porteuses de sens.**
Alors que les séquences
- ***"Ambitieuse établit'', " طموح يبني "***
- ***'' L'ambitieuse '', " الطموح "***
- **..... …etc. sont insignifiantes.**

Table 2.13 Représentation de texte par phrase

4.3. Représentation des textes avec des racines lexicales (Stemming)

Dans la description du modèle précédent, chaque flexion d'un mot est considérée comme un descripteur différent; en particulier, les différentes formes un verbe sont autant de mots. Par exemple, les mots « يدرس », « درسوا », et « درست » sont considérés comme des descripteurs différents alors qu'il s'agit de trois formes conjuguées du même verbe « درس ».

Pour remédier à ce problème, il faut de considérer uniquement la racine des mots plutôt que les mots entiers (on parle de stem en anglais). Plusieurs algorithmes ont été proposés pour substituer les mots par leur racine ; l'un des plus connus pour la langue arabe est l'algorithme de stemming de S. Khoja Stemmer et Tim Buckwalter Morphological analyzer. [13]

4.4. Représentation des textes avec des lemmes (lemmatisation)

La lemmatisation consiste à utiliser l'analyse grammaticale afin de remplacer les verbes par leur forme infinitive et les noms par leur forme au singulier.

La lemmatisation est donc plus compliquée à mettre en œuvre que la recherche de racines, puisqu'elle nécessite une analyse grammaticale des textes.

L'objectif de la lemmatisation est d'associer à chaque mot, une entrée dans le lexique.

D'une manière générale, on définit un lexique comme un ensemble de lemmes, ce qui correspond plus ou moins à un dictionnaire. Par exemple le lemme de "mangeaient" est "manger". Depuis la fin des années 80, les lemmatiseurs sont capables d'associer à chaque mot d'un texte son lemme grâce à un étiquetage morpho-syntaxique (nom, verbe, adjectif, etc.). [14]

Cette représentation est simple mais présente plusieurs problèmes :

- ✓ La perte de l'information donnée par le contexte syntagmatique, nécessaire à la distinction des lemmes polysémiques (« prix ثمن » n'a pas le même sens dans « prix violent ثمن شديد », « grand prix » ou « prix d'une marchandise ثمن سلعة »).
- ✓ La présence de synonymes, considérés comme des lemmes différents même s'ils font référence au même concept (« mission مهمة » et « délégation انابة » peuvent dénommer la même entité dans un article de journal). [15]

4.5. Représentation par n-grammes

La notion de n-grammes et plus particulièrement bi-grammes et trigrammes (c'est -à-dire avec respectivement n=2 et n=3). La notion de n-grammes introduit par (Shannon, 1948) dans le cadre de systèmes de prédiction de caractères en fonction des autres caractères précédemment entrés. La notion n-gramme de X définit comme une séquence de n X consécutifs. X peut alors être un caractère ou bien un mot.

la construction de n-grammes de caractères et de mots par la notion de déplacement de fenêtre. Ce déplacement se fait par étape, une étape étant soit un caractère ou bien un mot. Les caractères (ou mots) contenus dans la fenêtre ainsi définie constituent les descripteurs d'un corpus. Nous avons par exemple dans la figure suivante les descripteurs de la phrase sous forme de bi-grammes de mots qui sont : "**يوجد عين**", "**عين في**" et "**في الجنة**".

Nous présentons ci-dessous les deux types de n-grammes, caractères et mots :

Figure 2.2 Exemple de N-grammes de mots et de caractères

Les n-grammes de caractères sont les premiers à avoir été utilisés pour une tâche utilisant des données textuelles. Les n-grammes de caractères sont très utilisé dans l'identification de la langue ou encore la recherche documentaire.

De nombreux travaux ont montré l'efficacité des n-grammes comme méthode de représentation des textes pour la classification. Cette technique présente plusieurs avantages, comparativement à d'autres techniques, les n-grammes capturent automatiquement les racines des mots les plus fréquents sans passer par l'étape de recherche des racines lexicales.

5. Sélection de descripteurs (Réduction)

5.1. Pourquoi réduire ?

S'attaquer au problème de la classification automatique de textes signifie aussi s'attaquer à des difficultés du traitement automatique de la langue naturelle. La taille impressionnante du vocabulaire peut s'avérer un obstacle à l'utilisation d'algorithmes plus complexes, pourquoi ? Si l'on utilise directement le vocabulaire contenu dans les textes et que l'on crée un attribut pour chaque mot qu'il contient, on se retrouve avec un espace vectoriel de dimension très élevée. Chacun des textes sera représenté par un vecteur ayant autant de termes qu'il y a de mots dans le vocabulaire. Le traitement d'un tel espace vectoriel demanderait beaucoup de mémoire et de temps de calcul et même il pourrait nous empêcher d'utiliser des algorithmes de classification plus complexes. Utiliser tous ces mots influencerait aussi négativement sur la précision de la classification. En effet plusieurs mots sont vides de sens. Aussi si un mot est présent dans plusieurs documents, c'est donc il ne permettra pas de départager l'appartenance d'un texte qui le contient à l'une ou l'autre catégorie. Ainsi, Il est nécessaire de diminuer d'

avantage et choisir les descripteurs les plus appropriés (ceux qui assureraient les meilleures performances au classifieur), qui vont être utilisés comme vecteurs d'entrées avant de pouvoir utiliser un modèle d'apprentissage. [16]

En effet, les principaux objectifs de la réduction de dimension sont :

- ✓ faciliter la visualisation et la compréhension des données,
- ✓ réduire l'espace de stockage nécessaire,
- ✓ réduire le temps d'apprentissage et d'utilisation,
- ✓ identifier les facteurs pertinents.

5.2. Le nombre de descripteurs conservés

Nous cherchons donc, à supprimer des termes de la représentation des textes, tout en sachant que chaque suppression de terme entraîne une perte d'information ; il faut trouver le bon compromis entre, d'une part, la nécessité de réduire l'espace des descripteurs avec moins de redondances possibles et, d'autre part, le nécessité de garder suffisamment d'informations.

Plusieurs chercheurs dans le domaine ont essayé de réaliser ce bon compromis, comme par exemple (Dumais & all, 1998) construit son modèle à base des SVM en prenant en considération seulement 300 termes sur le corpus Reuters, par contre (Joachims, 1998) pense autrement en considérant que tous les termes du corpus, fournis après les prétraitements (Presque 10000 termes), sont informatifs et sa conservation en entrée dans son modèle est nécessaire. [17]

5.3. Méthodes de sélection de descripteurs

Il existe plusieurs méthodes ou techniques de sélection de termes : pour chaque terme qui représente son importance pour le document où il figure ou pour le corpus complet, puis à sélectionner ces termes les plus importants.

Principalement ces techniques sont développées pour réduire la dimension du vocabulaire. Et sont divisées en deux grandes familles ; **la sélection d'attributs** et **l'extraction d'attributs.**

a. Sélection d'attributs :

Cette technique prend les attributs (mots dans notre cas) d'origine et conserve seulement ceux jugés utiles à la classification, selon bien sur une certaine fonction d'évaluation.

b. Extraction d'attributs :

Contrairement à la sélection, cette méthode crée de nouveaux attributs à partir des attributs de départ, en faisant soit des regroupements ou des transformations.

6. Pondération

La pondération des termes est une mesure statistique, le principe de pondération s'appuie sur l'observation suivante (Rij, 1979) (Sal & McG, 1983) : « la fréquence d'apparition des mots dans les textes en langage naturel est significative de l'importance de ces mots dans le seul but de représenter le contenu de ces textes ». L'intérêt de cette pondération est mieux exploiter l'information contenue dans le document pour améliorer les performances d'un système de classification de textes.

Il y a beaucoup de méthodes pour calculer le poids sachant que, pour chaque terme, il est possible de calculer non seulement sa fréquence dans le corpus, mais aussi le nombre de documents contenant ce terme.

6.1. Formules de pondération

6.1.1. Term frequency (TF)

- Un terme qui apparait plusieurs fois dans un document est plus important qu'un terme qui apparaît une seule fois
- w_{ij}= Nombre d'occurrences du terme t_i dans le document d_i

  TF_{ij}= Fréquence du terme $t_i$dans le document d_i

$$TF_{IJ} = \frac{w_{ij}}{|dj|} \quad (1)$$

6.1.2. Inverse document frequency (IDF)

- Un terme qui apparaît dans peu de documents est un meilleur discriminant qu'un terme qui apparaît dans tous les documents
  - df_i= nombre de documents contenant le terme t_i
  - d= nombre de documents du corpus

$$IDF = \log \frac{d}{df_i} \quad (2)$$

6.1.3. TF-IDF

- TF-IDF signifie Term Frequency x Inverse Document Frequency :

« Mesure l'importance d'un terme dans un document relativement à l'ensemble des documents ».

$$TF_{IDF} = tf_{i,j} * \log \frac{d}{df_i} \quad (3)$$

6.2. Modèles de représentation de document

La représentation de document est une étape préliminaire qui consiste en la représentation de chaque document par un **vecteur**, dont les composantes sont par exemple les mots contenus dans le texte, afin de le rendre exploitable par les algorithmes d'apprentissage. Une

collection de textes peut être ainsi représentée par **une matrice** dont les lignes sont les termes qui apparaissent au moins une fois et les colonnes sont les documents de cette collection. [18]
Plusieurs modèles ont étés proposés pour la représentation de texte, tel que le modèle probabiliste, le modèle séquentiel et le modèle le plus utilisé, qui est le modèle vectoriel.

6.2.1. Le modèle vectoriel

Le modèle vectoriel permet une analyse très efficace de grandes collections de documents. Il a une structure de données simple, sans utiliser aucune information sémantique explicite.
Un grand nombre de chercheurs dans le domaine ont choisi d'utiliser une représentation vectorielle dans laquelle chaque texte est représenté par un vecteur de n termes pondérés.

6.2.1.1. Représentation binaire

Cette représentation est la plus simple et la plus ancienne, elle ne s'intéresse que sur la présence ou la non-présence d'un terme dans le texte, il consiste à utiliser une pondération binaire : 1 si le terme est présent une ou plusieurs fois dans le document, 0 dans le cas contraire.

Cette façon de représenter un texte, est peu informative car elle ne donne pas les informations nécessaires ni sur les occurrences d'un terme dans le document qui peut être une information importante pour l'opération de classification, ni sur la longueur du texte.

6.2.1.2. Représentation fréquentielle

Cette représentation consiste à présenter le texte sous forme de vecteur dont les éléments renseignent non seulement sur la présence ou l'absence d'un terme comme dans un vecteur binaire mais aussi informe sur le nombre de présences du terme dans le texte (fréquence d'apparition des termes).

6.2.1.3. Vecteur TF-IDF

L'idée de base est de représenter les documents par des vecteurs et de mesurer la proximité entre documents par l'angle entre les vecteurs, cet angle étant donc supposé représenter une distance sémantique.
Le principe est de coder chaque élément du sac de mot par un scalaire (nombre) appelé TF-IDF (présenté précédemment) pour donner un aspect mathématique aux documents textes.
Cette représentation donne plus de poids aux **termes qui apparaissent avec une haute fréquence dans peu de documents**. L'idée est que de tels mots aident à discriminer entre textes ayant différent sujet. [19]

	Doc1	Doc2	Doc1	Doc2	Doc1	Doc2	DF	Doc1	Doc2
هذا	0	1	1	1	0.18	0	1	0	0.055
فوج	1	0	0	0	0.09	0.11	2	0.09	0
فرقة	1	0	0	0	0.09	0	2	0	0
و	0	1	1	1	0	0.22	1	0.045	0.011
من	1	0	0	1	0.09	0	1	0.09	0
حياة	0	1	1	1	0	0.11	1	0.09	0.11
عالم	1	1	1	0	0.09	0	1	0	0
كرة	0	1	0	1	0	0	1	0.09	0.11
كيف	1	0	1	1	0.09	0.11	2	0.045	0.055
2014	1	0	1	1	0.09	0.11	2	0.045	0.055
vocabulaire	**Vecteur binaire**		**Vecteur fréquentiel**		**Vecteur fréquentiel normalisé**		**DF et Vecteur TF_IDF**		

Table 2.14 Représentations vectorielles des documents

7. Outils de prétraitement des textes

Nous employons « **WEKA** (Environnement De Waikato Pour L'analyse De La Connaissance) http://www.cs.waikato.ac.nz/ml/weka » et « **RapidMiner** http://rapid-i.com » pour le prétraitement et la classification des textes. WEKA est une suite populaire du logiciel d'étude de machine écrit dans Java, développé à l'université de Waikato. C'est logiciel libre disponible sous GNU General Public License. WEKA fournit une grande collection d'algorithmes d'étude de machine pour le prétraitement de données, la classification, groupement, les règles d'association, et la visualisation, qui peut être appelée par une interface utilisateur graphique commune.

8. Conclusion

Au cours de ce chapitre, nous avons présenté différentes approches de représentation de texte et les avantages de chacune, les méthodes les plus utilisé pour pondérer les termes, et nous avons cité également les principales méthodes de réduction de la dimension.

Comme nous avons déjà souligné, l'étape de représentation de texte est très importante pour la catégorisation de texte pour avoir de bons résultats. La troisième phase du processus de construction du modèle de prédiction est aussi importante.

Dans le chapitre suivant nous présenterons une idée générale et l'intérêt de la classification. Les grandes approches successives, l'approche linguistique, basé sur la construction d'un dictionnaire et l'approche statistique indique les différents algorithmes d'apprentissage, les plus utilisés dans le domaine de la classification des textes.

CHAPITRE 3
CLASSIFICATION AUTOMATIQUE DES TEXTES

I. Introduction

Dans ce contexte, nous allons exposer la classification automatique de texte, plus en détail la classification des textes d'opinions, est définie comme un processus permettant d'associer une catégorie à un texte, en fonction des informations qu'il contient. Elle a pour but d'automatiser l'association d'un texte à une catégorie. C'est un élément important des systèmes de gestion de l'information.

Ce chapitre présente deux points principaux, d'abord la classification de textes qui contient quelques définitions sur les différents jeux de mots: classification, catégorisation et le clustering, ensuite le processus général de la classification de textes avec toutes ses étapes, et les problèmes spécifiques aux textes lors de l'apprentissage automatique. Et après nous présentons quelques applications de la classification de textes Parmi ceux-ci l'analyse de sentiment et la fouille d'opinion plus en détail dans la deuxième point, Qui présent aussi quelque définitions et situé deux grand approches de classification. L'approche lexicale pour construire un dictionnaire comme un modèle d'apprentissage à la classification des textes, et l'approche statistique qui représente les différents algorithmes de classification.

II. Classification des textes

1. Définition de la classification, catégorisation, segmentation

Les termes 'classification' et 'catégorisation' ont des histoires et des origines très différentes. A ce jour, il semble persister une confusion entre ces termes. Aucune définition scientifique n'a pu être trouvée, hormis le grand robert qui a définit la catégorisation comme un processus de classement par catégories et la classification comme étant l'action de distribution par classes. [20]

La classification automatique de documents est un problème connu en informatique, il s'agit d'assigner un document a une plusieurs catégories ou classes. Le problème est différent selon la nature des documents en question, en effet la classification de textes diffère de la classification de documents images, vidéo ou encore son.

La classification de textes est une tâche générique qui consiste à regrouper de manière automatisée des documents, et La classification de textes est définie comme une opération qui

identifie des classes d'équivalence entre des segments de textes en tenant compte de leur contenu informationnel (mots, n-gram, etc.).

La catégorisation automatique de textes correspond à la procédure d'affectation d'une ou de plusieurs catégories ou classes prédéfinies à un texte. Elle correspond à la *classification supervisée* pour l'apprentissage automatique et à la *discrimination* en statistiques alors que la recherche d'informations utilise des termes plus proches de l'application concernée : *filtrage* ou *routage*.

Cette problématique utilise largement des méthodes issues de l'apprentissage automatique et beaucoup d'algorithmes d'apprentissage supervisé lui ont été appliqués (Naïve bayes, K-plus proches voisins KPPV, arbres de décision, machines à vecteurs support SVM, réseaux de neurones, etc...)

La segmentation ou **clustering** c'est L'apprentissage non supervisé consiste à apprendre à classer sans supervision. Au début de processus nous ne disposons ni de la définition des classes, ni de leurs nombres. C'est l'algorithme de classification qui va déterminer ces informations. Nous ne disposons pas non plus de données en entrée qui sont déjà classées, c'est aussi à l'algorithme de découvrir par lui-même la structure plus ou moins cachée des données et de former des groupes d'individus dont les caractéristiques sont communes. Le clustering consiste donc, à diviser les objets (dans notre cas des textes) en groupes sans connaître à priori leurs classes d'appartenance. [21]

2. Pourquoi automatiser la classification ?

On assiste aujourd'hui à un accroissement de la quantité d'information textuelle disponible et accessible d'une manière exponentielle. D'après les derniers chiffres, on parle de plus de 200 millions de serveurs hôtes sur Internet et plus de 3 milliards de pages, la taille des corpus tests utilisés est passée de quelques Méga-octets à plusieurs Giga-octets. [22]

Le nombre de textes à classifier étant énorme, il serait très difficile de pouvoir déterminer de combien de temps a besoin un expert pour associer un texte à une catégorie ? En pratique, il s'agit d'une question difficile à répondre. Assurément, plusieurs variables influencent le phénomène et les lignes qui suivent porteront sur certaines d'entre elles.

Certainement une grande partie du temps consommé pour classer un document est employé dans sa lecture, puis éventuellement à sa relecture. On peut aussi imaginer que la longueur des textes à classer est assez déterminante du temps qui va être requis pour cette opération, et sans doute, d'une personne à une autre, la vitesse de lecture varie. Une fois cette étape achevée, il faut trancher à quelle(s) catégorie(s) ce texte appartient. Au temps de réflexion exigé s'ajoute,

certainement, le temps de se référer à la description des classes et éventuellement de consulter d'autres textes préalablement associés à certaines classes, pour valider la décision. D'autres facteurs interviennent également comme, comme par exemple le nombre de classes qui peut faire la différence : plus il y a de classes différentes, autrement dit plus il y a d'étiquettes possibles pour un texte donné, plus il est difficile de faire un choix parmi celles-ci. Aussi, plus la sémantique des catégories est précise, fine, détaillée, plus il faut faire attention avant d'y associer un document. À cet égard, classer des documents appartenant soit à la catégorie «informatique» soit à la catégorie «mathématiques» est vraisemblablement plus aisée que celle de classer des documents appartenant à l'une ou l'autre des catégories «Intelligence artificielle», «Génie logiciel» et «Système d'information». [23]

En conséquence, nous pouvons résumer les contraintes majeures qui s'opposent au traitement manuel de classification des documents textuels dans les trois points suivants :

- La réalisation manuelle de cette tâche par un expert est extrêmement coûteuse en terme de temps et personnel car il s'agit de lire attentivement chaque texte, au vu de la quantité phénoménale de textes aujourd'hui accessibles (par le biais du réseau Internet en particulier)
- Les traitements manuels sont peu flexibles et leur généralisation à d'autres domaines est quasi impossible; c'est pourquoi on cherche à mettre au point des méthodes automatiques
- Cette opération peut être perçue comme subjective puisque basée sur l'interprétation du document, deux experts peuvent classer différemment un même document, ou encore un même expert peut classer différemment un même document soumis à deux instants différents

3. Architecture du système de classification

Le système que nous réalisons a pour but essentiel de permettre la classification de textes en langue arabe dans un but de catégorisation et d'indexation. Pour ce faire nous nous proposons de définir un processus de traitement permettant d'avoir en entrée un texte brut et de présenter en sortie la catégorisation de ce dernier. Cette catégorisation peut se faire par rapport à une référence existante ou par rapport à un autre texte en entrée. Notre utilisation de la théorie de la distance intertextuelle pour la mise en place d'une métrique de classification nous contraint à l'intégration d'un processus de lemmatisation des textes (Prétraitement). Cette étape est nécessaire car elle prépare les textes en les décomposant ce qui permet l'exploitation des structures grammaticales dans la détection des classes d'équivalence entre segments de textes.

Nous avons exploité la richesse de la grammaire de la langue arabe pour intégrer la notion

de classes grammaticales au niveau du lemmatiseur ainsi que dans la métrique, de cette manière le lemmatiseur fonctionne indépendamment du jeu de la structure adopté et nous introduisons des poids associés aux classes grammaticales au niveau de la métrique. [24]

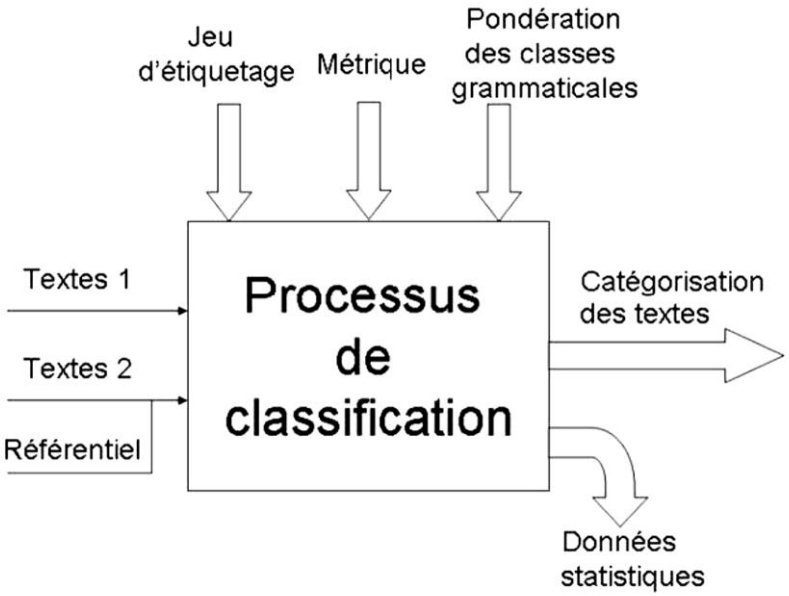

Figure 3.1 Définition du processus

4. Quelques problèmes rencontrés dans la catégorisation de textes arabes

Beaucoup difficultés peuvent s'opposer au processus de catégorisation de textes. Des problèmes connus dans la discipline liés à l'apprentissage automatique supervisé comme la subjectivité de la décision prise par les experts, le sur-apprentissage, etc.. mais aussi des problèmes particuliers liés à la nature des données traitées à savoir des données textuelles comme la polysémie, la redondance, Les variations morphologiques ou même L'homographie, etc.. Nous allons signaler les huit principales Dans ce qui suit :

4.1. Sur-apprentissage

Le sur-apprentissage s'explique par le fait que le modèle de prédiction n'arrive pas à bien classer les nouveaux textes, pourtant il l'a bien fait dans la phase d'apprentissage en classant correctement les textes de la base d'apprentissage.

Pour limiter le sur-apprentissage, on doit sélectionner des termes pour réduire la dimensionnalité. D'après les expériences antérieures, le nombre de termes doit être limité par rapport au nombre de textes de la base d'apprentissage.

Quelques auteurs recommandent d'utiliser au moins 50 à100 fois plus de textes que de termes. En général le nombre de textes d'apprentissage est limité, c'est pour cela on cherche à

agir sur le nombre des termes utilisés en les diminuant, pour éviter ce sur-apprentissage. Sans bien sûr pénaliser le système en supprimant des termes pertinents. [25]

4.2. L'homographie (signification de mot)

Deux mots sont dits homographes si 'ils s'écrivent de la même façon sans forcément avoir la même prononciation. L'homographie est une sorte d'ambigüité supplémentaire. (Ex : ***avocat*** en tant que fruit et ***avocat*** en tant que juriste c'est en français), et en arabe le mot (**قلب**) a trois significations comme un nom

signification de mot	**Phrase**
noyau	في **قلب** الحدث
cœur	اجرى عملية **قلب** مفتوح
centre, moyen	الكرة في **قلب** الملعب

Table 3.1 La signification du mot (قلب) comme nom. [26]

4.3. Polysémie (Ambiguïté)

Un mot possède, dans différents cas, plus d'un sens et plusieurs définitions lui sont associées. Par conséquent, à cause de la polysémie, les mots seuls sont parfois de mauvais descripteurs ; un mot arabe peut avoir plusieurs significations ; Prenons à titre d'exemple le mot arabe non voyelles **ذهب** qui a au moins deux significations : **ذَهَبَ**« aller », **ذَهَبٌ**« or », l'absence des voyelles dans l'arabe couramment écrite génère une ambiguïté sur certains mots qui pénalise la performance des systèmes de traitement de texte arabe.

4.4. Les mots composés

La non prise en charge des mots composés comme : Arc-en-ciel, peut-être, sauve-qui-peut en langue française et إحدى عشر, واحد و عشرونen arabe etc.., Dont le nombre est très important dans toutes les langues, et traiter le mot Arc-en-ciel et le mot واحد و عشرون par exemple en étant 3 termes séparés réduit considérablement la performance d'un système de classification néanmoins l'utilisation de la technique des n-grammes pour le codage des textes atténue considérablement ce problème des mots composés.

4.5. Redondance(Synonymie)

La redondance et la synonymie permettent d'exprimer le même concept par des expressions différentes, plusieurs façons d'exprimer la même chose.

Cette difficulté est liée à la nature des documents traités exprimés en langage naturel contrairement aux données numériques. Exemples des synonymes en arabe sont (اعطى منح بذل وهب) qui signifie (donner), (عائلة اسرة) qui signifie (famille), et (صف فصل) qui signifie (classe). Lors d'une représentation vectorielle d'un document, ces termes sont représentés séparément, et les occurrences du concept sont dispersées. Il est alors important de rassembler ces termes en un groupe sémantique commun.

4.6. La forme de mot selon son cas

La forme de quelques mots arabes peut changer selon leurs modes de cas (nominatif, accusatif ou génitif). Par exemple le pluriel du mot (مسافر) qui signifie que (voyageur) peut être la forme (مسافرون) dans le cas du nominatif (مرفوعة) et la forme (مسافرين)dans le cas d'accusatif/génitif (منصوبة/مجرورة). Refouler arabe de lumière peut manipuler ces caisses.

4.7. Présence-Absence de termes

La présence d'un mot dans le texte indique un propos que l'auteur a voulu exprimer, on a donc une relation d'implication entre le mot et le concept associé, quoique on sait très bien qu'il y a plusieurs façons d'exprimer les mêmes choses, de lors l'absence d'un mot n'implique pas obligatoirement que le concept qui lui est associé est absent du document.

Cette réflexion pointue nous amène à être attentifs quant à l'utilisation des techniques d'apprentissage se basant sur l'exclusion d'un mot particulier.

4.8. Subjectivité de la décision

Parmi les problèmes classiques usuels dans le domaine de l'apprentissage supervisé c'est la subjectivité de la décision prise par les experts qui décident de la classe à laquelle le texte va être attribué.

Certainement après la lecture du texte à classer, l'expert va trancher à quelle(s) catégorie(s) ce texte appartient en se basant sur le contenu sémantique et le contexte du texte et même en consultant d'autres textes préalablement associés à certaines classes, pour valider la décision prise qui ne peut être que subjective.

Les experts humains ne lisent pas de la même manière ! Ne réfléchissent pas de la même manière ! Donc ne classent pas de la même manière !

Ainsi un même document peut être classé différemment par deux experts, ou encore un même document peut être classé différemment par le même expert, soumis à deux instants différents. [27]

5. Applications de la classification

La classification automatique est une technique utilisée dans plusieurs domaines. Sa

capacité prédictive la rend rapide et efficace. Parmi les applications où la classification est utilisée, nous trouvons le filtrage de spam, en effet il s'agit de traiter les messages électroniques textuels, identifier leurs caractéristiques et les classer en deux groupes messages désirés ou non désirés.

Une autre application est la détermination automatique du sujet d'un texte pour le classer automatiquement afin de notifier des personnes intéressées par ce sujet de la présence d'un nouveau texte…, il existe également d'autres applications comme, la catégorisation des documents multimédia, l'organisation des documents, l'indexation automatique des textes, **l'analyse des sentiments et la fouille d'opinion** quand est une côte plus importante dans le temps actuel. Aussi on peut résoudre, par cette technique de classification des problèmes tels que l'identification de la langue d'un document et l'ambiguïsation des termes.

III. Analyse de sentiment et la fouille d'opinion

L'analyse de texte en termes d'étude des sentiments, opinions ou points de vue n'est pas récente. Cependant le domaine de la fouille d'opinion et de l'analyse des sentiments a pris une grande place dès le début des années 2000 avec l'arrivée du Web communautaire et la multiplication des forums sur la toile. Depuis ce jour, le domaine est devenu un enjeu majeur pour toute entreprise désireuse de mieux comprendre ce qui plait et déplait à ses clients ainsi que pour les clients qui souhaitent comparer les produits avant de les acquérir. Par exemple, Morinaga et al. [28] expliquent comment ils vérifient les réputations de produits ciblés en analysant les critiques des clients. Ils recherchent tout d'abord les pages Web parlant du produit concerné et extraient les phrases qui expriment de l'opinion. Ils classent ensuite les phrases selon qu'elles expriment une opinion **négative** ou une opinion **positive** et en déduisent la popularité du produit.

1. Définition :

1.1. La fouille d'opinion

Peut être définie comme une tache ou discipline de l'informatique linguistique, ou bien des foyers permettent d'extraire les opinions des personnes à partir d'un ensemble de documents pertinents pour un sujet donnée ou par le web. L'expansion récente du web encourage des utilisateurs à contribuer et s'exprimer par l'intermédiaire des blogs, des vidéos, des emplacements sociaux de gestion de réseau (*social networking*), etc... Toutes ces plateformes fournissent une quantité énorme de l'information valable que nous sommes intéressée d'analyser. Donné un morceau de texte, les systèmes opinion-mining analysent:

- quel partie est exprimer d'opinion;

- qui a écrit l'opinion;
- ce qui est commenté.

1.2. L'analyse de sentiment

D' autre part, est une sous-tâche de la fouille de donnée Elle consiste de façon générale à déterminer la subjectivité, la polarité (*P*ositive ou *N*égative) et la force de polarité (*F*aiblement *P*ositive, *M*odérément *P*ositive, *F*ortement *P*ositive, etc...) du document sur le sujet – en d'autres termes:

Quel est l'opinion de l'auteur.

2. Les applications d'opinion mining

La fouille d'opinion et l'analyse de sentiment couvrent une étendue large des applications :

- logiciel de cartographie d'Argument (Argument Mapping Software) aide à organiser de manière logique les rapports politiques, par explicitant les liens logiques entre eux. Sous le champ de la délibération en ligne de recherche, des outils comme de Compendium, Debatepedia, Cohere, Debategraph ont été développés pour donner une structure logique à un certain nombre de déclaration de politique générale, et de lier les arguments avec l'évidence pour la soutenir.
- Le Vote Conseillent Des Applications (Voting Advise Applications) aide les électeurs à comprendre quel parti politique (ou autre électeurs) ont des positions plus étroits avec eux. Par exemple : SmartVote, demande à l'électeur de déclarer son degré d'accord avec un certain nombre de décideurs politiques, assortit alors sa position avec les parties politiques.
- L'analyse de contenu automatisée aides traitant la grande quantité de données qualitatives. Il y a aujourd'hui sur le marché beaucoup d'outils qui combinent l'algorithme statistique avec la sémantique et les ontologies, ainsi que la machine Learning avec la surveillance humaine. Ces solutions sont capables d'identifier des commentaires pertinents et attribuer des connotations positives ou négatives (le sentiment que l'on appelle).

Les deux premiers points reflètent les domaines d'application matures, tandis que le troisième secteur est en train d' émerger et les questions de recherche pertinentes. Nous donc nous concentrerons principalement sur ce secteur pour les issues de recherches.

3. Pour quoi elle importe dans les gouvernements

Les applications de la fouille d'opinion sont l'infrastructure de base à grande échelle de collaboration de décideurs des politiques. Ils aident à donner un sens à des milliers d'interventions. Ils aident à détecter un système d'alerte précoce de la perturbation possible

d'une façon opportune, en détectant la rétroaction tôt des citoyens. Traditionnellement, des aperçus ads-hoc sont employés pour rassembler la rétroaction d'une façon structurée. Toutefois, ce type de collecte de données est coûteuse, car elle mérite un investissement dans la collecte de conception et de données ; il est difficile, car les gens ne sont pas intéressés à répondre à des sondages ; et, finalement, il n'est pas très utile, car il détecte " problèmes connus" à travers des interviewes et des questions prédéfinis, mais ne parvient pas à détecter les problèmes les plus importants. La fouille d'opinion est utile pour identifier les problèmes par l'écoute, plutôt que par demander, là en assurant une réflexion plus précise de la réalité. Logiciel de cartographie de l'argument (Argument Mapping software AMS) est alors utile pour s'assurer que les discussions de politique sont logiques et évidence, et fondées sur des données probantes, et ne répétez pas les mêmes arguments à plusieurs reprises. Ces outils seraient finalement utiles non seulement pour les décideurs politiques mais également pour les citoyens qui pourraient plus facilement comprendre les points essentiels d'une discussion et participer au processus de décideurs politiques.

4. Les approches de classification d'opinion

Puisque nous avons besoin d'associer des notes à des textes, nous nous intéressons ici uniquement à la classification d'opinions. Deux grands types des méthodes sont utilisés pour cette tâche. Il y a tout d'abord les approches plutôt linguistiques qui consistent à répertorier le vocabulaire porteur d'opinion, puis à établir des règles de classification selon la présence ou l'absence des mots appartenant à ce vocabulaire. Il existe également les approches mettant en œuvre des outils issus du domaine de l'apprentissage automatique. Nous intéressons ici uniquement à celles de la deuxième famille, qui sur nos données se sont avérées nettement plus efficaces. Les méthodes utilisées dans ce cadre sont issues de la classification dite supervisée, où un classifieur est appris à l'aide d'exemples de données (ici de texte de commentaire) dont on connaît déjà la classe (ici la polarité).

Ces grands deux approches sont résumés dans qui suit :

4.1. Approche linguistique

La principale tâche dans cette approche est la conception de lexiques ou dictionnaires d'opinion. L'objectif de ces lexiques ou dictionnaires est de répertorier le plus de mots porteurs d'opinion possible. Ces mots permettent ensuite de classer les textes en deux (positif et négatif) ou trois catégories (positif, négatif et neutre).

4.1.1. Construction du dictionnaire

Cette méthode lexicale nécessite donc la construction d'un dictionnaire d'opinion. Pour

construire un tel dictionnaire, trois genres de techniques sont possibles :

- la méthode manuelle ;
- la méthode basée sur les corpus ;
- la méthode basée sur les dictionnaires.

La méthode manuelle demande un effort important en terme de temps mais il faut savoir que toutes les autres méthodes nécessitent également de créer initialement, de façon manuelle, un ensemble de mots et expressions porteurs d'opinions. Cet ensemble de mots est appelé *graine*. Il est ensuite utilisé afin de trouver d'autres mots et expressions porteurs d'opinions. Une solution afin d'agrémenter cet ensemble de mots est donc l'utilisation de corpus de textes. Turney [29] propose la méthode suivante : afin de déterminer la polarité de mots ou expressions non classés, il compte le nombre de fois où ces mots ou expressions apparaissent dans le corpus à côté de mots ou expressions déjà classés. Un mot apparaissant plus souvent à côté de mots positifs sera donc classé dans la catégorie positif et inversement. Yu et Hatzi-vassiloglou [30] proposent une méthode similaire, mise à part qu'ils utilisent la probabilité qu'un mot non classé soit proche d'un mot classé afin de mesurer la force de l'orientation du premier nommé.

La méthode basée sur les dictionnaires consiste à utiliser des dictionnaires de synonymes et antonymes existants tels que WordNet. Afin de déterminer l'orientation sémantique de nouveaux mots, on utilise ces dictionnaires afin de prédire l'orientation sémantique des adjectifs. Dans le **WordNet**, les mots sont organisés sous forme d'arbres (voir figure 3.3). Afin de déterminer la polarité d'un mot, ils traversent les arbres de synonymes et d'antonymes du mot et, s'ils trouvent un mot déjà classé parmi les synonymes, ils affectent la même polarité au mot étudié, ou bien la polarité opposée s'ils trouvent un mot déjà classé parmi les antonymes. S'ils ne croisent aucun mot déjà classé, ils réitèrent l'expérience en partant de tous les synonymes et antonymes, et ce jusqu'à rencontrer un mot d'orientation sémantique connue.

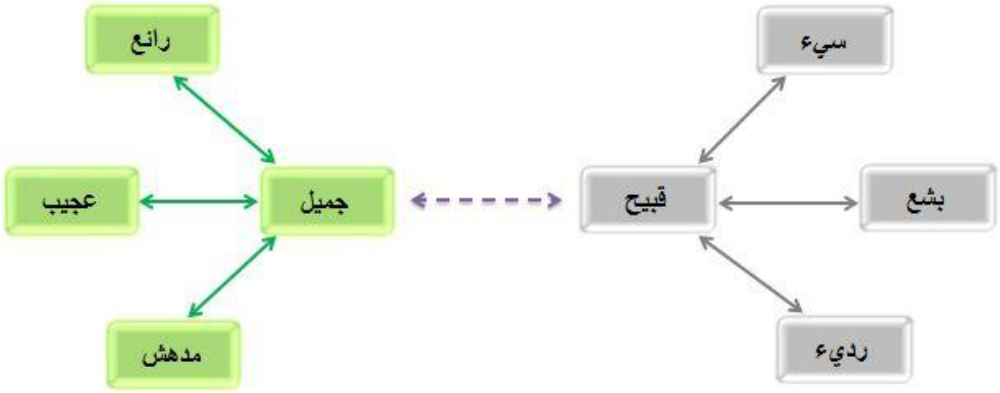

Figure 3.2 Exemple d'arbre de synonymes et antonymes présents dans WordNet (flèche pleine =synonymes, flèche hachurée = antonymes)

La construction du dictionnaire constitué le suivant :

- **Marqueur :** la table du marqueur contient tous les prédicats, les adjectifs et les adverbes construisent à partir du corpus avec leur polarité et l'intensité.
  - Prédicat : أحبaimer, كرهdétester, ظنpenser.
  - Adjectifs : جميلةbien fait, رائعةmagnifique, ركيكةlâche.
  - Adverbe : غنيةriche, مضجرةfatigante, مفيدةintéressante.
- **Intensité :** كثيرا beaucoup, جداtrès, مئة بالمائةcent pour cent
- **Négation :** لا (non, ni, pas), لم لن ليسpas.

4.1.2. La base lexicale WordNet

Le WordNet est un système de référence lexical en ligne. Sa conception est fortement inspirée par les théories psycholinguistiques récentes sur la mémoire sémantique humaine. Les noms, verbes, adjectifs et adverbes sont organisés en ensemble de synonymes (synset), chacun représente un concept lexical. Différentes relations lient les synsets en un réseau sémantique. Plusieurs versions de WordNet existent, la version anglaise est la plus complète. Une version en langue arabe est disponible sur internet (Arabic WordNet). [31]

Arabic Wordnet :

L'Arabic WordNet est une base de données lexicale. Sa conception basé sur Princeton WordNet [32] est construite suivant des méthodes développées pour EuroWordNet est reliée avec l'ontologie SUMO (*Suggested Upper Merged Ontology*). L'ontologie Arabic WordNet contient 9228 concepts « *synsets* » (6252 nominales et2260 verbales, 606 adjectival, et 106 adverbales), contient 18,957 expressions et 1155concepts nommés le fichier base de l'AWN sous format XML [33] contient les quatre balises :

- **Item** : Contient les concepts (*synsets*), les classes et les instances de l'ontologie.
- **Word** : Contient les mots arabes vocalisés.
- **Form** : Contient les Racines des mots arabes « *root* ».
- **Link** : Contient les relations entre les concepts.

4.1.3. Le problème de négation

Les négations paraissent logiquement être des termes importants à détecter, en plus des adjectifs et des verbes, car ils permettent d'inverser la polarité d'une phrase. On propose par exemple d'ajouter des mots dans le dictionnaire d'opinion comme "***ليس جيد***" qui sont utilisés lors de la détection d'un couple ***bien*-négation**. Les négations peuvent être ***ليس, غير, لا,***...

Le problème de la détection de la négation reste un problème très ouvert, les méthodes existantes n'étant pas réellement convaincantes. Ceci est aussi dû aux différentes façons d'utiliser la négation comme le sarcasme ou l'ironie.

Pour finir, il s'agit ensuite de déterminer la polarité d'une phrase à l'aide de ces dictionnaires. La solution la plus simple consiste à compter le nombre de mots positifs et le nombre de mots négatifs présents. S'il y a une majorité de termes positifs, la phrase est déclarée positive.

À l'inverse, si les mots négatifs sont les plus nombreux, la phrase est déclarée négative. Les phrases possédant autant de mots négatifs que de mots positifs peuvent être déclarées neutres.

4.2. Approche statistique

Les méthodes statistiques les plus utilisées sont les méthodes à apprentissage supervisé. Ce type de méthode consiste à représenter chaque commentaire comme un ensemble de variables, puis à construire un modèle à partir d'exemples de textes dont on connaît déjà le label. Le modèle est ensuite utilisé pour attribuer sa classe à un nouveau commentaire non étiqueté.

Pang et al. [34] montrent que des techniques d'apprentissage automatiques offrent de meilleurs résultats que les méthodes linguistiques décrites précédemment. Ils précisent toutefois que les dictionnaires d'opinion utilisés ne sont peut-être pas optimaux. Pour faire leurs comparaisons, ils ont basé leurs expérimentations sur trois méthodes de classification automatique : un classifieur naïf bayésien, un algorithme de Machines à Vecteurs Support et un classifieur basé sur le principe d'entropie maximale.

Mais en fait, peu de travaux sont basés uniquement sur des méthodes statistiques. Le plus souvent, des prétraitements linguistiques sont effectués sur les textes, soit pour réduire le nombre de variables, ou encore pour sélectionner uniquement les traits grammaticaux susceptibles d'exprimer une opinion et ainsi éviter le bruit avec des mots inutiles pour ce type de classification.

4.2.1. Les algorithmes d'apprentissage

L'apprentissage automatique consiste en l'acquisition de connaissances à partir d'observation de phénomènes. Le plus souvent, l'apprentissage est associé à la construction d'un modèle à partir duquel de nouvelles entrées pourront être identifiées. Nous distinguons deux types d'apprentissage : le supervisé et le non supervisé. Ces deux notions sont souvent utilisées pour accomplir des tâches distinctes. Nous présentons ci-dessous quelques algorithmes d'apprentissage supervisé couramment utilisés dans le cadre de catégorisation automatique de texte :

4.2.1.1. K plus proches voisins (KPPV)

k-PPV (*K Plus Proches Voisins*, ou KNN pour *K Nearest Neighbours*) a prouvé son efficacité face au traitement de données textuelles. La phase d'apprentissage consiste à stocker les exemples étiquetés. Le classement de nouveaux textes s'opère en calculant la distance entre la représentation vectorielle du document et celle de chaque exemple du corpus. Dans les *k-PPV*, pour chaque nouvel exemple à catégoriser, le classifieur calcule la similarité du document avec l'ensemble des autres exemples du corpus d'apprentissage. Il sélectionne ensuite les *k* documents les plus proches de *d.* La catégorie de *d* est attribuée par le vote (pondéré ou non par leur similarité) de ces *k* documents.

$$f(d) = \text{argmax}_{cj \in c}(\sum_{i=1}^{k} c(d_i, c_j)) \qquad (4)$$

Ou

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n}(a_r}\,(x_i) - a_r(x_j))^2 \quad (5)$$

L'algorithme ci-dessous montre comment classer un nouvel exemple par la méthode K plus proche voisin KPPV :

Algorithme : algorithme de classification par K-PPV

paramètre : le nombre K de voisin

contexte : un échantillon de l textes classés en $C = c_1, c_2, \dots c_n$classes

1 : **pour** chaque texte *t* **faire**

2 : transformer le texte t en vecteur $t = (x_1, x_2, \dots, x_m)$.

3 : déterminer les k plus proches textes du texte *t* selon une métrique de distance

4 : combiner les classes de ces k exemples en une classe c

5 : **fin pour**

Sortie : le texte t associé à la classe c.

Figure 3.3 Algorithme de KPPV

Pour que cet algorithme soit efficace, il faut une bonne mesure de similarité entre les documents, notamment afin que les attributs non discriminants ne soient pas pris en compte et que ceux qui sont discrimants le soient.

Remarque :

« Calcules de la distance **»**

Les textes étant représentés sous forme vectorielle, comme des points dans un espace à n dimensions, nous déterminons les voisins les plus proches en calculant la distance entre ces points.

Il existe différentes mesures permettant le calcul de la similarité comme : *Le coefficient de Jaccard, La mesure de Cosinus, Le produit d'Inner, et La similarité de Dice.*

4.2.1.2. Naïve Bayes

C'est une méthode de classification probabiliste. Elle consiste à utiliser les probabilités jointes des mots et des catégories pour estimer la probabilité d'une catégorie sachant un texte à classer. Le classifeur de type Naïve Bayes est un catégoriseur de type probabiliste fondé sur le théorème de Bayes Considérons $vj = (v_{j1}, \dots, v_{jd}, \dots, v_{jd})$un vecteur de variables aléatoires représentant un document d_j et Cun ensemble de classes.

En s'appuyant sur le théorème de Bayes, la probabilité que ce dernier appartienne à la classes $c_i \in C$est définie par : $\mathrm{P}(\mathrm{c_i}|\mathrm{v_j}) = \frac{\mathrm{P(c_i)P(v_j|c_j)}}{\mathrm{P(v_j)}}$ (6)

La variable aléatoire v_{jk} du vecteur v_j représente l'occurrence de l'unité linguistique k retenue pour la classification dans le documentd_j.

La classe c_k d'appartenance de la représentation vectorielle v_j d'un document $d_j$est définie comme suit :

$$\mathrm{c_k} = \mathrm{argmax}\, \mathrm{P(c_i \in C)} \prod_{\mathrm{k}} \mathrm{P(v_{jk}|c_j)} \quad (7)$$

En d'autres termes, le classificateur Naïve Bayes affecte au document d_j la classe ayant obtenue la probabilité d'appartenance la plus élevée.

Alors, $P(c_i)$ est définie de la façon suivante :

$$\mathrm{P(c_i)} = \frac{\text{nombre de documents} \in \mathrm{c_i}}{\text{nombre totale de documents}} \quad (8)$$

En faisant l'hypothèse que les $\boldsymbol{v_j}$sont indépendantes, la probabilité conditionnelle $\mathrm{P}(\mathrm{v_j}|\mathrm{c_i})$ est définie ainsi :

$$P(v_j|c_i) = P(v_{jk}|c_i) \quad (9)$$

Pour réellement déterminer à quelle catégorie un document appartient, il faut calculer $P(c_i|v_j)$ Pour chacune des catégories. Étant donné que $P(v_i)$ reste constant pour toutes les catégories, déterminer $P(c_i|v_j)$ se résume juste au calcul de $P(v_j|c_i) * P(c_i)$. (10)

Une telle hypothèse d'indépendance des v_j peut néanmoins dégrader qualitativement les résultats obtenus avec une telle approche.

$\boldsymbol{P(c_i)}$	est la probabilité qui associe le document $\boldsymbol{v_j}$ à la catégorie $\boldsymbol{c_i}$ indépendamment du contenu du document.
$\boldsymbol{P(c_i \mid v_j)}$	représente la probabilité d'appartenance du document $\boldsymbol{v_j}$ à la catégorie $\boldsymbol{c_i}$
$\boldsymbol{P(v_j \mid c_i)}$	est la probabilité selon laquelle, pour une catégorie donnée, les mots du document $\boldsymbol{v_j}$ sont associés à la catégorie $\boldsymbol{c_i}$.
$\boldsymbol{P(v_i)}$	est la probabilité propre du document $\boldsymbol{d_j}$.

Table 3.2 Table d'abréviations

Algorithme : algorithme de classification par Naïve Bayes

Formation

de formation de données *D*, extraire un vocabulaire*V*

$\boldsymbol{N}$←nombre de documents dans *D*

Calcule de paramètre $\boldsymbol{P(c_i)}$ et $\boldsymbol{P(v_j \mid c_i)}$

pour chaque$\boldsymbol{c_i}$in $\boldsymbol{C}$**faire**

$\boldsymbol{N_i}$←nombre de documents dans$\boldsymbol{c_i}$

$$P(c_i) \leftarrow \frac{N_i}{N} \quad (11)$$

$\boldsymbol{text_i}$←le texte de tous de tous les documents dans la classe $\boldsymbol{c_i}$

pour chaque mot$\boldsymbol{v_j} \in \boldsymbol{V}$:

$T_{ji} \leftarrow nombre$ d'occurrences $de\ \boldsymbol{v_j} en\ text_i$

$$P(X = v_j \mid c_i) = \frac{T_{ji}+1}{\sum_i (T_{ji}+1)} \quad (12)$$

Test

Position $\boldsymbol{S}$←toutes les positions dans le document courant qui contiennent des mots dans $\boldsymbol{V}$

Retour$\boldsymbol{c_k}$, où

$$c_k = \text{argmax}\, P(c_i \in C) \prod_{k \in S} P(v_{jk} \mid c_i) \quad (13)$$

Figure 3.4 Algorithme de Naïve Bayes

4.2.1.3. Machines à support de vecteurs

Les « Supports Vectors Machines » appelés aussi « Maximum Margin Classifier» est des techniques d'apprentissage supervisé basées sur la théorie de l'apprentissage statistique ou automatique, est très rapide et efficace pour les problèmes de classification de texte. L'algorithme des SVM est originalement un algorithme mono-classe permettant de déterminer si un élément appartient (qualifié de positif) ou non (qualifié de négatif) à une classe, si on a un document $\boldsymbol{d}$ est représenté par un vecteur($\boldsymbol{t_{d1}}, \boldsymbol{t_{d2}}, \ldots \boldsymbol{t_{dn}}$) des mots qui le compose. Un SVM simple peut seulement séparer deux classes : une classe positive $\boldsymbol{L_1}$(indiqué par$\boldsymbol{y = +1}$) et une classe négative $\boldsymbol{L_2}$(indiqué par$\boldsymbol{y = -1}$). Le but de SVM est de trouver un classificateur qui sépare au mieux les données et maximise la distance entre ces deux classes. Avec SVM, ce classificateur est un classificateur linéaire appelé **hyperplan** comme le montre dans la **Figure 3.5** qui suit, cet hyperplan sépare les deux ensembles de points.

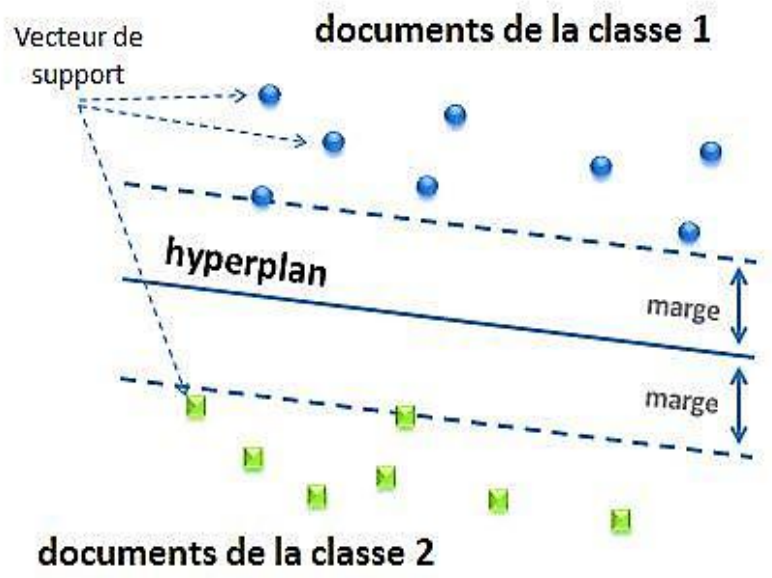

Figure 3.5 Le principe de SVM

4.2.1.4. Arbres de décisions

Les arbres de décision sont plus populaires des méthodes d'apprentissage. Les Algorithmes connus sont ID3 et C4.5. Ils sont également populaires pour la classification de document.
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.

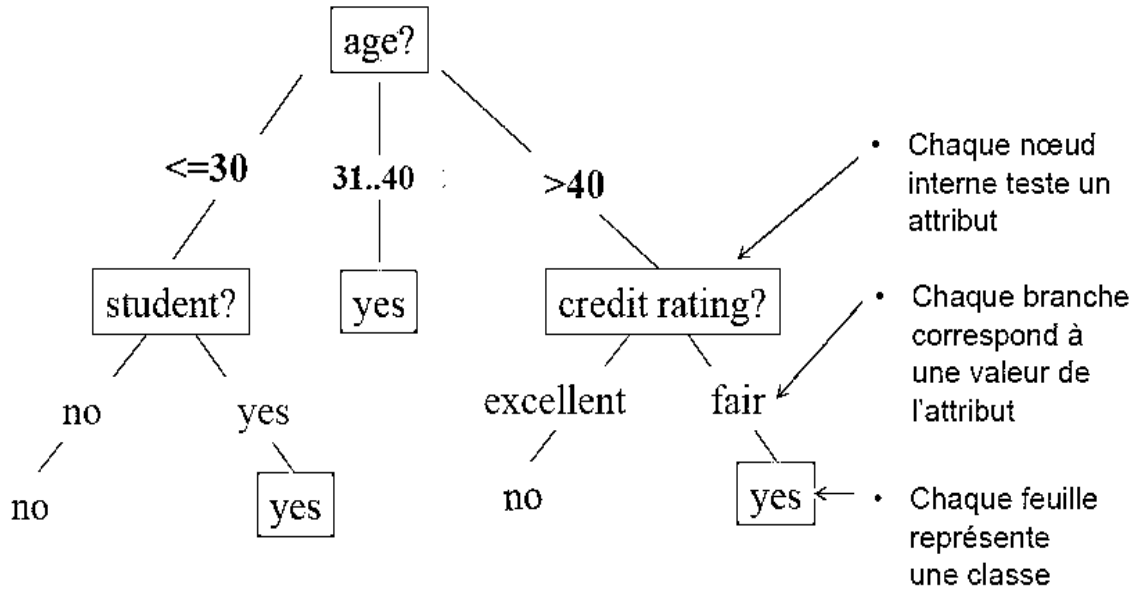

Figure 3.6 L'arbre de décision

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 créer autant de descendants que de valeurs possibles pour ce test.

4.2.1.5. Réseaux de neurones

Les réseaux de neurones : c'est une structure constituée de suite successive de couches de nœuds et qui permet de définir une fonction de transformation non linéaire des vecteurs d'entrées (composés dans le cas de classification des mots pondérés de leur poids) en vecteur de catégories. La disposition des neurones dans le réseau ainsi que le nombre de couches utilisées ont une influence sur le résultat de classification.

Comparés aux autres méthodes de classification par apprentissage supervisé, les réseaux de neurones ont l'inconvénient que le coût d'apprentissage est assez élevé.

Les réseaux de neurones artificiels sont habituellement utilisés pour des tâches de classification. Par analogie avec la biologie, ces unités sont appelées neurones formels. Un neurone formel est caractérisé par :

- Le type des entrées et des sorties ;
- Une fonction d'entrée ;
- Une fonction de sortie.

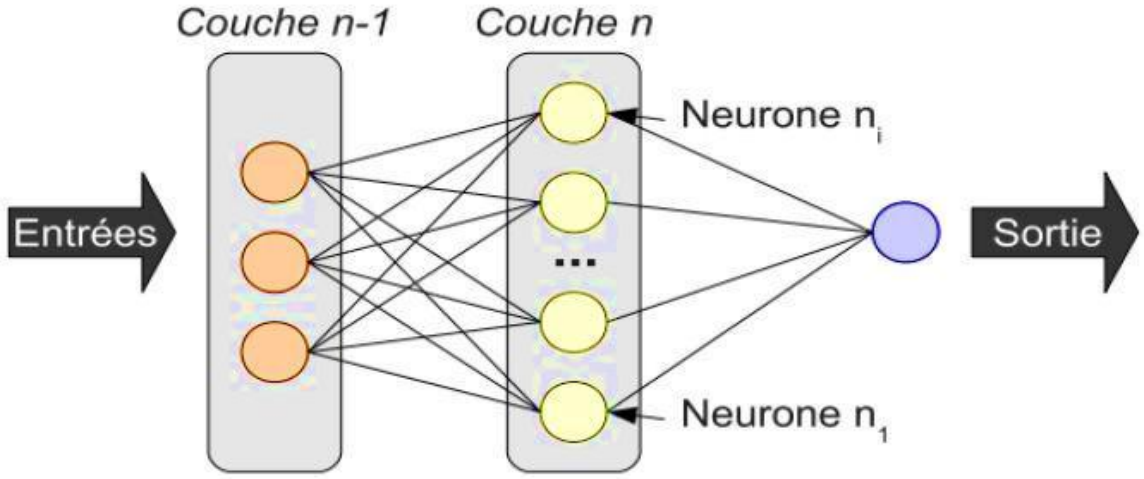

Figure 3.7 Architecture générale d'un réseau de neurones artificiels

5. Classification des opinions en arabe

La disponibilité des corpus annotés pour la formation et le test est très importante pour permettre le progrès sur des systèmes d'identification de sentiment. Le rassemblement de ces données (et en particulier des annotations) peut être très de main-d'oeuvre. Heureusement, un certain nombre de groupes de recherche ont élaboré et publié les corpus d'analyse de sentiment arabe, et nous passons en revue ces travaux ici.

- AWATIF est un multi-genre, corpus de multi-dialecte pour SSA arabe construit par Abdul-Mageed et Diab [35, 36]. AWATIF est extrait à partir de trois ressources différentes: la Penn arabe Treebank (PATB), qui est une collection existante d'histoires de fil de nouvelles dans différents domaines (par exemple sports, la politique, finances, etc.),
- Dans un travail relatif, Abdul-Mageed et Diab [37] utilisent une machine Learning de traduction pour traduire les lexiques anglais disponibles (par exemple, à partir de [38] et [39]) en arabe. Ils recherchent 229,452 entrées, y compris des expressions généralement utilisées dans des médias sociaux.
- Le corpus d'opinion pour l'arabe (OCA) est un corpus de texte de sites de movie

review par Rushdi-Saleh et autres. [40], et inclut une version anglaise parallèle appelée EVOCA. Le corpus se compose 500 revues, à moitié négatives et à moitié positives. Les revues crues ont contenu un certain nombre de défis ce que les auteurs ont essayé de fixer manuellement, y compris le filtrage hors de faux et indépendants commentaires, romanisation de l'arabe, revues de multi-langue, et movie review qui étaient plus d'opinion des thèmes culturels et politiques d'un film. L'OCA et l'EVOCA ont exécuté le prétraitement standard sur le corpus, y compris la correction des erreurs d'épellation et supprimer les caractères spéciaux, et également ont fait l'unigram, le bigram, et les trigrams disponibles pour l'ensemble de données.

- Lexique subjectif de MPQA et corpus de support d'opinion arabe: un autre corpus pour support d'opinion arabe et le lexique de subjectivité est proposé par Elaranoty et autres [41], qui ont créé un Corpus de nouvelles arabes. Ils ont rampé le mb 150 de nouvelles arabes et ont annoté manuellement 1 mb de corpus pour le support d'opinion. Le corpus de titulaire d'opinion a été annoté par trois personnes différentes.

- Arabe Appelé L'identification d'Entité (ANER) [42] a été employée pour extraire des noms à partir des documents. L'arabe proposé le lexique de subjectivité contient des indices forts et bien que des indices subjectifs faibles en traduisant manuellement de Lexique de MPQA [43].

- Un lexique arabe pour Revues D'Affaires a été proposé par Elhawary et Elfeky [44]. El-Hawary et Elfeky [44] ont utilisé le graphe de similarité de construire un lexique arabe. Le graphe de similarité est un type de graphe où les deux mots ou phrases ont un avantage si elles sont semblables à polarité ou sens. Le poids de l'arête représente le degré de similitude entre deux nœuds. Habituellement, ce graphique est construit d'une manière non supervisée basée sur lexicales co-occurrences de grande Web corpus.

- L'utilisation de la grammaire locale est une autre méthode qui peut être utilisé pour Extraire des caractéristiques de sentiment [45] . Ahmed et al. [46,47] ont appliqué cette approche aux documents du domaine des nouvelles financières. . Ils ont évalué le système manuellement et ont atteint des taux de précision entre 60-75% pour extraire le sentiment portant phrases.

- Evoca corpus. Ils utilisent à la fois Support Vector Machines (SVM) et Naive Bayes (NB) classificateurs, déclaré 90% de F-mesure sur l'OCA et de 86,9% sur Evoca utilisant SVM.

- Elhawary et Elfeky [44] présenter un système d'analyse de sentiment sur arabe des revues d'affaires, avec l'objectif spécifique de la construction d'un moteur de recherche Web qui serait annoter automatiquement et renvoyé les pages avec des scores de sentiment. Le système comporte plusieurs éléments.
- Le papier rapports une f-mesure de 81,70% en moyenne dans tous les domaines de documents positifs et 78,09% F-mesure pour les documents négatifs. La meilleure F-mesure est obtenue dans le domaine de l'éducation (85,57% pour la classe positive et 82,86% pour la classe négative).
- D'autres techniques utilisent les caractéristiques linguistiques de la langue arabe afin d'effectuer l'analyse des sentiments, par l'analyse de la structure grammaticale de l'arabe [48] et les caractéristiques morphologiques arabe spécifiques [35, 49,50].

Farra et al. [48] ont proposé la classification de sentiment au niveau de la phrase arabe, à l'aide de deux différentes approches: une approche grammaticale et une approche sémantique. L'approche grammaticale est basée sur la structure grammaticale arabe et combine la phrase de structure verbale et nominale dans une forme générales fondées sur l'idée de l'acteur / l'action. Dans cette approche, les sujets verbaux et les phrases nominales sont les acteurs et les verbes sont des actions. Ils marquent manuellement des balises d'action et acteur dans les phrases et utilisé ces balises que des fonctionnalités.

La deuxième approche proposée par Farra et al. [48] combinée les caractéristiques syntaxique et sémantique en extrayant certaines caractéristiques comme la fréquence des mots positifs, négatifs et neutres, la fréquence des caractères spéciaux (par exemple, "!"), la fréquence des mots importants (par exemple, «vraiment» et "En particulier"), la fréquence des mots concluants et contradiction, etc.

IV. Conclusion

La catégorisation de textes s'est avérée au sur des dernières années comme un domaine majeur de recherche pour les entreprises comme pour les particuliers.

Ce dynamisme est en partie dû à la demande importante des utilisateurs pour cette technologie. Elle devient de plus en plus indispensable dans de nombreuses situations où la quantité de documents textuels électroniques rend impossible tout traitement manuel.

La catégorisation de textes a essentiellement progressé ces dix dernières années grâce à l'introduction des techniques héritées de l'apprentissage automatique qui ont amélioré très significativement les taux de bonne classification.

Nous avons tenté dans ce chapitre de définir la classification de texte, précisément la classification automatique d'opinion, ainsi que les notions nécessaires pour l'entame de la suite de ce mémoire.

Dans le chapitre suivant, nous présentons l'approche proposée pour réaliser un système de classification des textes d'opinion.

CHAPITRE 4
APPROCHE PROPOSEE

1. Introduction

Dans ce chapitre, nous présentons un état de l'art pour mieux situer et expliquer la méthode que nous avons développée pour la classification des textes d'opinion. Ainsi que quelques algorithmes implémentés. Puis nous présentons, notre corpus, utilisé pour l'évaluation**.**

2. Présentation de l'approche de classification

L'approche linguistique est une approche de classification permettant de classer des textes selon l'opinion qu'ils expriment. Elle consiste à étiqueter et répertorier le plus de mots porteurs d'opinion, ces mots permettant par la suite de classer les textes.

2.1. Dictionnaire

Nous avons fait le choix de construire deux lexiques distincts. Le premier d'entre eux contient tous les mots porteurs d'opinion positive et le second tous les mots porteurs d'opinion négative. Pour trouver les mots exprimant une opinion et les classer, nous avons tout d'abord séparé le corpus d'apprentissage en plusieurs parties en fonction des notes attribuées à chaque commentaire. Pour commencer nous avons appliqué, sur chacun des dix sous corpus, un analyseur syntaxique afin de lemmatiser et étiqueter chaque mot du texte. Nous nous sommes basés sur l'hypothèse que les adjectifs et les verbes étaient les deux traits grammaticaux les plus utilisés pour exprimer des opinions. Nous avons donc filtré les mots selon leur trait grammatical et leur fréquence dans chaque sous corpus, et conservé les adjectifs et les verbes ayant le plus d'occurrences. Les lexiques ont ensuite été nettoyés manuellement afin de supprimer les termes n'exprimant a priori aucune opinion, ou encore les termes ambigus. Par exemple, le mot "*terrible*" "مهولة" n'apparait dans aucun des lexiques car il peut exprimer les deux types d'opinion.

Nous avons fait le choix de construire les dictionnaires d'opinion manuellement pour qu'ils ne contiennent que des mots vraiment spécifiques au corpus étudié. Nous pensons en effet que les lexiques d'opinion construits à l'aide des méthodes basées sur les dictionnaires (tels que WordNet), où l'on détermine la polarité des mots en fonction de leur synonymie, sont un peu trop aléatoires car beaucoup de mots peuvent avoir plusieurs sens selon le contexte. Nous

avons aussi jugé que le corpus étudié n'est pas adapté aux constructions de lexiques d'opinion basées sur les corpus, les commentaires étant en règle générale très courts. [51]

Au final, **3560** mots a priori porteurs d'opinion ont été classés dans deux catégories. Le lexique de mots positifs contient **1560** éléments et le lexique de mots négatifs en contient **2000**. La construction de notre Dictionnaire (lexiques) a été réalisée à l'aide de [52] [53] [54].

La table 4.1 présente des exemples de termes contenus dans les deux lexiques.

Mots ayant une Tonalité Universelle positive	سريع, ثقة, ساعد, اقتصاد, سهل, نعمة, فضل, شكر, نجاح, فاخر, جميل, ليس ضعيف, جيد, فرح, احترام, مسؤول, متانة, صبر, صمود, صلابة, مقاومة, مهذب, منور, ممتاز, جيد, لطيف, رقي, رائع, رزين, صفاء, ...
Nbr des mots positive	1560 Mots
Mots ayant Une Tonalité universelle négative	شائبة, تافه, بليد, جرم, فشل, خوف, كسل, قساوة, ضعف, توتر, رديء, سيء, حرمان, عائق, مانع, عسر, حرج, ضيق, صعوبة, مخاطرة, زلة, ركود, كساد, غلطة ,هفوة, مجازفة, ذعر, سرق, ذل , هوان, تعب, سوء, ملل, هلع, رعب, ...
Nbr des mots négative	2000 Mots
Les mots outils	من, على, الذي, التي, بين, ثم, الذين, اللواتي, إلى, ذلك, هم, هن, هما, مارس, أفريل, ماي, الجمعة, السبت, الاحد,....
Nbr des mots outils	385 Mots
Mots ayant une tonalité d'une force (intensité)	حقا, جدا, طبعا, جد, ابدا, قطعا ,يمكن ,ربما ,على الأرجح ,من المحتمل , اطلاقا,..
Les mots de négation	لا, ليس, غير, لن, لم, عدم, عديم, فقد, دون, خالي,...
Nbr des mots de négation	25
Les mots exceptionnels	nom de Majesté 'الله'
Nbr de mots exceptionnels	1
$\sum$ **Motes de dictionnaire**	3971 Mots

Table 4.1 Dictionnaire des termes ayant une tonalité positive ou négative

2.2. Les algorithmes utilisés

a. Module de préparation des données ‘ Prepare_data’

Comme son nom l’indique, cette procédure prépare et traite les données avant la modélisation, c’est l’étape préliminaire.

Debut

Pour chaque document du corpus faire
Diviser le document en unités lexicales (mots) ;
Extraire les mots vides (stop words) ;
Supprimer les signes de ponctuation ()[]{}=:?!;-,_"+*♥/.",<>≥≤%«»&, ;
Supprimer les chiffres ;
Supprimer les caractères latins [a-zA-Z];
Appliquer la Normalisation Morphologique sur le document ;
Appliquer l’algorithme de ‘Light Stemming’ / ‘Stemming’ ;
Calculer l’occurrence pour chaque mot dans le document ;
Représenter chaque document du corpus considéré sous forme d’un vecteur caractéristique constitué des N mots sélectionnés précédemment ;
Construction de la matrice Terme-Document [document, terme] ;
Enregistrer la matrice ;

Fin

b. Module de classification

Ce module permettant de classier un texte d’opinion selon leur polarité (positive, négative). Tout d’abord il utilise le module précédent (Prepar_data) puis calcule le nombre de mots le plus fréquents dans le texte après la comparaison entre les termes de ce texte et les termes du dictionnaire, sachant que les termes ayant des termes précédents de type négation aussi sont calculés avec le nombre totale des termes extraits.

Début

Prepare _data ;

Entrer le Dictionnaire de polarité

Entrer l'ensemble des points (la matrice terme-Document) ;

Pour chaque terme ∈ document **faire**

Fait la comparaison entre la matrice et le dictionnaire ;

Si le mot ∈ dictionnaire positive **alors**

Incrémenter le nombre des termes positive

Si non

Si le mot ∈ dictionnaire négative **alors**

Incrémenter le nombre des termes négatif ;

Si le Nbr_tp > Nbr_tn **alors**

classe= positive ;

Si non

Si Nbr_tp < Nbr_tn **alors**

classe= négative ;

Si non class= neutre ;

Si ∃ négation avant le mot **alors**

Inverser la classe (classe= inverse) {

Si classe= positive **alors** classe= négative

Si non classe= négative **alors** classe= positive}

Fin pour

Enregistrer les résultats ;

Fin.

2.3. Présentation de notre corpus de textes d'opinion

la classification supervisée nécessite des exemples (données étiquetées) afin de construire le « corpus d'apprentissage ».Ce corpus ayant un impact direct sur l'apprentissage des règles, et par conséquent sur la classification, il est nécessaire que les exemples soient représentatifs de l'apprentissage des règles , et par conséquent sur classification, il est nécessaire que les exemples soient représentatifs de l'ensemble des données. Cette hypothèse est généralement difficile à vérifier. En classification d'opinion, ou plus généralement en classification de textes, les corpus étudiés sont assez restreints car l'étiquetage des exemples

est souvent effectué à la main. Ceci entraîne un coût élevé et ne permet donc pas l'obtention d'un gros corpus d'apprentissage.

Pendant notre travail, nous avons utilisés un corpus de plusieurs commentaires sur des articles, recueillis à partir des journaux arabe disponibles sur le net (BBCNewsبالعربي, Eshoroukالشروق, Al_Khabarالخبر, SFU_Review_Corpus_Raw [55]). L'ensemble touchait plusieurs thèmes différents (économique, politique, culture, sport,…). Notre but était effectué une classification pour pouvoir déterminer la polarité des commentaires : Positive, Négative. C'est cela on a appliqué un ensemble des prétraitements manuels sur ce corpus.

Exemples des commentaires :

La table suivante présente un exemple de chaque type d'opinion que nous devons classifier.

	Polarité	Les commentaires en français	Les commentaires en arabe
01	**Positive**	Merci pour ce que tu as écrit Mr amine zaoui celui qui dirige ministère de la culture en Algérie est celui que le dépense sans dignité.	شكرا على ماكتبت سيد الأمين زاوي إن وزارة الثقافة بالجزائر يستلمها من يصرف المال العام بغير وجه حق.
02	**Positive Fort**	Félicitation à nous tous, vous avez fait le bon choix. La journaliste Leila bouzidi est réellement compétente et d'une personnalité professionnelle lui permettant de bien gérer sa courrière que dieu soit avec elle.	مبروك علينا وعليكم لقد احسنت الاختيار. الصحفية القديرة ليلى بوزيدي حقا هي متمكنة وذات شخصية مهنية متحكمة في إدارة مهنتها أعانها الله ووفقها.
03	**Négative**	Franchement, c'est article non professionnel et rigueur aussi de la part d'un expert connu voulant la réussite de l'équipe national	مقال غير احترافي صراحه من اعلامي كبير يحرص على نجاح المنتخب وقاسي كذلك
04	**Négative Fort**	Franchement, je n'ais pas aimé le style de l'écrivain, et aussi son point de vues envers ce sujet et j'ai haï ses moqueries pour l'art.	صراحة لم أحب أبدا أسلوب الكاتب ولا رأيه في الموضوع وكرهت استهزاءه بالفن.
05	**Neutre**	Non, c'est son point de vue personnel.	كلا انه رأيه الشخصي.

Table 4.2. Exemples de commentaires. [56]

3. Conclusion

Nous avons tenté, tout au long de ce chapitre, de présenter la technique de l'approche

linguistique. Cette technique consiste à construire un dictionnaire d'opinion manuellement avec l'aide de techniques simples de Traitement Automatique des Langues. Ce dictionnaire permet ensuite de classer les textes selon leur polarité, positive ou négative.

Les techniques traditionnelles ne permettent de traiter les cas de conflit, où le texte contient les mots positifs et négatifs. Nous avons proposé une solution simple consiste à attribuer à la classe dont le nombre de terme est majoritaire.

Notre deuxième contribution est de traiter le cas de négation par article et négation sémantique, où les mots comme « عدم » joue le même rôle que « sans دون » qui n'est pas pris en compte dans les approches lexicales proposées pour la langue arabe.

Dans le chapitre suivant, nous décrivons enfin le système résultant de notre réalisation par notre approche proposée.

CHAPITRE 5

IMPLEMENTATION ET EXPEREMENTATION

1. Introduction

Nous présentons dans ce chapitre, les outils exploités pour le développement du logiciel tels que le choix du langage de programmation, l'environnement de programmation, ainsi que l'ensemble des résultats des expérimentations par toutes les approches proposées. Nous terminons par une conclusion.

2. Configuration matérielle et logicielle

- ✓ Un PC Pentium 2 à 3GHZ et 4Go de RAM.
- ✓ Microsoft Office 2010 Professionnel.
- ✓ NetBeans IDE 6.8.

3. Les Outils de Développements

3.1. Création des programmes avec java

a. Introduction à java :

Java a été conçu par James Gosling en 1994 chez Sun. L'idée était d'avoir un langage de Développement simple, portable, orienté objet, interprété. Java reprend la syntaxe de C++ en le simplifiant. Java offre aussi un ensemble de classes pour développer des applications de types très variés (réseau, interface graphique, multi-tâches, etc.)

b. Caractéristiques du langage Java

- Il est la langue forte contient de nombreux outils pour les aider dans l'écriture de logiciels.
- Java est un langage de programmation moderne développé par Sun Microsystems (aujourd'hui racheté par Oracle). Il ne faut surtout pas le confondre avec JavaScript (langage de scripts utilisé principalement sur les sites web), car Java n'a rien à voir.
- le fait que le langage Java langue moderne a permis d'éviter les inconvénients de plusieurs langues avant, le plus important de ces défauts accès direct à des emplacements de mémoire pour le programme, qui mène à la faiblesse de la confidentialité des informations et facilement détruit.
- Une de ses plus grandes forces est son excellente portabilité : une fois votre programme créé, il fonctionnera automatiquement sous Windows, Mac, Linux, etc.
- On peut faire de nombreuses sortes de programmes avec Java :
  - des applications, sous forme de fenêtre ou de console ;

- des applets, qui sont des programmes Java incorporés à des pages web ;
- des applications pour appareils mobiles, avec J2ME ;
- et bien d'autres ! J2EE, JMF, J3D pour la 3D...

Langage Java, comme tous les autres langages de programmation n'est pas sans défauts, et le langage Java peut être considéré comme relativement lent. L'avantage de la vitesse est importante, mais elle doit sacrifier certaines caractéristiques pour obtenir les caractéristiques les plus importantes.

c. Les différentes éditions et versions de Java

- Sun/Oracle fournit gratuitement un ensemble d'outils pour permettre le développement de programmes avec Java. Ce kit, nommé JDK, est librement téléchargeable sur le site web de Sun http://java.sun.com
- Le JRE (Java Runtime Environment) contient uniquement l'environnement d'exécution de programmes Java. Le JDK contient lui-même le JRE. Le JRE seul doit être installé sur les machines où des applications Java doivent être exécutées.
- Sun définit trois plateformes d'exécution (ou éditions) pour Java pour des cibles distinctes selon les besoins des applications à développer :
  - Java Standard Edition (J2SE / Java SE) : environnement d'exécution et ensemble complet d'API pour des applications de type desktop. Cette plate-forme sert de base en tout ou partie aux autres plateformes.
  - Java Enterprise Edition (J2EE / Java EE) : environnement d'exécution reposant intégralement sur Java SE pour le développement d'applications d'entreprises.
  - Java Micro Edition (J2ME / Java ME) : environnement d'exécution pour le développement d'applications sur appareils mobiles et embarqués dont les capacités ne permettent pas la mise en œuvre de Java SE.
- Sun fournit le JDK, à partir de la version 1.2, sous les plates-formes Windows, Solaris et Linux.
- La version 1.3 de Java est désignée sous le nom Java 2 version 1.3.
- La version 1.5 de Java est désignée officiellement sous le nom J2SE version 5.0.
- La version 6 de Java est désignée officiellement sous le nom Java SE version 6.

Chaque version de la plate-forme Java possède un numéro de version et un nom de projet. A partir de la version 5, la plate-forme possède deux numéros de version :

- Un numéro de version interne : exemple 1.5.0.
- Un numéro de version externe : exemple 5.0.

3.2. Stockage de données dans des fichiers txt

L'enregistrement des données se fait dans des fichiers de type txt. Et tous les textes de notre corpus sont représentés avec le codage (UTF-8 : le supporté par le langage java).

4. Interface principale

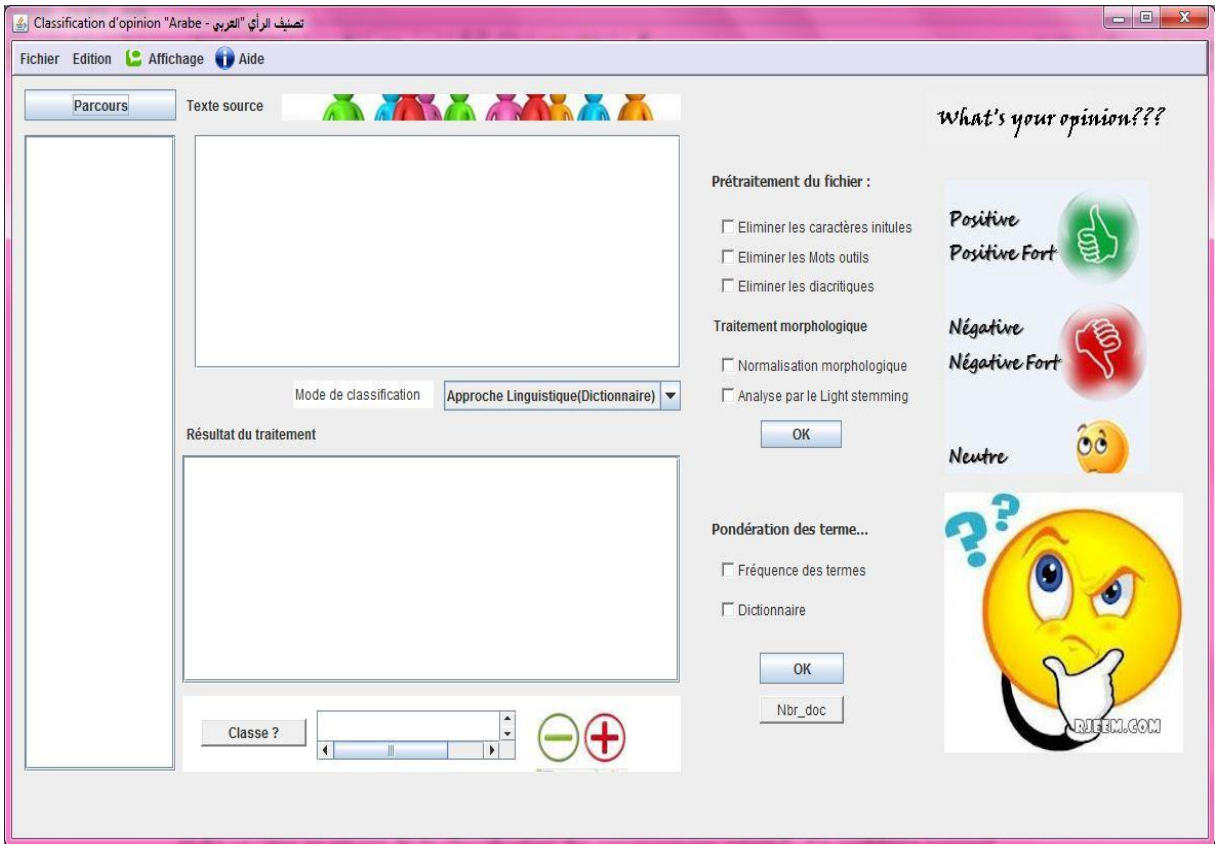

Figure 5.1 Interface principale

5. Résultats expérimentant

L'objectif des travaux que nous avons menés est d'effectuer une classification monologue de textes d'opinion arabes en utilisant l'approche de classification linguistique basée sur le dictionnaire, et une étude comparative des approches de prétraitement et représentation des textes.

Les performances des approches par les expérimentations sont les suivantes:

5.1. Résultats du prétraitement des textes

Tous les documents sont transformés en une matrice des occurrences des termes dont les colonnes correspondent à tous les termes du vocabulaire et les lignes correspondent aux documents du corpus.

Après la réduction, la matrice résultante contient tous les mots des textes avec ses attributs

<table>
<tr><td colspan="11">■ Prétraitement de texte</td><td></td></tr>
<tr><td colspan="11">مبروك علينا وعليكم, لقد احسنت الاختيار. الصحفية القديـــرة ليلى بوزيدي حقا هي متمكنة وذات شخصية مهنـــية متحكمة في إدارة مهنتها أعانها الله ووفقـــها.</td><td>Texte 01</td></tr>
<tr><td colspan="11">1- Après la segmentation et l’élimination des caractères inutiles</td><td rowspan="14"></td></tr>
<tr><td>مبروك</td><td>علينا</td><td>وعليكم</td><td>لقد</td><td>أحسنت</td><td>الاختيار</td><td>الصحفية</td><td>القديـــرة</td><td>ليلى</td><td>بوزيدي</td><td>حقا</td></tr>
<tr><td>هي</td><td>متمكنة</td><td>وذات</td><td>شخصية</td><td>مهـــنية</td><td>متحكمة</td><td>في</td><td>إدارة</td><td>مهنتها</td><td>أعانها</td><td>الله</td></tr>
<tr><td>ووفقـها</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="11">2- Après l’élimination des stopwords</td></tr>
<tr><td>مبروك</td><td>أحسنت</td><td>الاختيار</td><td>الصحفية</td><td>القديـــرة</td><td>ليلى</td><td>حقا</td><td>متمكنة</td><td>شخصية</td><td>مهـنية</td><td>متحكمة</td></tr>
<tr><td>إدارة</td><td>مهنتها</td><td>أعانها</td><td>الله</td><td>وفقـــها</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="11">3- Après la normalisation</td></tr>
<tr><td>مبروك</td><td>احسنت</td><td>الاختيار</td><td>الصحفيه</td><td>القديره</td><td>ليلى</td><td>حقا</td><td>متمكنه</td><td>شخصيه</td><td>مهنيه</td><td>متحكمه</td></tr>
<tr><td>اداره</td><td>مهنتها</td><td>اعانها</td><td>الله</td><td>وفقها</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="11">4- Après le stemming</td></tr>
<tr><td>مبروك</td><td>احسن</td><td>اختيار</td><td>صحف</td><td>قدير</td><td>ليلى</td><td>حق</td><td>تمكن</td><td>شخص</td><td>مهن</td><td>حكم</td></tr>
<tr><td>ادار</td><td>مهن</td><td>اعان</td><td>الله</td><td>وفق</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="11">5- Occurrence pour chaque mot dans le texte</td></tr>
<tr><td>مبروك</td><td>احسن</td><td>اختيار</td><td>صحف</td><td>قدير</td><td>ليلى</td><td>حق</td><td>تمكن</td><td>شخص</td><td>مهن</td><td>حكم</td></tr>
<tr><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>2</td><td>1</td></tr>
<tr><td>ادار</td><td>اعان</td><td>الله</td><td>وفق</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>1</td><td>1</td><td>1</td><td>1</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="11">6- Nombre des termes extraits :</td></tr>
<tr><td></td><td></td><td></td><td colspan="6">16 Mots</td><td></td><td></td></tr>
</table>

Table 5.1 Résultat du prétraitement des textes

5.2. Résultats de l’approche de classification

Nous avons testé notre système sur un corpus de textes et Cette méthode nous a permis de classer presque tous les commentaires présents dans le corpus de test.

Nous avons trouvé qu’un texte peut comporter une ou plusieurs opinions comme il ne peut rien contenir.

Afin de pouvoir comparer les résultats obtenus, nous avons décidé que tous les commentaires que la méthode n'a pas réussi à classer sont des commentaires négatifs. En d'autres termes, tous les commentaires qui ne sont pas positifs sont négatifs.

Dans le but d'évaluer ces résultats, nous calculons les valeurs suivantes: la précision, le rappel et le F-score. Formellement, pour chaque classe Ci, on calcule deux probabilités qui peuvent être estimées à partir de la matrice de contingence correspondante, ainsi ces mesures peuvent être définies de la manière suivante :

- **La Matrice de contingence :**

Pour évaluer un système de classification de ce type, nous utilisons un corpus étiqueté de documents (corpus d'apprentissage) pour lequel on connaît la vraie catégorie de chaque document, et le résultat obtenu par le processus de classification. Pour ce corpus, nous pouvons construire la matrice de contingence pour chaque classe (Voir **Table 5.2**), qui fournit 4 informations essentielles :

- Vrai Positif (**VP**) : Le nombre de documents attribués à une catégorie convenablement. (Documents attribués à leurs vraies catégories)
- Faux Positif (**FP**) : Le nombre de documents attribués à une catégorie inconvenablement. (Documents attribués à des mauvaises catégories)
- Faux Négatif (**FN**) : Le nombre de documents inconvenablement non attribués. (Qui auraient dû être attribués à une catégorie mais qui ne l'ont pas été).
- Vrai Négatif (**VN**) : Le nombre de documents non attribués à une catégorie convenablement (Qui n'ont pas à être attribués à une catégorie, et ne l'ont pas été)

Classe Ci	**Classe positive**	**Classe négative**
Classe positive	$\boldsymbol{VP_i}$	$\boldsymbol{FP_I}$
Classe négative	$\boldsymbol{FN_i}$	$\boldsymbol{VN_i}$

Table 5.2 Matrice de contingence de la classe Ci

- **Le rappel** étant la proportion de documents correctement classés dans par le système par rapport à tous les documents de la classe Ci.

$$Rappel(Ci) = \frac{nombre\ de\ documents\ bien\ classées\ dans\ Ci}{nombre\ de\ documents\ de\ la\ classe\ Ci} \quad (14)$$

$$Rappel = \frac{VP_i}{VP_i + FN_i} \quad (15)$$

Le rappel mesure la capacité d'un système de classification à détecter les documents correctement classés. Un rappel fort ou faible n'est pas suffisant pour évaluer les performances d'un système. Pour cela, on définit la précision.

- **La précision** est la proportion de documents correctement classés parmi ceux classés par le système dans Ci.

$$Précision(Ci) = \frac{nombre\ de\ document\ bien\ classées\ dans\ Ci}{nombre\ totale\ de\ document\ classé\ dans\ Ci} \quad (16)$$

$$Précision = \frac{VP_i}{VP_i + FP_i} \quad (17)$$

Les indicateurs les plus célèbres à savoir le rappel et la précision, sont une estimation courante de la performance d'un système de classification. **F-score : F** permet de combiner, les deux mesures classiques le Rappel (R) et la Précision (P) pour obtenir une moyenne harmonique entre ces deux indicateurs, définit par :

$$F = \frac{2*P*R}{P+R} \quad (18)$$

Les résultats de cette expérimentation sont présentés dans la **Table 5.3**

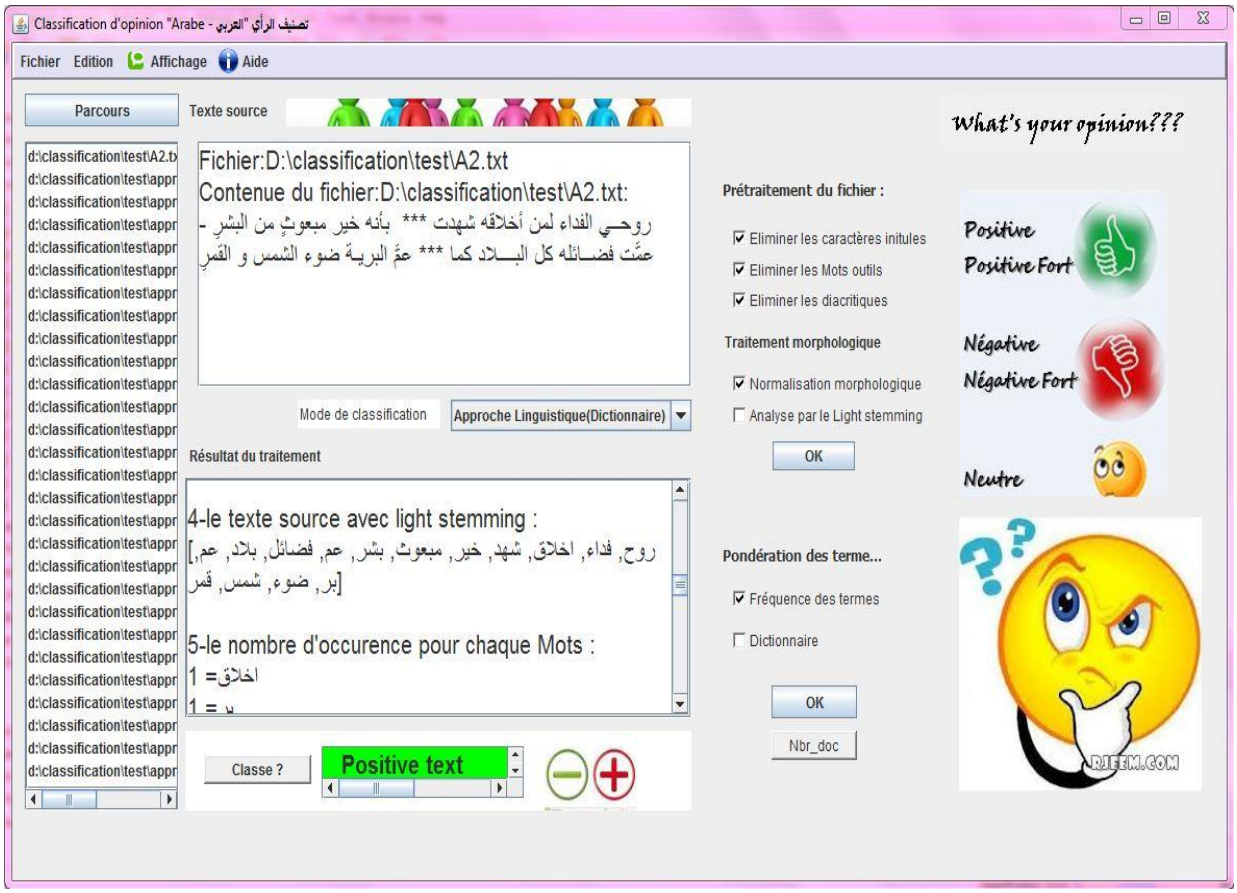

Figure 5.2 Résultat final de processus de classification de textes par l'interface

Classes	Textes de notre corpus	Rappel	Précision	f-score	Temp_ESP	Temp_EAP
Positive + Positive fort	800	75%	77.2%	76.1%	93 seconds	4996 seconds
Négative + Négative fort	800	75.7%	77.9%	76.8%		

Table 5.3 Comparaison des performances de la classification

6. Interprétation

Les taux de performance sont acceptables aux facteurs suivants :

- La richesse et la complexité de traitement de la langue arabes.
- Les mots ambiguës.
- Les erreurs de de l'algorithme de light steming.
- Les dialectes locaux et les erreurs d'orthographe.
- Les autres styles d'expression tels que l'ironie et le sarcasme.
- Pour chaque facteur, un travail de recherche reste à faire pour la langue arabe

Notre approche peut être comme un serveur de base pour les travaux futurs.

7. Conclusion

Au cours de ce chapitre, nous avons présenté les expérimentations d'approche proposée avec toutes ses performances, cette approche de classification est l'approche linguistique qui tient compte le conflit positive et négative et la négation.

CONCLUSION GENERALE

1. Conclusion générale

La classification de textes s'est avérée au cours des dernières années comme un domaine majeur de recherche pour les entreprises comme pour les particuliers. Ce dynamisme est en partie dû à la demande importante des utilisateurs pour cette technologie. Elle devient de plus en plus indispensable dans de nombreuses situations où la quantité de documents textuels électroniques rend impossible tout traitement manuel.

La catégorisation de textes a essentiellement progressé ces dix dernières années grâce à l'introduction des techniques héritées de l'apprentissage automatique qui ont amélioré très significativement les taux de bonne classification. Il reste néanmoins difficile de fournir des valeurs chiffrées sur les performances qu'un système de classification peut actuellement atteindre.

Nous avons entamé notre mémoire en proposant une approche dans le domaine de classification « **Classification automatique des textes Arabe** », ce travail consistait essentiellement à développer une application qui permet la détection de polarité d'opinions en langue arabe.

Cette application propose des méthodes issues du Traitement Automatique du Langage Naturel (TALN), et une méthode linguistique basée sur le dictionnaire pour la classification.

Nous avons décrit dans ce mémoire notre méthode de classification automatique de textes, dont voici les marques principales :

- La transformation ou le codage des documents est la préparation à « l'informatisation » de ces derniers.
- Nous avons expliqué différentes approches de classification d'opinion. Et nous avons choisi l'approche linguistique comme une approche de réalisation de notre système de classification. Cette approche consiste à construire un dictionnaire d'opinion manuellement avec l'aide de techniques simples de traitement automatique des langues. Le dictionnaire de l'approche permet ensuite de classer les textes selon leur polarité, positive ou négative et l'approche reste cependant intéressante car elle nécessite peu d'exemples (textes déjà classés) par rapport aux autres méthodes.
- Les expérimentations effectuées (la construction et le prétraitement du corpus) sur des commentaires issus de blogs d'opinion (twitter) et à partir des journaux arabes et SFU_Review_Corpus_Raw [57].

- Pour pouvoir comparer les résultats obtenus dans les différentes expérimentations, on a utilisé les mesures de performance Rappel, Précision et F-score.

2. Perspectives

Nous citons quelques perspectives de notre travail préliminaire :

- diminuant les dimensions des profils des documents et catégories avec élimination des termes très fréquents (mots outils) et mots très rares sachant que la sélection de descripteurs est un des principaux enjeux du système, puisque du choix des descripteurs et de la connaissance précise de la population va dépendre la mise au point du processus de classification. Ces entrées non discriminantes doivent être supprimées pour deux raisons différentes : réduire le temps de calcul et diminuer le sur-apprentissage.
- Clusterisation de tout le corpus de notre travail, pour avoir un catalogue complet de la base, facilitant son exploitation. Néanmoins le traitement d'un tel espace vectoriel demanderai beaucoup de mémoire et de temps de calcul; d'où la nécessité de recours aux grilles de calcul.
- Utilisation du thésaurus UMLS [58], qui est une base de vocabulaire, comme solution de réduction la taille du vocabulaire.
- Enrichissement du dictionnaire des termes positives et négatives par les méthodes statistiques et apprentissage automatique (cooccurrence, règles d'association et clustering,…) et l'adaptation pour autre domaines (analyse de sentiments, les mauvaises et bon nouvelles,…)
- Adaptation de l'approche pour la détection d'autres tâches plus complexes tel que l'ironie et le sarcasme et les opinions artificielles (spam opinions).
- La spécialisation de dictionnaire par domaine exemple : مسرحية, لعبة ayant la polarité Neutre mais dans le domaine politique est considéré comme un terme négative.
- Le couplage des méthodes lexicale et statistique pour mieux classier l'opinion et pour prendre en compte le contexte (voisinage des mots) pour les ambiguës.
- Elargir la méthode pour tenir compte d'autre styles d'expression tel que spécialisation, généralisation (par exemple l'exception inverse de la polarité…)

Enfin, ce travail était l'occasion de mettre en application l'ensemble des acquis de ma formation de Master « Système d'Information Avancé », et m'a permis d'acquérir de nouvelles compétences sur le plan organisationnel.

Bibliographie

[1], [3] H. Dahmani, « Classification des documents médicaux basée sur le Text Mining » Mémoire de Master, Département de l'informatique, Université de saâd dahlab blida, 2012.

[2] CRISP DM. «Cross Industry Standard Process for Data Mining», 1999.

[4] M. Hearst. «What Is Text Mining? », 2003.

[5] H. Cherfi, «Etude et réalisation d'un système d'extraction de connaissances à partir de textes», 2014.

[6], [7], [8] Chapitre 8, «Fouille de textes », (PDF).

[9] R. Duwairi, « Arabic text Categorization ». Departement of computer informatique systems, University of science and technologie, Jordan, April 2007.

[10] R. Al-Shalabi, G. Kanaan, H. Al-sarhan, « New Approche For Extracting Arabic Roots ». In proceedings of the international arab conference on information technology (ACIT2003), alexandria, Egypt, 2003, pp. 42-59.

[11], [12], [17] H. Matallah, « Classification Automatique de Textes Approche Orientée Agent ». Mémoire de magister, Département de l'informatique, université d'Aboubekr Belkaid-Tlemcen, Février 2011.

[13] K. Abidi, **«**La Categorisation De Texte Multilingue **»,** Mémoire De Magistère D'informatique, Option : SIC, Ecole National Supérieur d'Informatique, 2010/2011.

[14] S. Jaillet, « Catégorisation Automatique De Documents » LIRMM UMR 5506, 161 rue Ada, 34392 Montpellier Cedex 5 – France URL, 2004. L'URL :
http://www.lirmm.fr/doctiss04/art/I02.pdf

[15] C. Ignat, « Représentation De Textes A L'aide D'étiquettes Sémantiques Dans Le Cadre De La Classification Automatique », European Commission, IPSC, Joint Research Centre – 21020 Ispra (VA) - Italie LICIA, LGECO – INSA -67084 Strasbourg CEDEX – France URL, 2007. L'URL :
http://langtech.jrc.it/JRC_Publications.html

[16], [18] H. Dahmani, « Classification des documents médicaux basée sur le Text Mining » Mémoire de Master, Département de l'informatique, Université de saâd dahlab blida, 2012.

[19] R. Jalam, « Apprentissage Automatique Et Catégorisation De Textes Multilingues », 2003.

[20], [31] K. Abidi, **«**La Catégorisation De Texte Multilingue **».** Mémoire De Magistère D'informatique, Option : SIC, Ecole National Supérieur d'Informatique, 2010/2011.

[21] B. Ameni, « Catégorisation automatique de news à l'aide de techniques d'apprentissage supervisé », Rapport de Projet de Fin d'Etudes, Université Nice Sophia Antipolis.

[22] http://www.cuy.be/html/typoweb/chap1.htm

[23] H. Matallah, « Classification Automatique de Textes Approche Orientée Agent », Mémoire de magister, Département de l'informatique, université d'Aboubekr Belkaid-Tlemcen, Février 2011.

[24] R. Ayadi, W. Jaoudi, « La distance intertextuelle pour la classification de textes en langue Arabe », Tunisie, 2009.

[25] F. Sebastiani, « Machine learning in automated text categorization », 2002.

[26] M.K. Saad, W. Ashour, « Arabic Morphological Tools for Text Mining ». Faculty of Information Technologie and computer Engineering, Islamic University of Gaza, Palastine, 2010.

[27] J. Clech, D.A. Zighed, « Une technique de réétiquetage dans un contexte de catégorisation de textes », 2014.

[28] S. Morinaga, K. Yamanishi, K. Tateishi, et T. Fukushima, « Mining Product Reputations On The Web », In KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, 2002, pp. 341–349. ACM.

[29] P.D. Turney, « Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews ». In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 417–424.

[30] Yu, H. et V. Hatzi-vassiloglou, « Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences ». In Proceedings of the 2003 conference on Empirical methods in natural language processing, Morristown, NJ, USA, 2003, pp. 129–136. Association for Computational Linguistics.

[32] http://www.globalwordnet.org/AWN/

[33] La base sous format *XML* et une autre *MySQL* disponible dans le lien suivant : http://www.globalwordnet.org/AWN/DataSpec.html

[34] Pang. B, L. Lee, et S. Vaithyanathan, « Thumbs up ?: sentiment classification using machine learning techniques ». In EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Morristown, NJ, USA, 2002, pp. 79–86. Association for Computational Linguistics.

[35] M. Abdul-Mageed and M. Diab. AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), 2012.

[36] M. Abdul-Mageed and M. T. Diab. Subjectivity and sentiment annotation of modern standard arabic newswire. In Proceedings of the 5th Linguistic Annotation Workshop, LAW V '11, pages 110–118, 2011.

[37] M. Abdul-Mageed and M. Diab. Toward building a large-scale Arabic sentiment lexicon. In Proceedings of the 6th International Global Word-Net Conference, Matsue, Japan, 2012.

[38] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06, volume 6, pages 417–422. Citeseer, 2006.

[39] M. Abdul-Mageed, M. Korayem, and A. YoussefAgha. "Yes we can?": Subjectivity annotation and tagging for the health domain. In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP, Hissar, Bulgaria, 2011.

[40] M. Rushdi-Saleh, M. Mart´ın-Valdivia, L. Ure˜na-L´opez, and J. Perea-Ortega. Oca: Opinion corpus for Arabic. Journal of the American Society for Information Science and Technology, 62(10):2045–2054, 2011.

[41] M. Elarnaoty, S. AbdelRahman, and A. Fahmy. A Machine Learning Approach For Opinion Holder Extraction Arabic Language. CoRR, abs/1206.1011, 2012.

[42] S. AbdelRahman, M. Elarnaoty, M. Magdy, and A. Fahmy. Integrated machine learning techniques for arabic named entity recognition. International Journal of Computer Science Issues IJCSI, 7(4):27–36, 2010.

[43] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 347–354. Association for Computational Linguistics, 2005.

[44] M. Elhawary and M. Elfeky. Mining Arabic Business Reviews. In Proceedings of International Conference on Data Mining Workshops (ICDMW), pages 1108–1113. IEEE, 2010.

[45] Z. Agi´c, N. Ljube˘si´c, and M. Tadi´c. Towards sentiment analysis of financial texts in croatian. In Proceedings of the Seventh International Conference on Language Resources and

Evaluation (LREC'10), 2010.

[46] K. Ahmad, D. Cheng, and Y. Almas. Multi-lingual sentiment analysis of financial news streams. In Proceedings of the 1st International Conference on Grid in Finance, 2006.

[47] Y. Almas and K. Ahmad. A note on extracting sentiments in financial news in English, Arabic & Urdu. In Proceedings of Workshop on Computational Approaches to Arabic Script-based Languages, 2007.

[48] N. Farra, E. Challita, R. Assi, and H. Hajj. Sentence-Level and Document-Level Sentiment Mining for Arabic Texts. In Proceedings of International Conference on Data Mining Workshops (ICDMW), pages 1114–1119. IEEE, 2010.

[49] M. Abdul-Mageed, M. Diab, and M. Korayem. Subjectivity and sentiment analysis of modern standard Arabic. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, pages 587–591. Association for Computational Linguistics, 2011.

[50] M. Abdul-Mageed and M. Korayem. Automatic Identification of Subjectivity in Morphologically Rich Languages: The Case of Arabic. In Proceedings of the 1st workshop on computational approaches to subjectivity and sentiment analysis (WASSA), pages 2–6, 2010.

[51] D. Poirier, F. Fessant, C. Bothorel Émilie Guimier de Neef, M. Boullé, «Approches Statistique et Linguistique Pour la Classification de Textes d'Opinion Portant sur les Films », France Telecom R&D, TECH / EASY, 2 avenue Pierre Marzin, 22300 Lannion, France.

[52] Sentiment Lexicons, OpinionFinder: 2006 positive words, 4783 negative words sur l'URL :
http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[53] Liste de 879 sentiments répartis en 10 catégories émotionnelles, Jean-Philippe Faure – décembre 2006.

[54] Encarta Dictionnaire, 2009.

[55], [57] http://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html

[56] A. Ziani, Y. Tlili Guiassa, N. Azizi, « Détection de polarité d'opinion dans les forums en langue arabe par fusion de plusieurs SVMs », Département de l'informatique, université d'Annaba (Algérie), 2013.

[58] http://www.nlm.nih.gov/research/umls/

ملخص

العمل المنجز في اطار هذه المذكرة يتعلق بتصميم برنامج يسمح بمتابعة عملية المعالجة الالية و التصنيف الالي للنصوص العربية بناءا على انواع معروفة مسبقا وهي الآراء والتعليقات الشخصية حول موضوع معين في مختلف المجالات سياسة, اقتصاد, ثقافة, فن , .. و ما شابه ذلك, وهذا من خلال استخدام الطريقة التي تعتمد على قاموس الكلمات السلبية والايجابية ومعالجة حالات التضاد السلبي والايجابي وحالات النفي. التطبيق انجز بلغة java.

الكلمات المفتاحية : تنقيب النصوص, التصنيف الالي, النصوص العربية, تنقيب الرأي,

Résumé

Le travail effectué dans le cadre de ce mémoire s'intéresse à la réalisation d'un système permet de suivre pas à pas le processus de prétraitement et de classification automatique des textes arabes sur la base des types déjà connus à l'avance qui spécifique les opinions et les commentaires personnels sur un sujet particulier dans différents domaines, politique, économie, culture, art, …etc, et ce grâce à l'utilisation de la méthode qui base sur un dictionnaire des mots ayant une tonalité positive ou négative et traitement des cas de conflit entre les mots négative et positive et le problème de négation. L'application effectue par le langage java.

Mots clés : Fouille de textes, Classification automatique, Textes arabes, Fouille d'opinion,

Abstract

The work carried out within the framework of this memory is interested in the realization of a system makes it possible to follow step by step the process of pretreatment and automatic classification of the texts Arabs on the basis of already known type in advance which specific personal opinions and comments on a particular subject in various fields, policy, economy, culture, art... etc, and this by to the use of the method which based on a dictionary of the words having a positive or negative tonality and treatment of the cases of conflict between the words negative and positive and the problem of negation. The application carries out by the language java.

Keywords : Texts mining, Automatic classification, Arabic texts, Opinion mining.