الجمهورية الجزائرية الديمقراطية الشعبية

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE YAHIA FARES DE MEDEA

FACULTE DES SCIENCES

Département Mathématique et Informatique



Projet de fin d'études

Présenté pour l'obtention du Diplôme de

Master en Informatique

Spécialité Ingénierie des Systèmes et Technologie du Web

THEME

Analyse d'opinion contextuelle dans les réseaux sociaux pour la langue Arabe

Proposé et dirigé par :

Mr. Mazari A.C

et

Melle Ould Kebaili H

Préparé et présenté par :

Melle Metidji Battoul et Melle BenHelal Asmaa

Année Universitaire 2016-2017

Remerciement

Avant toutes choses, nous remercions Dieu le tout puissant qui nous a donné la volonté et la patience pour aboutir à la réalisation de ce travail.

Nous souhaiterons manifester nos reconnaissances particulièrement à MR Mazari A.c, Ould Kebaili Hadjer, Khelili Lotfi et tous les enseignants de université de Médéa qui nous ont aidés.

Nous exprimons nos profonds remercîments aux membres de jury pour leurs participations dans le jugement de notre travail.

Nos sincères remercîments vont également à l'ensemble des enseignants du département de Réseaux et systèmes informatique pour leur encadrement universitaire. Et leurs aides considérables et leurs en leurs encouragements.

Nous remercions aussi nos amis pour leur aide et leurs encouragements.

Que tous ceux qui nous ont aidés de près ou de loin, trouvent ici l'expression de notre gratitude.

Médéa, juin 2017

Battoul, Asmaa



Dédicace

J'ai l'honneur de dédier ce modeste travail à mes chers parents qui ont enduré tous les peines de la vie afin que je puisse continuer dans la voie du savoir et de l'éducation, à mes sœurs Imene, Amira, Meriem et mon fiancé Lotfi qui m'ont tellement aidé et encouragé,

A ma chère partenaire : Asmaa

A mes amis.

A tous mes amis et toute personne qui de près ou de loin nous a aidé à réaliser notre projet et a participé à notre formation.

Aux membres de ma promotion.

Qu'ils trouvent ici mes sincères salutations

Battoul

Dédicace

J'ai l'honneur de dédier ce modeste travail à mes chers parents qui ont enduré tous les peines de la vie afin que je puisse continuer dans la voie du savoir et de l'éducation, à mes sœurs Chaimâa qui m'a tellement aidé, Amina, Bouchra et mon frère Abd El Aziz

A ma chère partenaire : Battoul

A mes amis.

A tous mes amis et toute personne qui de près ou de loin nous a aidé à réaliser notre projet et a participé à notre formation.

Aux membres de ma promotion.

Qu'ils trouvent ici mes sincères salutations

Asmâa

Résumé

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

Dans ce présent mémoire, nous avons proposé une approche de classification d'opinions basée sur une analyse symbolique des textes, qu'ont testés sur un ensemble de tweets de la langue arabe. Nous avons inspiré notre approche, après une étude comparative de trois approches utilisées en classification de sentiments : l'approche symbolique basée sur l'analyse syntaxique des textes, l'approche statistique basée les techniques d'apprentissage automatique, et la troisième approche, est une hybridation des deux premières. Son fonctionnement est basé sur l'extraction des expressions subjectives qui reflètent des jugements personnels sur des sujets divers.

Notre approche, se base sur un modèle de représentation d'opinion, qui considère qu'une opinion est conformée d'un ensemble d'élément : *Feature*. L'identification de ce élément nécessite un ensemble des méthodes bien définies et une ressource externe. A l'aide de ce dernier nous avons ajouté à ce modèle deux éléments « *Intensité* » et « *polarité* » pour calculer l'orientation sémantique globale de l'opinion.

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

Mots Clés : Langue Arabe, TALN, Opinion, Analyse, Objectivité, Subjectivité, Classification

ملخص

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

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

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

اعتمدنا في طريقتنا على نموذج لتمثيل الرأي و يعتبر هذا النموذج أن الرأي يتكون من مجموعة من عنصر و هو: الميزة ، يتم تحديد هذا عنصر وفقا لمجموعة من القواعد اللغوية مبنية على أسس صحيحة و على مصدر خارجي، و اعتمادا على هذا الأخير تم إضافة عنصرين لهذا النموذج و هما: الشدة و القطبية لحساب التوجه المعنوي الكلي للرأي وفقا للعناصر المشكلة له..

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

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

Abstract

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

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

Our approach, bases on a model of representation of opinion, which considers that an opinion is formed of group of element: *Feature*. The identification of this element requires a set of linguistic rules well defined and extern resource. Based on this resource we added two elements called *intensity* and *polarity* to this model to calculate the total semantic orientation of the opinion according to these components.

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

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

Table de Matières

Introduction générale 01
Problématique et objectif 02
Organisation du mémoire..... 03

Chapitre -I- : Etat de l'art

1. Introduction..... 04

PARTIE 01

2. Fouille d'opinion 05
2.1. Définition 05
2.1.1. Définition formelle d'une opinion 05
2.1.2. Définition de la fouille d'opinion 05
2.2. Source de données 06
2.2.1. Web 2.0 06
2.2.2. Réseaux sociaux..... 06
2.3. Différent types d'opinion..... 07
2.3.1. Opinions comparatives et réguliers..... 07
2.3.2. Opinions explicites et implicites 08
2.4. Subjectivité/ Emotion 08
2.4.1. Subjectivité/objectivité 08
2.4.2. Emotion..... 09
2.5.Traits/ Features 10
2.5.1. Présence 10
2.5.2. Fréquence 10
2.5.3. Position 10
2.6. Echelle et niveaux d'analyse 10
2.7. Outils d'analyse 12
3. Classification d'opinion 12
3.1. Définition 12
3.2. Polarité de l'opinion 13
3.3. Orientation sémantique 13
4. Difficultés de la fouille d'opinion..... 13

PARTIE 02

5. TALN 15
5.1. Les niveaux d'analyse de TALN 15
5.2. Difficulté de TALN : ambiguïté et implicite 17

PARTIE 03

6. La langue arabe..... 19

6.1. Composition du lexique arabe : les formes de bases 22

6.1.1. Les verbes 22

6.1.2. Les noms 23

6.1.3. Les pronoms..... 24

6.1.4. Les mots outils 25

6.2. Difficultés de traitement automatique de la langue arabe 26

6.2.1. Difficultés de l'analyse morphologique de l'arabe 27

7. Conclusion..... 29

Chapitre -II- : Les travaux Connexes

1. Introduction 30

2. Travaux de recherche de fouille d'opinion 31

3. Techniqques de classification de textes 33

3.1. Classification supervisée 34

3.1.1. Algorithme de classification supervisée 34

3.1.1.1. K plus proches voisins et ses ameliorations 34

3.1.1.2. Arbres de décisions 35

3.1.1.3. Naive bayes(ou simple bayes) 35

3.1.1.4. Réseaux de neurones 36

3.1.1.5. Machines a support de vecteurs (ou SVM) 36

3.1.1.6. Programmation génétique 37

3.2.Classification non supervisée 37

3.2.1. Quelques algorithmes de classification supervisée 37

4. Classification d'opinion 40

4.1. Travaux de recherches 40

4.2. Méthodes de classification 42

4.2.1. Méthode symbolique 42

4.2.2. Méthode statistiques 43

4.2.3. Méthode hybride 43

5. Trvauxx connexes sur la langue arabe 43

6. Conclusion 44

Chapitre -III- : L'approche proposée

1. Introduction 45

2. Description 46

3. Notre systéme de classifier une opinion 46

3.1. Collecte des opinions 47

3.1.1. Normalisation des textes 47

3.1.2. Encodage uniques des textes 47

3.2. Architecture générale 47
3.2.1. Prétraitement des données 48
3.2.1.1.Suppression de la répétition 49
3.2.1.2.Suppression des expressions supplementaire 49
4. Traitement des données 50
4.1. Lemmatisation (Radicalisation) 51
4.2. Recherche des Features 51
4.3. Extraire de polarité 51
4.4. Classification d'opinion 53
5. Conclusion 54

Chapitre -IV- : Implémentation et Evaluation

1. Introduction 55
2. Constitution de la abse de données 55
3. L'environnement de développement 55
3.1. Python..... 56
3.1.1. Flask 56
3.1.2. Les modules utilisés 56
3.2. HTML5..... 58
4. Presentation de l'application 58
4.1. Déroulement 61
5. Classification et évaluation 65
5.1. RapidMiner..... 65
5.2. Classification 65
5.3. Evaluation..... 69
6. Conclusion 70
Conclusion et perspectives 71
Annexe 1..... 72
Annexe 2..... 76
Bibliographie 77

Tables des figures

FIGURE 1.1 : *les niveaux d'analyse de TALN exemple anglais* 17
FIGURE 1.2 : *Exemple de dérivation de la racine* بتك *ktb* 20
FIGURE 1.3 : *Structure d'un mot arabe* 21
FIGURE 1.4 : *Modèle du mot graphique en arabe appliqué pour le mot «* 21 ..... *»* هنونرعنا
FIGURE 3.1 : *Architecture générale de notre système* 48
FIGURE 3.2 : *Processus de prétraitement proposé* 49
FIGURE 3.3 : *Processus de traitement proposé* 50
FIGURE 3.4 : *Résultat d'utilisation de n-gramme* 51
FIGURE 3.5 : *présentation des concepts de ressource Senticnet* 52
FIGURE 3.6 : *présentation de traitement de ressource* 52
FIGURE 3.7 : *présentation de lemmatisation de ressource* 53
FIGURE 4.1 : *Diagramme de classe* 60
FIGURE 4.2 : *Fenêtre d'accueil de notre site web* 61
FIGURE 4.3 : *Fenêtre de SEARCH* 61
FIGURE 4.4 : *Ecrire une opinion* 62
FIGURE 4.5 : *Résultat de SAERCH* 62
FIGURE 4.6 : *Fenêtre de CORPUS* 62
FIGURE 4.7 : *Résultats de CORPUS* 63
FIGURE 4.8 : *Top 20 opinions positives* 64
FIGURE 4.9 : *Top 20 opinions négatives* 64
FIGURE 4.10 : *design d'application de Naïve bayes* 66
FIGURE 4.11 : *Validation de modèle* 66
FIGURE 4.12 : *Résultats de classification selon Senticnet 1* 67
FIGURE 4.13 : *Résultats de classification selon Senticnet 2.1* 67
FIGURE 4.14 : *Résultats de classification selon Senticnet 2.2* 68
FIGURE 4.15 : *Résultats générale de classification* 68
FIGURE 4.16 : *Évaluation de différentes ressources pour la classification d'opinion* 69
FIGURE : *Hourglass de SenticNet* 75

Tables des tableaux

TABLEAU 1.1 : *Variation de la lettre* 19 .....ع
TABLEAU 1.2 : *Tableau de voyelles en arabe*.....19
TABLEAU 1.3 : *Interprétation du mot* بنك *sans voyelles*..... 20
TABLEAU 1.4 : *Schèmes de dérivés du mot* 20 .....لمح
TABLEAU 1.5 : *Principaux schèmes verbaux augmentés* 23
TABLEAU 4.1 : *Résultats générale de classification d'opinion* 69

Introduction général

La fouille d'opinion ou l'analyse des sentiments attirent de nos jours l'attention de plusieurs chercheurs dans le domaine de Traitement Automatique des Langages Naturels (TALN). L'importance de cet axe de recherche réside surtout dans le traitement de pages web dont le contenu est généré par l'utilisateur (*User-generated content Web*). Ce type d'information, tels que les réseaux sociaux, les blogs et les forums de discussion, a connu ces dernières années une évolution énorme et est devenu une composante influente dans la vie sociale, politique et même économique ; Ces ressources sont considérées comme référence pour les sociétés qui veulent savoir les opinions de leurs clients sur un produit ou un service, et les sondeurs qui veulent savoir les opinions publiques sur un sujet politique ou autres,…etc.

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

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

L'analyse d'opinion est considérée parmi les tâches les plus importantes dans le domaine de fouille d'opinion. Cette tâche permet d'attribuer à un énoncé subjectif une polarité positive ou négative. Ceci est très utile dans la mesure où il permet de fouiller les opinions des gens à propos d'un sujet donné (i.e. une organisation, un produit, un personnage, etc), et agir et prendre des décisions selon les résultats.

Dans ce mémoire, nous proposons une méthode de classification d'opinions à partir des textes des réseaux sociaux Arabe (par exemple Twitter) en analysant des polarité des traits (Features) par l'analyse du lexique, le contexte et d'appliquer les techniques d'apprentissage automatique, une partie d'expérimentation est indispensable pour mesurer les performances de l'approche proposée.

[1] Ceux qui utilisent les opinions d'autres individus ou d'autres organismes, par exemple pour améliorer la qualité d'un service ou d'un produit.

Problématique et objectifs

Ce projet présente un domaine : la fouille de données d'opinion (ang. Opinion-Mining)), dont l'objectif est de concevoir et de développer un système capable d'analyser et classifier les opinions dans les textes et plus précisément dans les textes arabes de différentes sources : forums de discussion, réseaux sociaux, sites de e-commerce, blogs, journaux, etc.,.

L'Opinion-Mining (OM) ou l'analyse des sentiments (ang. Sentiment Analysis), également appelé analyse de subjectivité (ang. Subjectivity Analysis), est une discipline en plein essor, réunit plusieurs disciplines : Traitement automatique des langages naturels (TALN), recherche d'information (RI), et la linguistique, dont l'objectif principal est d'extraire des opinions à partir des textes et de classifier ces opinions selon leurs orientations sémantiques en deux catégories principales : positive et négative.

La tâche d'analyse d'opinions est un ensemble de techniques qui fait partie du domaine du traitement automatique du langage naturel et plus précisément de la recherche de l'information, qui consiste à extraire les opinions disponibles sur des masses textuelles importantes, et les classifier en catégories. La classification d'opinions est un axe d'actualité qui cherche à déterminer les avis ou les critiques des individus et des organismes à propos d'un sujet ou d'un objet. Et elle nécessite une analyse lexicale et syntaxique profonde, surtout dans notre cas, où la langue traitée est l'arabe, cette dernière, qui se caractérise par une morphologie complexe, présente l'un des grands challenges que nous devons faire face.

La quantité énorme des informations disponibles notamment sur le Web rend aujourd'hui les systèmes actuels chaque jour plus préoccupés par la manière de gérer la surcharge d'information, en s'assurant que l'utilisateur aura l'accès aux meilleures sources avec le moindre effort. Ces dernières années, une attention particulière a été donnée pour résoudre ce problème. Le secteur de e-commerce est l'un des plus influencé par la quantité de données produites par les internautes sous forme de commentaires et critiques sur des produits commercialisés ,ou des sujets politiques cette augmentation a été remarquée surtout après l'apparence du Web 2.0.

Les opinions des internautes représentent une source d'information qui ne devrait pas être maltraitée ou ignorée par la communauté de recherche. Ce travail souligne la nécessité de développer une approche capable de profiter pleinement de ces données, dont l'objectif est d'exploiter les sources d'opinions en vue de faciliter la prise de décision plus rapidement avec le moindre coût.

Dans cette thèse, nous présentons une approche d'analyse d'opinion, et le classifier selon un domaine quelconque.. Pour analyser les expressions d'opinion, nous utilisons un lexique de sentiments externe, où chaque mot de ce lexique peut être un adjectif, un verbe ou un adverbe, et pour chaque mot ou expression de ce lexique, une polarité positive ou négative est associée, avec des taux exprimant le degré de positivité et négativité.

Organisation du mémoire

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

Selon la nature de notre projet, nous avons organisé ce présent mémoire en 4 chapitres comme suit :

Le premier chapitre est sur l'état de l'art, dans lequel on explique en trois parties les différents concepts de fouille d'opinions et sa difficultés, nous parlons d'une manier brève dans traitement automatique de la langue et la langue arabe et leur difficultés.

Dans le deuxième chapitre, nous expliquons les différentes techniques d'analyse d'opinion et ces différentes approches faites.

Le troisième chapitre est dédié pour expliquer notre approche d'analyse des sentiments d'opinions, en expliquant les techniques utilisées pour prétraiter les textes arabes, ainsi que l'approche utilisée pour classifier toutes les opinions extraites.

Dans ces chapitres nous avons abordés, entre autres, les points suivants :

1. Comment collecter les opinions ?
2. Quels sont les différentes étapes pour l'analyse d'opinion ?
3. Comment déterminer l'orientation sémantique d'une opinion ?
4. L'approche utilisée pour classifier les opinions analysées.

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

CHAPITRE I :
Etat de l’art

1. Introduction :

Avec l'explosion d'Internet et des réseaux sociaux est apparue la nécessité d'analyser des millions de postes, de tweets ou d'avis afin de savoir ce que pensent les internautes.

L'Opinion Mining permet d'analyser automatiquement toutes ces données textuelles et d'en faire ressortir les différentes opinions exprimées sur un sujet précis comme une marque, une actualité ou un produit.

Ces analyses sont précieuses pour les entreprises, notamment, qui peuvent ainsi savoir comment leur image est perçue et si leurs produits plaisent ou au contraire font un flop. C'est pourquoi de nombreuses applications d'Opinion Mining fleurissent depuis les années 2000.

Elle est en passe de devenir une véritable industrie, tout aussi stratégique que celle des sondages. Les promesses avancées sont impressionnantes : la puissance de calcul des outils informatiques permettrait de suivre toutes les évolutions de l'opinion sur le web en temps réel, quel qu'en soit le volume. Plus encore, les capacités de traitement linguistique permettraient de détecter les tonalités de tous les verbatims recueillis, grâce aux méthodes dites de « sentiment analysis ».

Mais comme dans toutes applications qui traitent du texte, la difficulté majeure est d'arriver à interpréter le langage humain qui reste, encore aujourd'hui, un objet complexe, tant par sa richesse que par ses subtilités. Il est très difficile par exemple de détecter de façon informatique.

L'objectif de ce chapitre c'est présenté le domaine de fouille d'opinion ou la première partie, nous allons présenter les différents termes de ce domaine en particulier nous décrivons les notions de fouille d'opinion et ses difficultés, la deuxième partie nous avons traitement automatique de la langue et sa difficultés et pour la troisième partie et la dernière présentation de la langue arabe, ses caractéristiques et ses obstacles.

PARTIE 01

2. Fouille d'opinion (ang. Opinion Mining) :

L'internet social a récemment fait exploser la disponibilité de documents textuels exprimant des opinions ou des sentiments, par exemple dans les groupes de discussions, les blogs, forums et autres sites spécialisés dans les critiques de produits. Les opinions disponibles sur l'internet ont un impact considérable sur les internautes [Gillot ,2010].

2.1. Définition:

2.1.1. Définition formelle d'opinion :

L'opinion (aussi appelle avis) peut être définie comme l'expression des sentiments d'une personne envers une entité, étant subjective et opposée à l'expression de faits [Gillot ,2010].

Une opinion est composée de 3 composants basic [J. Wiebe et al., 2005]:

- Détenteur de l'opinion (ang. opinion holder) : Le détenteur d'un avis est la personne ou l'organisation qui exprime l'opinion. Les détenteurs d'opinion sont également appelés sources d'opinion.
- Objet (ang. object): Un objet o est une entité qui peut être un produit, une personne, un événement, une organisation ou un sujet. sur lequel l'opinion a été exprimée.
- Opinion : un point de vue de détenteur de l'opinion sur un objet, un avis positif ou négatif, une attitude, une émotion ou une appréciation.

2.1.2. Définition de fouille d'opinion :

Le terme « fouille d'opinions » est utilisé pour évoquer le traitement automatique des opinions, des sentiments et de la subjectivité dans les textes, ce domaine est connu sous les noms d'opinion mining, sentiment analysis [Gillot ,2010].

Fouille d'opinions (ang. Opinion Mining), aussi appelée l'analyse automatique des opinions, concerne l'extraction d'un sentiment dans une source telle qu'un texte sans structure prédéfinie.

Les sentiments reconnus peuvent être classés soit positifs soit négatifs, soit en des classes définies plus finement.

L'objectif de la fouille d'opinion est d'analyser une grande quantité de données afin d'en déduire les différents sentiments qui y sont exprimés. Les sentiments extraits peuvent ensuite faire l'objet de statistiques sur le ressenti général d'une communauté.

Elle se réduit à distinguer les opinions positives, les opinions négatives, les tonalités, et éventuellement à classer dans une catégorie neutre les opinions

indécidables (parce qu'elles sont mixtes par exemple, ou qu'il n'y a pas d'opinion exprimée dans le document, ou parfois encore parce que la méthode ne peut trancher clairement dans l'ensemble du document) [Boullier et Lohard, 2012].

2.2. Source de données :

2.2.1. Web 2.0 :

La popularité des réseaux sociaux comme Twitter ou Facebook vient de commencer à augmenter au cours des dernières années. Une des principales raisons de son succès est qu'une interface simple est présentée aux utilisateurs, afin qu'ils puissent partager leurs sentiments, leurs idées et leurs expériences de manière rapide et simple. En outre, avec les dernières avancées dans les technologies mobiles, les utilisateurs peuvent réellement interagir à tout moment de n'importe où; L'information en temps réel est devenue une réalité. Ce processus collaboratif et interactif entre utilisateurs, visant à partager l'information, a reçu le nom de Web 2.0.

L'expression « Web 2.0 » désigne l'ensemble des techniques, des fonctionnalités et des usages qui ont suivi la forme originelle du web, www ou (World Wide Web), caractérisée par plus de simplicité et d'interactivité (sociabilité). Elle concerne en particulier les interfaces et les échanges permettant aux internautes ayant peu de connaissances techniques de s'approprier de nouvelles fonctionnalités du web. Les internautes peuvent d'une part contribuer à l'échange d'informations et interagir (partager, échanger, etc.) de façon simple, à la fois au niveau du contenu et de la structure des pages, et d'autre part entre eux, créant notamment le Web social. L'internaute devient, grâce aux outils mis à sa disposition, une personne active sur la toile.

L'expression « Web 2.0 » utilisée en 2003 par Dale Dougherty, un cadre de la société O'Reilly Media, et diffusée par Tim O'Reilly en 2004 puis consolidée en 2005 avec l'exposé de position « What Is Web 2.0 » [Site1, 2009].

Donc le Web 2.0 est l'évolution du Web vers l'interactivité à travers une complexification interne de la technologie mais permettant plus de simplicité d'utilisation, les connaissances techniques et informatiques n'étant pas indispensables pour les utilisateurs.

2.2.2. Réseaux sociaux :

Le terme de réseaux sociaux désigne généralement l'ensemble des sites internet permettant de se constituer un réseau d'amis ou de connaissances professionnelles et fournissant à leurs membres des outils et interfaces d'interactions, de présentation et de communication.

Réseaux sociaux (ang. Social media) Mode d'interactions sociales qui facilite la création et l'échange d'information et de contenus entre des individus et entre individus et organisations. Désigne aussi les plateformes qui rendent ces interactions possibles (telles que Facebook, Twitter, etc.).

Les réseaux sociaux les plus connus sont Facebook, Twitter, LinkedIn, Viadeo, Pinterest, etc. YouTube peut également être considéré partiellement comme un réseau social dans la mesure où le service a développé des outils d'interactions entre ses membres.

2.3. Diffèrent types d'opinion :

2. 3.1. Des opinions comparatives et régulières :

Les types d'opinion que nous allons discuter sont le type régulier et le type comparatif.
Nous pouvons aussi classer les opinions selon la façon de leur exprimé dans le texte, opinion explicite et opinion implicite [Liu, 2012].

A. Une opinion régulière :

Un avis régulier est souvent considéré comme une opinion dans la littérature. Il exprime un sentiment seulement sur une entité particulière ou un aspect de l'entité, par exemple "Coke tastes very good," qui exprime un sentiment positif sur le goût aspect de Coke [Liu, 2012].

Il a deux sous-types principaux :

- ✓ **Opinion directe :**

Un avis direct renvoie à un avis exprimé directement Sur une entité ou un aspect d'entité, par exemple, « The picture quality is great »,

- ✓ **Opinion indirecte :**

Un avis indirect est une opinion exprimée indirectement sur une entité ou un aspect d'une entité en fonction de ses certaines autres entités. Ce sous-type se produit souvent dans le domaine médical. Par exemple, la phrase « After injection of the drug, my joints felt worse « décrit un effet indésirable de la drogue sur« my joints », Ce qui donne indirectement une opinion négative ou un sentiment à la drogue. Dans le cas, l'entité est le médicament et l'aspect est l'effet sur les articulations (my joints).

Une grande partie de la recherche actuelle se concentre sur les opinions directes. Car Ils sont plus simples a traité par rapport à les opinions indirectes qu'elles sont souvent plus difficiles à traiter. Par exemple, dans domaine des drogues, il faut savoir si dès l'état indésirable est avant ou après l'utilisation du médicament. Par exemple, la phrase « Since my joints were painful, my doctor put me on this drug » n'exprime pas un sentiment ou une opinion sur

la drogue parce que «painful joints» (qui est négatif) est arrivé avant d'utiliser le médicament.

B. Une opinion comparative :

Opinion comparative exprime une relation de Ressemblances ou différences entre deux ou plusieurs entités et / ou une Préférences du détenteur de l'opinion sur la base de certains aspects [Jindal et Liu, 2006]. Par exemple, les Phrases, « Coke tastes better than Pepsi » et « Coke tastes the best » deux opinions comparatives. Une opinion comparative est habituellement exprimée en utilisant la forme comparative ou superlative d'un adjectif ou adverbe, bien que pas toujours (par exemple, préfèrent). Opinions comparatives également ont de nombreux types.

2.3.2. Des opinions explicites et implicites :

Une opinion explicite :

Une opinion explicite est une affirmation subjective qui donne une opinion régulière ou comparative, par exemple, «Coke tastes great », et «Coke tastes better than Pepsi. ».

Une opinion implicite (implicite) :

Une opinion implicite est une déclaration objective qui implique un avis régulier ou comparatif.

Une telle objective expression exprime habituellement un fait souhaitable ou indésirable, par exemple, « I bought the mattress a week ago, and a valley has formed », et « The battery life of Nokia phones is longer than Samsung phones. » Les opinions explicites sont plus faciles à détecter et à classer que les opinions implicites. Une grande partie de la recherche actuelle s'est concentrée sur des opinions explicites que les opinions implicites.

2.4. Subjectivité/ Emotion:

Il existe deux concepts importants qui sont profondément liés au sentiment et opinion, c'est-à-dire subjectivité et émotion.

2.4.1. Subjectivité / Objectivité :

Ces deux notions ne sont pas équivalentes, bien qu'ils aient une grande intersection. La tâche de déterminer si une phrase est subjective ou objective est appelée classification de subjectivité, nous devrions les noter comme suit:

A. Subjectivité :

Pour Jan Wiebe, la subjectivité est « the linguistic expression of somebody's opinions, sentiments, emotions, evaluations, beliefs, speculation (private states)» [J. Wiebe et al., 1999] . Elle reprend ensuite la définition du « private

state » de Quirck, Greenbaum, Leech et Svartvic : un « private state is a state that is not open to objective observation or verification ».

Les expressions subjectives sont nombreuses les opinions, les allégations, les désirs, les croyances, les soupçons et les spéculations ; Un exemple d'une phrase subjective est : "I like iPhone."

Elles expriment certains sentiments, vues ou croyances personnelles.

Mais parfois une phrase subjective ne peut exprimer aucun sentiment. Par exemple, "I think that he went home" est une phrase subjective, mais n'exprime pas tout un sentiment positif ou négatif à propos de quoi que ce soit.

B. Objectivité :

Une phrase objective présente une information réelle sur le monde ; Un exemple d'une phrase objective est : "iPhone is an Apple product."

Les phrases objectives peuvent impliquer des opinions ou des sentiments et des faits indésirables .Par exemple, les deux phrases ces dessous qui indiquent que certains faits impliquent clairement des sentiments (qui sont des opinions implicites) sur leurs produits respectifs parce que les faits sont indésirables:

"The earphone broke in two days."

"I brought the mattress a week ago and a valley has formed".

2.4.2. Emotion :

Les émotions sont nos sentiments et nos pensées subjectives. Les émotions ont été étudiées dans des multiples domaines, par exemple, la psychologie, la philosophie, et la sociologie.

Les émotions sont étroitement liées aux sentiments. La force d'un sentiment ou l'opinion est typiquement liée à l'intensité de certaines émotions, par exemple : la joie et la colère. Les opinions que nous étudions dans l'analyse des sentiments sont pour la plupart des évaluations (mais pas toujours). Selon la recherche sur le comportement des consommateurs, les évaluations peuvent être classées en deux catégories: les évaluations rationnelles et des évaluations émotionnelles.

- **Evaluation rationnelle :**

Ces évaluations sont du raisonnement rationnel, tangible croyances et attitudes utilitaristes. Par exemple, les phrases suivantes exprimes des évaluations rationnelles: "The voice of this phone is clear," "This car is worth the price," et "I am happy with this car."

- **Evaluation émotionnelle :**

Ces évaluations sont de nature non tangible et les réponses émotionnelles aux entités qui vont profondément dans l'état d'esprit des gens. Par exemple, les phrases suivantes expriment des evaluations émotionnelles: "I love iPhone," "I am so angry with their service people" et "This is the best car ever built." [Liu, 2012].

2.5. Traits /Features opinions :

Chaque document est représenté par un vecteur de traits relatifs aux éléments du lexique. Ces traits décrivent la présence de l'élément, son nombre d'apparition, ou encore la position dans le document.

2.5.1. Présence :

Le concept de présence est binaire, soit l'élément du lexique apparaît, soit il est absent. Ce trait a montré son supériorité par rapport aux autres dans des travaux passés.

Si un trait *f* ou l'un de ses synonymes apparaît dans une phrase *p*, *f* est appelée trait explicite dans *p*. Si ni *f*, ni aucun de ses synonymes apparaissent dans *p*, alors *f*, est appelée trait implicite dans *p* [Liu, 2010]. Par exemple le trait « la qualité vocale de téléphone » dans l'exemple suivant est explicite ''The voice quality of this phone is clear'' mais dans cet exemple ''This phone is small'' le trait « taille » n'apparaît pas dans la phrase donc il est implicite.

2.5.2. Fréquence :

Nous testons aussi la fréquence de l'élément dans le document. Cette fréquence est le rapport du nombre de fois que l'élément est rencontré sur le nombre total d'éléments rencontrés dans le document. Il a été vu que ce trait utilisé avec des uni-grammes de mots n'aide pas à produire de meilleurs résultats.

2.5.3. Position :

Dans le domaine du cinéma, l'opinion est souvent statuée soit au début, soit à la fin de la critique, le milieu du texte étant le plus souvent réservé à des détails descriptifs. Nous expérimentons l'information des positions des éléments (ex. début milieu, fin, début et fin…) dans le document, pour représenter les critiques [Liu, 2012].

2.6. Echelle et niveau d'analyse :

L'analyse de sentiment « basique » se focalise sur une seule dimension : le sentiment général est-il positif ou négatif ? Les techniques visant à déterminer un (ou plusieurs) sentiments généraux (l'envie, la colère, la frustration ou la joie).

L'analyse du sentiment a été étudiée principalement à trois niveaux :

- ✓ **Au niveau du document :**

La tâche à ce niveau est de classer si un document d'opinion entier exprime un sentiment positif ou négatif [Pang et al., 2002]. Cette tâche est communément connue sous le nom de classification des sentiments au niveau du document. Ce niveau d'analyse suppose que chaque document exprime des opinions sur une seule entité (par exemple, un seul produit). Ainsi, elle ne s'applique pas aux documents qui évaluent ou comparent plusieurs entités.

Cette analyse fonctionne bien pour des documents qui présentent un point de vue précis, mais moins pour des comparaisons car elle ne fera pas la différence entre les sujets abordés.

- ✓ **Au niveau de la phrase :**

La tâche à ce niveau va aux phrases et détermine si chaque phrase a exprimé une opinion positive, négative ou neutre (Neutre signifie habituellement pas d'opinion). Ce niveau d'analyse est étroitement lié à la classification de la subjectivité [J. Wiebe et al., 1999] qui distingue des phrases (appelées phrases objectives) qui expriment des informations factuelles à partir de phrases appelées phrases subjectives qui expriment des opinions subjectives. Cependant, nous devons noter que la subjectivité n'est pas équivalente au sentiment, car de nombreuses phrases objectives peuvent impliquer des opinions.

- ✓ **Au niveau des aspects (aussi appelé Feature level) :**

Les analyses au niveau du document et de la phrase ne permettent pas de découvrir ce que les gens ont aimé et n'a pas aimé. Le niveau d'aspect effectue une analyse plus fine. Le niveau de l'aspect était plus tôt appelé le niveau de fonctionnalité (l'extraction de l'opinion par caractéristiques et le résumé) [Hu et Liu, 2004]. Au lieu de regarder les constructions du langage (documents, paragraphes, phrases, clauses ou phrases), le niveau d'aspect regarde directement l'opinion elle-même. Elle repose sur l'idée qu'une opinion se compose d'un sentiment (positif ou négatif) et d'une cible (d'opinion).

L'objectif de ce niveau d'analyse est de découvrir des sentiments sur des entités et / ou leurs aspects. Par exemple, la phrase "The iPhone's call quality is good, but its battery life is shorts", évalue deux aspects : la qualité des appels et la durée de vie de la batterie, de l'iPhone. Le sentiment sur la qualité d'appel de l'iPhone est positif, mais le sentiment sur sa vie de la batterie est négatif. La qualité de l'appel et la durée de vie de la batterie de l'iPhone sont les cibles d'opinion.

Une cible d'opinion :

Un avis sans que son objectif soit identifié est d'une utilité limitée. Réaliser l'importance des cibles d'opinion nous aide aussi à mieux comprendre le problème de l'analyse du sentiment. Par exemple, bien que la phrase "although the service is not that great, I still love this restaurant", nous ne pouvons pas dire que cette phrase est tout à fait positive parce qu'il a un ton positif. En fait, la phrase est positive sur le restaurant, mais négatif sur son service. Dans de nombreuses applications, les cibles d'opinion sont décrites par des entités et / ou leurs différents aspects [Liu, 2012].

2.7. Outils d'analyse :

Il existe des outils permettant d'identifier le sentiment dégagé par un texte. Voici une liste non exhaustive des outils les plus connus :

- ***Werfamous*** : outil d'analyse en ligne gratuit, donnant un score de sentiment sur une échelle de -100 à 100, ainsi qu'un niveau de confiance lié à ce score.
- ***AFINN*** : évalue la positivité/négativité d'un mot à l'aide d'un dictionnaire contenu dans une archive ;
- ***General Inquirer*** : lemmatise les mots, effectue une analyse graphique et statistique et produit un rapport contenant des phrases avec les mots les plus significatifs ;
- ***SenticNet*** : analyse avancée de la polarité de mots en prenant en compte leurs nuances ;
- ***WordNet*** : permet de savoir à l'aide de groupe de synonymes si un mot est positif ou non ;
- ***SentiWordNet*** : il s'agit d'une extension à WordNet ; il attribue à chaque groupe de synonymes provenant de WordNet, trois scores de sentiment : la positivité, la négativité, l'objectivité ;
- ***SentiSense*** : il s'agit également d'un travail basé sur WordNet permettant de polariser les mots de façon plus précise ;
- ***Subjectivity Lexicon*** : permet d'obtenir la polarité d'un mot tout en prenant soin d'analyser son contexte d'utilisation ;
- ***MicroWNOp*** : se base à la fois sur General Inquirer et WordNet ; il fournit des ensembles de mots positifs, négatifs et objectifs qui sont synonymes.

3. Classification d'opinion :

3.1. Définition:

La classification du texte est une technique qui attribue une étiquette à un ensemble des objets d'entrée non marqués. La classification d'opinions est une tâche importante

en opinion-mining, il s'agit de regrouper les opinions des utilisateurs envers un objet selon leurs orientations sémantiques en trois catégories : *positive*, *négative*, et *neutre*.

3.2. Polarité de l'opinion :

La polarité est une sous-tâche de la détection d'opinions. Elle consiste de façon générale à déterminer si l'opinion du document sur le sujet est positive ou négative.

3.3. Orientation sémantique :

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

Exemple :

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

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

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

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

4. Difficulté de la fouille d'opinion :

L'étude des opinions, sentiments, émotions ou jugements d'évaluation à partir de données écrites ou orale se généralise de plus en plus alors que les connaissances sur ces sujets ne sont encore que très parcellaires.

En effet, la caractérisation et la modélisation de l'information subjective (opinions, sentiments, émotions, jugements d'évaluation, etc.) constituent un problème scientifique difficile à traiter. Les phénomènes subjectifs recouvrent des concepts riches, divers et dont les frontières sont difficiles à établir. Certains modèles se limitent à une classification en polarité (positif, neutre, négatif), d'autres proposent une modélisation plus fine en intégrant des arguments tels que la source (la personne qui exprime l'opinion, le sentiment ou l'émotion), la cible (l'objet de l'expression d'opinion) et l'expression de l'opinion/sentiment/émotion/jugement d'évaluation. Cependant, la modélisation du discours subjectif reste partielle, en effet, une analyse complète nécessite l'intégration d'autres informations indispensables telles que la classe sémantique (colère, peur, joie, plaisir, surprise, etc.) ou encore le degré d'engagement de la source [Site2, 2015].

Différentes problématiques sont envisageables :

- ✓ La modélisation linguistique et informatique de la subjectivité dans la langue (orale et écrite et signée),
- ✓ l'expression de la subjectivité en langue et en discours, à l'écrit ou à l'oral la construction,
- ✓ l'acquisition et la validation de ressources linguistiques (dictionnaires, lexiques, ontologies, corpus, etc.).

Pour l'analyse de la subjectivité :

- ✓ Les méthodes et les outils pour identifier, annoter et extraire des opinions, sentiments et émotions portées ou mentionnées par le langage dans des documents textuels ou audiovisuels ;
- ✓ la présentation synthétique de la diversité des contenus subjectifs.

Nous utilisons l'exemple suivant sur l'iPhone pour introduire le problème (un nombre est associé à chaque phrase pour faciliter la consultation) :

"(1) I bought an iPhone a few days ago. (2) It was such a nice phone. (3) The touch screen was really cool. (4) The voice quality was clear too. (5) Although the battery life was not long, that is ok for me. (6) However, my mother was mad with me as I did not tell her before I bought it. (7) She also thought the phone was too expensive, and wanted me to return it to the shop. … "

- La question est:

Ce que nous voulons extraire de cet examen? La première chose que nous pouvons remarquer, c'est qu'il y a plusieurs opinions dans cette revue.

Les phrases (2), (3) et (4) expriment des opinions positives, tandis que les phrases (5), (6) et (7) expriment des opinions négatives ou des émotions. Ensuite, nous remarquons également que les opinions ont tous des cibles ou des objets sur lesquels les opinions sont exprimées.

L'opinion contenue dans la phrase (2) concerne l'iPhone dans son ensemble et les opinions exprimées dans les phrases (3), (4) et (5) sont sur "touch screen" l'écran tactile, "voice quality" la qualité vocale et "battery life" la durée de vie de la batterie de l'IPhone.

L'opinion dans la phrase (7) est sur le prix de l'iPhone, mais l'opinion / émotion dans la phrase (6) est sur "me" "moi, pas iPhone. C'est un point important.

Dans une application, l'utilisateur peut être intéressé par des opinions sur certaines cibles ou objets, mais pas sur tous (par exemple, improbable sur "me" moi).

Enfin, nous pouvons également remarquer les sources ou les détenteurs d'opinions. La source ou le titulaire des opinions dans les phrases (2), (3), (4) et (5) est l'auteur de la revue ("I"), mais dans les phrases (6) et (7) est "my mother" ma mère.

Avec cet exemple à l'esprit, nous définissons maintenant formellement les problèmes de l'analyse des sentiments. Un des problèmes c'est l'objectif d'opinion.

En général, les opinions peuvent être exprimées sur quoi que ce soit, par exemple un produit, un service, un individu, une organisation, un événement ou un sujet. Nous utilisons le terme objet pour désigner l'entité cible qui a été commentée. Un objet peut avoir un ensemble de composants (ou de parties) et un ensemble d'attributs (ou de propriétés).

Chaque composant peut avoir ses propres sous-composantes et son ensemble d'attributs, et ainsi de suite. Ainsi, un objet peut être décomposé hiérarchiquement en fonction de la partie de la relation [Liu, 2010].

PARTIE 02

5. Traitement Automatique des Langues Natural TALN :

Le Traitement Automatique des Langues (TAL), appelé « Natural Language Processing » (NLP) en anglais, Le Traitement Automatique du Langage Naturel (TALN) est un domaine à la fois scientifique et technologique en plein essor qui débouche sur des applications très diverses : correction automatique des erreurs, analyse de textes, génération automatique de résumés, extraction de connaissances, interrogation de bases de données en langage naturel, aide à la traduction, etc. elle a pour objectif de traiter des données exprimées dans une langue dite naturelle. Le TAL cherche donc à concevoir des logiciels et des programmes pour traiter ces données linguistiques qui peuvent être écrites ou orales.

Parmi les applications du TAL, on retrouve, le traitement de la parole, la traduction automatique, la correction orthographique, la compréhension et la génération automatique de texte, la recherche d'information et la fouille de textes, l'extraction d'information, le résumé automatique de texte, les systèmes de questions/réponses…
Dans cette section, nous allons voir quelles sont les différentes étapes d'analyse réalisées lors du traitement automatique et ces difficultés [Cadilhac, 2010].

5.1. Les niveau d'analyse de TALN :

Le traitement automatique des langues effectue une analyse qui peut se découper en plusieurs étapes :

Analyse Morphologique :

Cette analyse se compose de deux sous-étapes [Cadilhac, 2010] :

La segmentation : permet d'identifier les morphèmes, appelés tokens en anglais. Les morphèmes sont les unités élémentaires du texte c'est-à-dire des mots simples ou composés du texte,

La lemmatisation : prend en entrée les résultats de la segmentation et identifie la racine des mots, appelée lemme, en supprimant les conjugaisons, les suffixes et les préfixes, les accords en genre et en nombre…

Analyse Lexicale :

Lors de cette étape, on définit la catégorie grammaticale (nom, verbe, adjectif,…) de chaque mot. Il y a beaucoup d'ambiguïtés pour trouver la catégorie grammaticale d'un mot car chaque mot pris hors contexte peut appartenir à une ou plusieurs catégories.

Pour résoudre ce problème, plusieurs méthodes sont envisageables. L'outil *Intex1* utilise les règles linguistiques, l'outil *TreeTagger2* utilise des modèles probabilistes, l'outil *Brill3* utilise un étiquetage par transformations [Cadilhac, 2010].

Analyse Syntaxique :

Cette étape permet de définir comment les mots sont organisés entre eux pour former des phrases. Elle permet de reconnaître des groupes de mots (nominaux, verbaux…) que l'on appelle *syntagmes*.

L'analyse syntaxique consiste donc à étudier les contraintes entre les groupes de mots qui doivent être respectées pour former des structures autorisées par la langue. Son but est de fournir un arbre syntaxique qui représente la structure de la phrase [Cadilhac, 2010].

Analyse Sémantique :

Cette analyse permet de donner un sens aux mots et aux phrases du texte. Elle peut se découper en deux sous-tâches [Cadilhac, 2010].

La sémantique lexicale : permet de reconnaître le sens des mots. Souvent les mots du dictionnaire ont plusieurs sens possibles donc il peut être difficile de trouver le sens particulier du mot tel qu'il est employé dans un certain contexte.

- *La polysémie :* propriété pour un mot d'avoir plusieurs sens différents qui possèdent des relations sémantiques.
- *L'homonymie :* propriété pour un mot de renvoyer à deux formes identiques pour lesquelles il n'y a pas de relation sémantique.
- *La synonymie :* propriété qui lie deux mots ayant le même sens.
- *L'hyponymie et l'hyperonymie :* propriétés qui expriment des rapports d'inclusion.
- *La méronymie et l'holonymie :* propriétés qui expriment des relations partie/tout.

La sémantique grammaticale : permet de reconnaître le sens des phrases. Cette analyse nécessite de résoudre plusieurs ambiguïtés sémantiques, telles que :

- *L'anaphore :* reprend un segment du discours antérieur par un autre élément grammatical pour éviter la répétition.
- *L'ellipse :* consiste à omettre un élément dans une construction qui aurait dû le contenir pour éviter la répétition.

Analyse Pragmatique :

Cette analyse prend en entrée le résultat de l'analyse sémantique et ajoute les informations liées au contexte énonciatif donné (texte, dialogue…) [Cadilhac, 2010].

La présupposition : information implicite d'un énoncé, c'est-à-dire information qui n'est pas explicitement formulée mais qui peut être déduite Par exemple, grâce à la phrase « Johnny stop smoking the cigarette. », nous pouvons déduire que Johnny smoked.

Le déictique : ensemble des informations liées à un contexte d'énonciation particulier. Par exemple, dans la phrase «We will go to the cinéma tomorrow. », il faut connaitre la date en cours pour savoir à quel jour fait référence « tomorrow ».

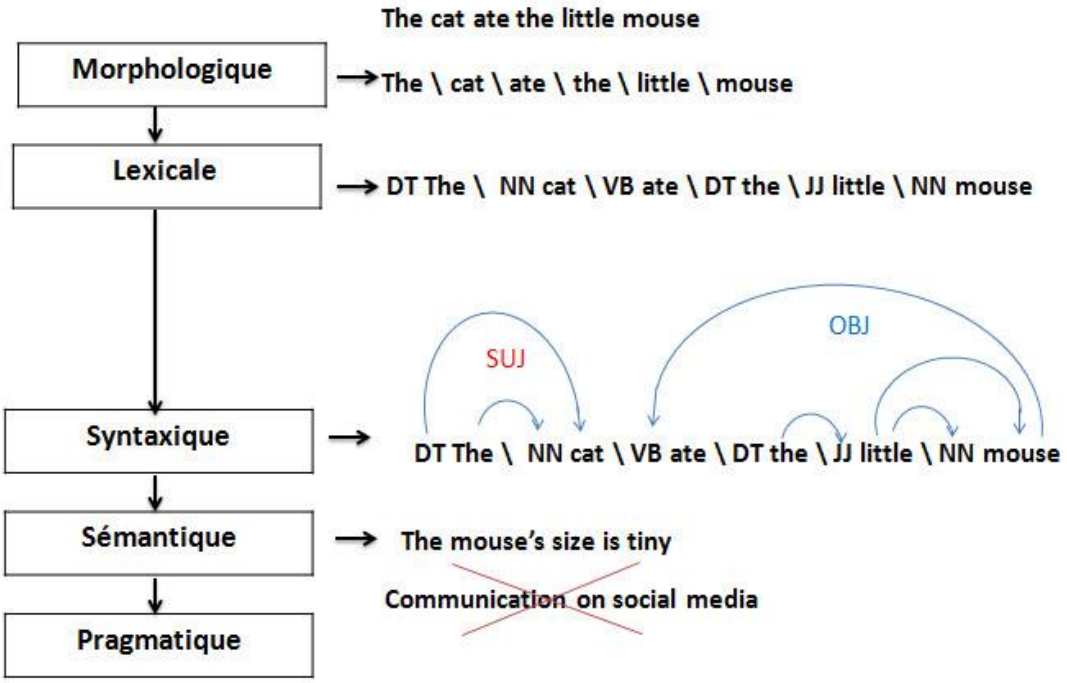


FIGURE 1.1: *Les niveaux d'analyse de TALN exemple anglais*

5.2. Les difficultés du TALN : ambiguïté et implicite :

Les difficultés que l'on rencontre en TALN sont principalement de deux ordres, et ressortent soit de l'ambiguïté du langage, soit de la quantité d'implicite contenue dans les communications naturelles.

Ambiguïté :

Le langage naturel est ambigu, et ce à quelque niveau qu'on l'appréhende. Cette ambiguïté, est un de ses traits caractéristiques. On peut d'ailleurs voir là le résultat d'un compromis inévitable entre d'un côté une capacité d'expression quasi illimitée, et de l'autre des contraintes liées à la limitation des ressources physiologiques mises en œuvre (taille de la mémoire à long et court-terme, densité de l'espace phonétique, contraintes articulatoires, etc.) [Yvon, 2007].

Cette ambiguïté se manifeste par la multitude d'interprétations possibles pour chacune des entités linguistiques pertinentes pour un niveau de traitement, comme en témoignent les exemples suivants :

- ✓ ambiguïté des graphèmes (lettres) dans le processus d'encodage orthographique.
- ✓ ambiguïté des terminaisons dans les processus de conjugaison et d'inflexion.
- ✓ ambiguïté dans les propriétés grammaticales et sémantiques (i.e. associées à son sens) d'une forme graphique donnée.
- ✓ ambiguïté de la fonction grammaticale des groupes de mots.
- ✓ ambiguïté de la portée des quantificateurs, des conjonctions, des prépositions.
- ✓ ambiguïté sur l'interprétation à donner en contexte à un énoncé.

Implicite :

L'activité langagière s'inscrit toujours dans un contexte d'interaction entre deux humains, sensément dotés d'une connaissance du monde et de son fonctionnement telle que l'immense majorité des éléments de contexte nécessaires à la désambiguïsation mais aussi à la compréhension d'un énoncé naturel peuvent rester implicites.

La situation change du tout au tout dès qu'une machine tente de s'insérer dans un processus de communication naturel avec un humain : la machine ne dispose pas de cette connaissance d'arrière-plan. Ce qui rend la compréhension complète de la majorité des énoncés difficile, voire impossible, si l'on ne dispose pas de bases de connaissance additionnelles, donnant accès à la fois à un savoir sur le monde (ou le domaine) en général (connaissance statique) et sur le contexte de l'énonciation (connaissance dynamique).

A l'absence de telles connaissances, bien d'autres problèmes de compréhension deviennent pratiquement insurmontables : pensez par exemple aux ellipses, aux métaphores, et, plus généralement, aux figures de style [Yvon, 2007].

PARTIE 03

6. La langue arabe :

L'arabe "العربية" la langue parlée à l'origine par les Arabes, est une langue sémitique dont l'origine est l'arabe classique c'est à dire la langue du Coran. Il correspond à une évolution moderne de l'arabe classique. Ainsi, même s'ils existent de nombreux dialectes, les locuteurs dont la langue arabe est la langue maternelle peuvent communiquer entre eux. L'arabe étant la langue officielle de 26 pays et de plusieurs organismes internationaux comme l'OMS (Organisation Mondiale de la Santé), de nombreux documents administratifs et techniques sont produits dans cette langue.

Nous allons parler brièvement de la langue arabe et ces caractéristiques.

Caractéristiques de l'arabe :

- L'alphabet de la langue arabe est un *abjad*[1] compte 28 consonnes.

- L'arabe s'écrit et se lit de droite à gauche.

- Les lettres changent de forme de présentation selon leur position.

A la fin d'une lettre Non joignable	A la fin	Au milieu	Au début
ع		ـعـ	

TABLEAU 1.1: *Variation de la lettre ع*

- Un mot arabe s'écrit avec des consonnes et des voyelles. Les voyelles sont ajoutées au- dessus ou au-dessous des lettres et nous comptons six voyelles, et une voyelle muette.

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

TABLEAU 1.2: *Tableau des voyelles en arabe*

[1] Abjed est appelé aussi alphabet consonantique, désigne un alphabet ne notant que des consonnes(ou notant principalement les consonnes), comme en arabe ou en hébreu.

Mot sans voyelles	1ere Interprétation		2eme Interprétation		3eme Interprétation	
بـتـك	بَـتـك	Il a écrit	بَـتـبْ ك	Il a été écrit	بَـتـبُ ك	Des livres

TABLEAU 1.3: *Interprétation de mot بتك sans voyelle*

- La langue arabe comprend trois catégories de mots: verbes, noms et particules.

- Les verbes et les noms sont le plus souvent dérivés d'une racine à trois consonnes.

- Une famille de mots peut être générée à partir d'une seule racine à l'aide de différents schèmes.

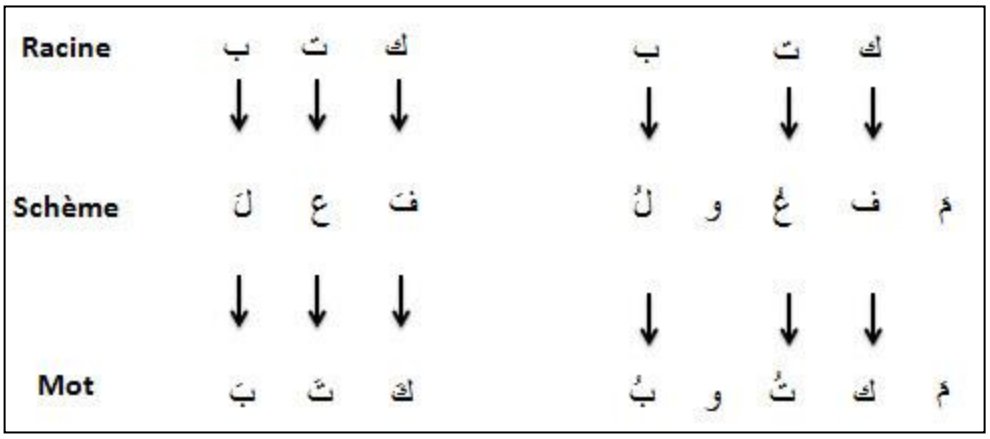


FIGURE 1. 2 : *Exemple de dérivation de la racine* ***بتك*** *ktb*

فعل	حمل	Notion de porter
فاعل	حَامِل	Porteur
فعل	حَمَلَ	A porté
مفعل	مَحمَل	Brancard
فعل	حُمِلَ	A été porté

TABLEAU 1.4 : *Schèmes de dérivés du mot* ***حمل***

- La représeŶtatioŶ suivaŶte schéŵatise uŶe structure possidle d'uŶ ŵot.

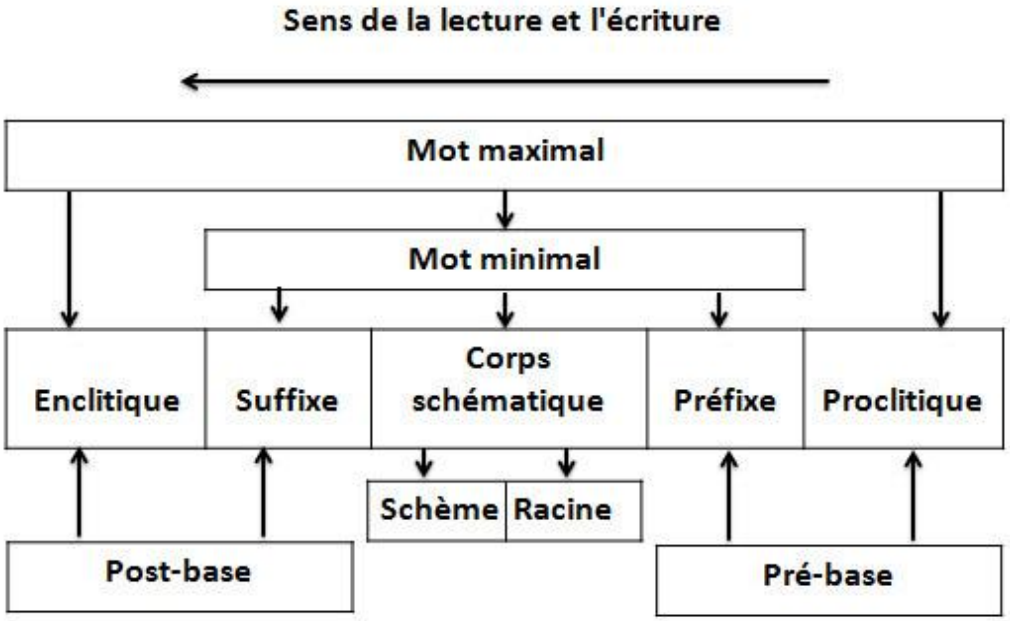


FIGURE 1.3 : *Structure d'un mot arabe*

-UŶ ŵot ŵaxiŵal est coŵposé d'uŶe ďase, pré-base et post-base. La pré-base est elle-même coŵposée d'uŶ proclitiɑue et d'uŶ préfixe. La post-ďase est coŵposée d'uŶ eŶclitiɑue et d'uŶ suffixe. A partir de la base ; nous pouvons extraire la racine et le schème. La figure précédente illustre ce modèle.

- Les proclitiques, les préfixes, les suffixes et les enclitiques sont des morphèmes qui expriment des informations et des traits grammaticaux sur les mots.

On a le mot arabique : هنونرعنأ

-Ce mot exprime la phrase en français: "Est-ce que vous le connaissez? "

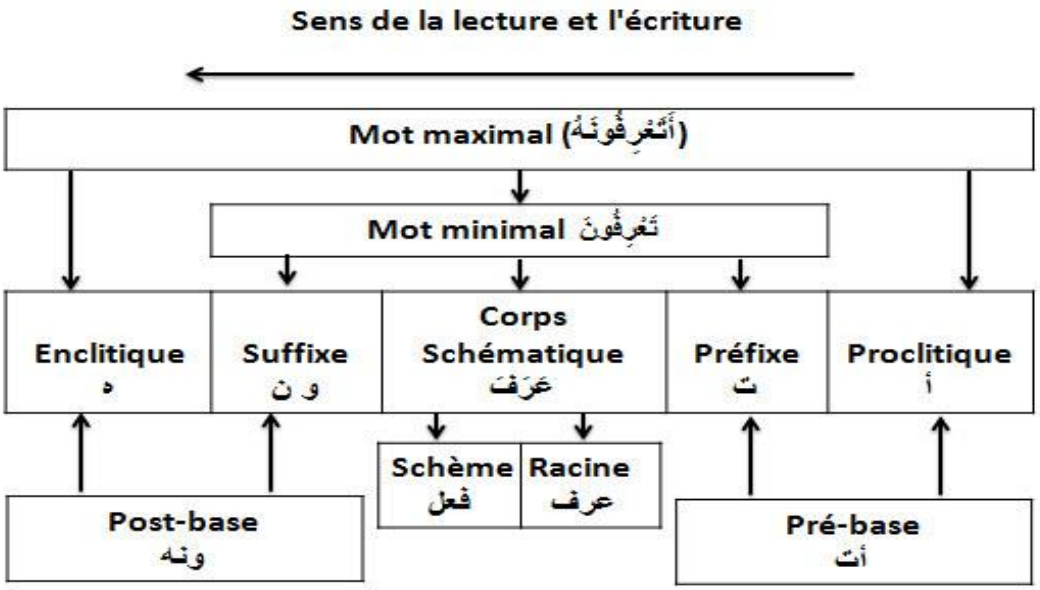


FIGURE 1.4 : *Modèle du mot graphique en arabe appliqué pour le mot «* هنونرعنأ *»*

Proclitique أ : Conjonction d'iŶterrogatioŶ.

Préfixe E : préfixe verbal du teŵps de l'inaccompli.

Corps Schématique فعـرع : dérivé de la racine فرع: selon le schème لعف

Suffixe ϥو : suffixe verbal exprimant le pluriel.

Enclitique ـه : pronom suffixe complément du nom.

6.1. Composition du lexique arabe : les formes de bases :

La grammaire traditionnelle arabe ne connaît que trois sous-ensembles : noms, verbes et particules. Ce classement montre, rapidement, ses limites quand il s'agit d'un traitement informatisé de la langue.

6.1.1. Les verbes :

Un verbe est une entité exprimant un sens dépendant du temps. La majorité des verbes arabes sont formés sur des radicaux de 3 consonnes tel est le cas du verbe "بـتـك " (kataba – écrire) et éventuellement 4 consonnes tel est le cas du verbe " Nحـدر " (dahrağa – glisser, faire glisser).

Ces racines peuvent donner naissance à plusieurs schèmes ou patrons à la suite d'une ou plusieurs transformations morphologiques (par exemple, le redoublement d'une consonne, l'allongement d'une voyelle, l'adjonction d'un morphème, etc.), il s'agit dans ce cas de racines à schème augmenté [Mesfar, 2008]. Il nous importe d'introduire un classement des verbes selon leurs radicaux :

- **Les verbes à racine simple:**

Les verbes à racine simple ont une base à trois consonnes appelées consonnes radicales. Ces verbes sont associés au schème verbal " لعـف " (faʿala). Lorsqu'aucune des consonnes radicales du verbe n'est une voyelle longue, il est dit sain. Ces radicaux peuvent comporter de causes de transformation ou défectuosité (ﻝﻼﻋ), nous citons :

✓ La présence d'un glide " أ " (*á – hamzä*), " ي " (*y - yâ'*) ou " و " (*w – wâw*) parmi les consonnes du radical. Selon la position de ce glide, nous distinguons différents types de verbes :

- Si l'une des consonnes radicales est " أ " (à, hamzä), indépendamment de sa position : *verbe hamzé* « ﻡﻮﻤﻬ » (mahmuwz) ;
- La 1ère consonne radicale est un " و " (w) ou " ي " (y) : verbe assimilé « ﻡﺎﻤﺜ » (mital) ;
- La 2ème consonne radicale est un " و " (w) ou " ي " (y) : verbe creux « أﻒﻮﺟ » (ağwaf) ;

- La 3ème consonne radicale est un " و " (w) ou " ي " (y) : verbe défectueux « صقان » (naãqis) ;

✓ La présence de deux consonnes identiques en deuxième et troisième position du radical est *verbe redoublé* « فعاضم » (*mudaā'af*).

Notons que, les causes de défectuosité peuvent être combinées pour avoir des transformations doubles ou triples au sein d'une même racine.

- **Les verbes à racine augmentée:**

Aux schèmes simples, décrits ci-dessus, vient s'ajouter une dizaine de schèmes verbaux augmentés. Ces schèmes sont formés à partir de racines simples par un ensemble d'opérations morphologiques pour donner un sens particulier aux verbes résultats, nous citons :

Schème verbal augmenté	**Opération morphologique associées**
« لعفل » (*fa''ala*)	Redoublement de la deuxième consonne radicale
« لعافل » (*faā'ala*)	Allongement de la première consonne radicale par l'ajout d'un " ا " (*ā – álif)*
« لعفأ » (*áfa'ala*)	Adjonction d'une " أ " (*á – hamzä*) au début de la racine
« لعفت » (*tafa'ala*)	Adjonction d'une " ت " (*t*) au début de la racine + redoublement de la deuxième consonne radicale
« لعافت » (*tafaā'ala*)	Adjonction d'une " ت " (*t*) au début de la racine + allongement de la première consonne radicale par l'ajout d'un " ا " (*ā – álif)*
« لعفتا » (ìfta'ala)	Adjonction d'un " ا " (ā – álif) au début de la racine + insertion d'un morphème mono-consonantique " ت " (t) à la suite de la première consonne
« لعفنا » (*ìnfa'ala*)	Adjonction d'un morphème bi-consonantique " نإ " (*ìn*) au début de la racine
« لعفستا » (ìstaf'ala)	Adjonction d'un morphème tri-consonantique " ستا " (ìsta) au début de la racine

TABLEAU **1.5 :** *Principaux schèmes verbaux augmentés*

6.1.2. Les noms :

Le système morphologique des noms arabes distingue trois sous-catégories :

- **Les noms primitifs:**

Ce sont des noms qui ne peuvent pas être rattachés à une racine verbale.
Ils paraissent bien constituer le glossaire fondamental de la langue concrète. Exemple : رأس (raás – tête), كرسي (kursiyy – siège), كبش (kabš – bélier) etc. Dans cette catégorie, nous incluons aussi les noms à racine bilitère tels que : دم (dam - sang), فم (fam - bouche), أب (áab – père), أخ (ákh – frère), etc [Mesfar, 2008].

- **Les noms dérivés ou déverbaux:**

Ce sont les noms qui peuvent être dérivés à partir d'une racine verbale. Le nombre et la nature de ces formes varient selon le statut du verbe auquel ils se rattachent. En tant que noms, ils peuvent recevoir les marques du cas, du genre et de l'indétermination.

- **Les nombres:**

Cette catégorie de noms est constituée par les numéros simples représentant les unités : de " صفر " (sifr- zéro, 0) à " تسعة " (tisʿat – neuf, 9) ; les dizaines : " عشرة " (ʿašarat – dix, 10) ,"عشرون " (ʿišruwn – vingt, 20), …et " تسعون " (tisʿuwn – quatre-vingt-dix, 90) ; les centaines, etc. ainsi que les numéros composés tels que les cardinaux de " أحد عشر " (áahada ʿašara - onze, 11) à " تسعة عشر " (tisʿat ʿašara - dix-neuf,19).
Dans leur décomposition, les grammairiens arabes ont assimilé les adjectifs à des noms vu qu'ils y prennent presque toutes les formes morphologiques ; ils peuvent, par exemple, être définis ou indéfinis et se fléchir suivant le cas, le nombre et le genre [Mesfar, 2008].

6.1.3. Les pronoms :

La classe des pronoms a été introduite en tant qu'extension du système de décomposition traditionnel du lexique de la langue arabe. Cette classe regroupe quelques formes qui étaient considérées comme des noms particuliers. Cet ensemble échappe à toute règle de dérivation. Les pronoms forment une liste fermée de mots. Dans cette liste, nous distinguons :

- **Les pronoms démonstratifs " أسماء الإشارة ":**

Ils représentent une sous-catégorie de pronom exprimant une idée de démonstration. Ils permettent d'indiquer que l'objet représenté se trouve, soit dans le texte, soit dans l'espace ou le temps, défini par la situation d'énonciation. Ils existent deux sous-ensembles : les démonstratifs de proximité (par exemple : " هذا " (hadaā – celui-ci), " هؤلاء " (haŵulaā' – ceux-ci) et les démonstratifs d'éloignement (par exemple : " ذلك " (dalika – celui-là), " أولئك " (ùuwlaāýika – ceux-là), etc.). Les démonstratifs ne sont déclinables qu'au duel.

- **Les pronoms relatifs " الأسماء الموصولة ":**

Ils se rapportent au nom ou au pronom personnel qui les précède et que nous désignons par antécédent.

Les relatifs s'accordent avec l'antécédent et ne sont déclinables qu'au duel (comme les démonstratifs).Parmi les pronoms relatifs, nous citons : " الذي " (al-ladiy - celui), " اللتين " (al-latayni – celles, féminin, duel), " اللذين " (al- ladayni – ceux, masculin, pluriel), etc.

- **Les pronoms personnels " الضمائر المنفصلة ":**

Ils servent à désigner les trois types de personnes grammaticales :

- ✓ La première personne, c'est-à-dire, l'énonciateur ou locuteur : " أنا " (ánaā - je) ou " نحن" (nahnu – nous) ;
- ✓ La deuxième personne, c'est-à-dire, le destinataire ou interlocuteur : " أنت " (áanta – tu, masculin), " أنت " (áanti – tu, féminin), " أنتما " (áantumaā – vous, duel), " أنتم " (áantum – vous, masculin, pluriel), " أنتن " (áantunna – vous, féminin, pluriel);
- ✓ La troisième personne, c'est-à-dire, la personne absente, celle dont on parle : " هو " (huwa – il), " هي " (hiya – elle), " هما " (humaā – ils, duel), " هم " (hum – ils), " هن " (hunna – elles) ;

6.1.4. Les mots outils :

Les mots outils sont des entités qui servent à situer des faits ou des objets par rapport au temps ou au lieu. Ils jouent également un rôle clé dans la cohérence et l'enchaînement d'un texte. Par exemple, nous avons des particules qui désignent un temps : بعدَ (baʿda – après) قبلَ (qabla – avant)، منذ (mundu – depuis) ou un lieu tel que حيثُ (haytu – où),…. Selon leu sémantique et leur fonction dans la phrase, ils peuvent jouer un rôle important dans l'interprétation d'une phrase en exprimant une introduction, explication, conséquence, etc. [Kadri et Benyamina, 1992]. Les mots outils incluent différentes catégories, nous citons :

- **Les prépositions:** par exemple, " في " (fiy – dans) ou " على " (ʿalaą – sur);
- **Les conjonctions de coordination:** par exemple, " ثم " (tumma – ensuite, puis) ;
- **Les adverbes:** par exemple, " أبدًا " (abadã – jamais) ou " بشكلٍ عادي " (bišaklĩ ʿaādiyyĩ – normalement, de façon normale) ;
- **Les quantificateurs:** par exemple, " كلّ " (kulla – tout) ou " بعض " (baʿda – un peu);

Les mots outils se répartissent en sous-groupes : ceux qui sont variables (quantificateurs) et ceux qui sont invariables (adverbes, prépositions, etc.).

6.2. Difficulté de TALN de la langue arabe :

Nous avons présenté d'une manière brève la langue arabe et ses caractéristiques, maintenant nous allons décrives les difficultés que les rencontres en cours de traitement automatique de la langue arabe.

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

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

- **Les Voyelles :**

L'ambiguïté vient du mot علم science ou drapeau alors que voyellé on aura عِلْم pour la science et عَلَم pour le drapeau.

La Hamza et le Alif :

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

Exemple :

- فسا signifie « il s'est écoulé » « il a coulé », du verbe ليسي - فسا (racine ليس).
- فسأ signifie « il a questionné » ou « il a posé une question », du verbe فسأي – فسأ (racine فسأ).

Beaucoup de mots minimaux ont un homographe [2] avec et vice-versa, et ils ne sont pas toujours de la même famille. Exemple :

- Le nom ةرامإ (*amarat*) signifie « un marque de… » ou « signifie de »
- Le nom ةرامأ (*imarat*) signifie «Emirat »

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

- قرا «liée deux… » ou « mariage ».
- قرآ « le Coran ».

Le Ya et le Alif Maqsûra :

Dans l'usage typographique [3] égyptien par exemple, les auteurs notent la lettre

[2] Homographe préfixe grec homo (même), racine grecque graphe (graphie, écrit) Des homographes sont des mots qui s'écrivent de la même façon.

[3] Typographe préfixe grec typo (empreinte, marque), racine grecque graphe (écrire) typographique désigne les différents procédés de composition et d'impression utilisant des caractères et des formes en relief

ﻯ à la fin des mots sans point dessous, ce qui la rend équivalente à ﻯ Alif maqsûra. L'absence des deux points change totalement le mot, et pose un vrai problème de reconnaissance de la forme écrite.

A l'image du premier point, la plus part des mots se terminent avec Alif Maqsûra ont un homographe avec le Ya. Exemple :

- Le mot outils ﻊﻣ .
- Le nom propre ﻲﻨﻋ .

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

- **Le Tanwin :**

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

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

- **Le caractère '-' :**

Les typographes[4] font un usage fréquent du caractère '-' (appelé Kashida), qui permet l'allongement du trait au milieu des mots, pour une meilleure lisibilité, pour limiter les espaces blancs sur une ligne justifiée, voire pour des raisons purement esthétiques. Or cet usage peut nuire aux analyses automatiques : ce caractère ne fait pas partie de l'alphabet arabe, il est considéré comme un intrus par le système d'analyse automatique. Il faut donc recourir à un sous-programme particulier afin de l'éliminer. Exemple : le mot ﻟﺎﻤﺘﺤﻣ : peut-être écrit de plusieurs façons : ﻟﺎﻤﺘـﺤﻣ, ﻟـﺎﻤﺘﺤﻣ, ﻟﺎﻤﺘـﺤﻣ, …etc. [Abbès, 2004].

6.2.1. Difficultés de l'analyse morphologique de l'arabe :

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

Le nombre de catégories auxquelles un mot peut appartenir dépend du jeu d'étiquettes choisies. En moyenne, le nombre d'étiquettes par unité lexicale

[4] Typographe un artisan ou un ouvrier travaillant dans le domaine de la typographie.

voyellée est de 9, alors que pour les unités lexicales non voyellées, elles peuvent atteindre 12 étiquettes par unité lexicale.

Un autre facteur pour la langue arabe, est le nombre d'unités lexicales ambigües dans un texte. Contrairement au français et à l'anglais, il peut concerner 66% des unités lexicales qui composent le texte. Ces difficultés peuvent poser plusieurs ambigüités au niveau sémantique et syntaxique [Abbès, 2004].

- **Ambiguïtés dérivationnelles et flexionnelles :**

La flexion est la variation de la forme des mots en fonction de facteurs grammaticaux tels que la conjugaison pour les verbes

Par exemple : le mot " ونأثرهم " (ils s'influencent) est le résultat de la concaténation du préfixe " ي" indiquant le présent et du suffixe " ون" indiquant le masculin pluriel d verbe " رثأ ".

Le problème en analyse morphologique de l'arabe se rapporte surtout au niveau de la dérivation qui est un phénomène plus complexe que la flexion. En effet, la dérivation est la formation de nouveaux mots à partir de mots existants. Dans le cas de la langue arabe, la plupart des mots sont dérivés à partir de racines trilitères ou quadrilatères.

Le mot arabe n'est pas le résultat d'une simple concaténation de morphèmes comme c'est le cas pour l'anglais par exemple : unfailingly = un+fail+ing+ly, mais c'est à partir d'une racine, d'une combinaison de voyelles, de préfixes, d'infixes, de suffixes et d'un schème morphologique qu'on obtient un mot par exemple : à partir de la racine " رثأ " (choisir/citer à) on peut dériver plusieurs verbes tel que " رثأن " (s'influencer) et plusieurs noms tel que " رثأنم " (Ému) [Hadrich et Chaâben, 2006].

- **Ambiguïtés d'agglutination :**

Contrairement aux langues latines, en arabe, les articles, les prépositions, les pronoms, etc. collent aux adjectifs, noms, verbes et particules auxquels ils se rapportent. Comparé au français, un mot arabe peut parfois correspondre à une phrase française [Souissi, 1997] par exemple : le mot arabe "هنونرعتأ " correspond en français à la phrase «Est-ce que vous le connaissez?».

Cette caractéristique engendre une ambiguïté morphologique au cours de l'analyse. En effet, il n'est pas toujours facile de distinguer un proclitique ou enclitique d'un caractère original du mot. Par exemple, le caractère " و " dans le mot " لصو " (il est arrivé) est un caractère original alors que dans le mot " حتفو " (et il a ouvert), il s'agit plutôt d'une proclitique.

- **Ambiguïtés dues au non voyellation :**

La morphologie arabe est assez régulière lorsque les mots sont présentés sous leurs formes voyellées. Cependant, la majorité des documents arabes sont non

voyellées sauf pour le Coran et pour certains ouvrages scolaires pour débutants et donc c'est pour cette raison que nous nous sommes intéressés à l'arabe non voyellé. En fait, les mots non voyellées engendrent beaucoup de cas ambigus au cours de l'analyse par exemple : le mot non voyellé " لصف " pris hors contexte peut être un verbe au passé conjugué à la troisième personne du singulier " لصف "(il a licencié), ou un nom masculin singulier " لصف " (chapitre/ saison), ou encore une concaténation de la conjonction de coordination " ف " (puis) avec le verbe " لص "
: impératif du verbe lier conjugué à la deuxième personne du singulier masculin.

7. Conclusion :

Dans ce premier chapitre nous avons présenté les notions fondamentales en opinion-mining (fouille d'opinion), en vue de simplifier la compréhension des concepts de base qui ont généralement une orientation philosophique que linguistique ou informatique ;

Le traitement automatique de la langue naturel (TALN) ses niveaux d'analyse et ses difficultés ;

Les difficultés du traitement automatique de la langue arabe ainsi que les différents types d'ambigüités qui peuvent être imposées, et quelques techniques de désambigüisation comme la segmentation de textes et la détection des racines ;
Le lexique utilisé (les préfixes, les suffixes, les enclitiques, les proclitique, le pré-bases, les postes-bases) en traitement automatique de l'arabe, nous permet de segmenter les mots pour nous permettre de séparer les unités lexicales étiquetés ;

Le chapitre suivant sera beaucoup plus technique et comportera les travaux connexes réalisée pour la classification d'opinion.

CHAPITRE II :
Travaux Connexes

1. Introduction

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

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

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

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

2. Travaux de recherches de fouille d'opinion :

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

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

- ✓ Identification des caractéristiques des produits sur lesquelles, les clients ont donné leurs opinions ;
- ✓ Pour chaque caractéristique, identifier les phrases qui portent des opinions positives ou négatives ;
- ✓ Construire un résumé sur le produit en utilisant les opinions extraites sur les caractéristiques.

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

- Pour la qualité d'image, le nombre d'opinions positives : 256, le nombre d'opinions négatives : 6.
- Pour la taille, le nombre d'opinions positives : 134, le nombre d'opinions négatives : 10.

L'utilisation de tels résumés sur des produits différents, permettent de comparer les opinions des clients sur les produits concurrents où les clients peuvent voir visuellement les forces et les faiblesses de chaque caractéristique, en utilisant bien sûr des représentations graphiques[1].

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

Pour remédier à ce problème, les auteurs ont proposé d'utiliser une fonction d'agrégation d'opinions : une phrase *S* peut contenir multiple caractéristiques[2] ($f_1,\ldots f_m$) sur un produit et multiple mots d'opinions ($w_1,\ldots w_n$).

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

[2] Nous désignons par caractéristique, un sous-composant de l'objet ;Feature en anglais.

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

- Segmenter la phrase S, en utilisant des mots clés comme « mais », « à l'exception de… », …etc.
- Supposer que la caractéristique f_i est dans un segment S_k.
- Calculer le score de l'orientation d'opinion pour la caractéristique f_i, dans le segment S_k :

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

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

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

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

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

(1) Règle de conjonction intra-phrase

Par exemple, nous avons la phrase suivante :

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

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

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

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

Parfois, on ne peut pas employer une conjonction explicite « et ». Employons la phrase suivante :
« La durée de vie de la batterie est *longue* ».
Nous n'avons aucune idée si « *longue* » est positif ou négatif pour « la durée de vie de la batterie ». Une stratégie semblable peut être appliquée. Par exemple, dans une autre critique on peut avoir:
« La durée de vie de la batterie est *longue*, c'est *magnifique* ».

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

(3) Règle de conjonction d'inter-phrases :

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

(4) Règle des synonymes et antonymes

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

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

3. Techniques de classification de textes :

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

3.1. Classification supervisée :

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

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

Soit $D = \{d_1, d_2, \ldots, d_i, \ldots, d_m\}$, un ensemble de documents représentés chacun par une description $\vec{d}_1$, $\vec{d}_{2,\ldots,}$ $\vec{d}_m$,et $C = \{C_1, C_2, \ldots, C_k, \ldots, C_c\}$, un ensemble de classes, la classification supervisée suppose connues deux fonctions. La première fait correspondre à tout individu d_i, une classe C_k, Elle est définie au moyen des couples (d_i, C_k) donnés comme exemples au système. La deuxième fait correspondre à tout individu d_i, sa description $\vec{d}_i$. La classification supervisée consiste alors à déterminer une procédure de classification f : $C^f : d_i \rightarrow C_k$, qui à partir de la description de l'élément détermine sa classe avec le plus faible taux d'erreurs. La performance de la classification dépend notamment de l'efficacité de la description. De plus, si l'on veut obtenir un système d'apprentissage, la procédure de classification doit permettre de classer efficacement tout nouvel exemple (pouvoir prédictif) [Site2, 2008].

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

3.1.1. Algorithmes de classification supervisée :

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

Dans le cas de la classification de documents, on fournit donc à la machine des exemples sous la forme (Document, Classe). Cette méthode de raisonnement est appelée inductive car on induit de la connaissance (le modèle) à partir des données d'entrée (les Documents) et des sorties (leurs Catégories). Grâce à ce modèle, on peut alors déduire les classes de nouvelles données : le modèle est utilisé pour prédire. Le modèle est bon s'il permet de bien prédire [Site2, 2008].

Il existe de nombreuses méthodes d'apprentissage supervisé :

3.1.1.1. K plus proches voisins et ses améliorations :

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

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

3.1.1.2. Arbres de décisions :

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

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

Concrètement, chaque nœud d'un arbre de décision contient un test (un IF...THEN) et les feuilles ont les valeurs Oui ou Non. Chaque test regarde la valeur d'un attribut de chaque exemple. En effet, on suppose qu'un exemple est un ensemble d'attributs/valeurs. Pour des documents, chaque attribut peut être un mot, et la valeur sera par exemple 0 ou 1 selon que ce mot appartient ou non au document [Site4, 2008].

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

3.1.1.3. Naïve Bayes (ou Simple Bayes) :

Nommés d'après le théorème de Bayes, ces méthodes sont qualifiées de « Naïve » ou « Simple » car elles supposent l'indépendance des variables. L'idée est d'utiliser des conditions de probabilité observées dans les données. On calcule la probabilité de chaque classe parmi les exemples. Ce sont les « prior probabilities ». Par exemple, si la classe « informatique » revient 2 fois sur les 5 documents donnés en exemple, sa « prior probability » sera de 2/5. En plus des "prior probas", l'algorithme calcule les fréquences d'apparition de chaque variable d'entrée avec celles de sortie. Pour classer des documents, les variables d'entrée sont les mots présents dans l'ensemble des documents. A chaque mot on calcule

le nombre de fois qu'il apparaît dans les documents classés dans une classe donnée. On calcule cette fréquence pour chaque classe.

Une variante des Naïve Bayes sont les réseaux Bayésiens : dans ce modèle, on ne suppose plus que les variables sont toutes indépendantes, et on autorise certaines à être liées. Cela alourdit considérablement les calculs et les résultats n'augmentent pas de façon significative [Site4, 2008].

3.1.1.4. Réseaux de neurones :

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

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

Enfin, les réseaux de neurones sont appropriés si la compréhension de la fonction apprise par le réseau n'est pas essentielle. Avec un arbre de décision, l'opérateur humain peut toujours visualiser l'arbre et « comprendre » comment la machine décide. Avec un réseau de neurone, des techniques de visualisation existent, mais elles demandent généralement plus d'expertise que l'analyse d'un arbre de décision (qui peut être visualisé sous forme de règles) [Site4, 2008].

3.1.1.5. Machines à support de vecteurs (ou SVM) :

Cette technique - initiée par Vapnik - tente de séparer linéairement les exemples positifs des exemples négatifs dans l'ensemble des exemples. Chaque exemple doit être représenté par un vecteur de dimension n. La méthode cherche alors l'hyperplan qui sépare les exemples positifs des exemples négatifs, en garantissant que la marge entre le plus proche des positifs et des négatifs soit maximale. Intuitivement, cela garantit un bon niveau de généralisation car de nouveaux exemples pourront ne pas être trop similaires à ceux utilisés pour trouver l'hyperplan mais être tout de même situés franchement d'un côté ou l'autre de la frontière. L'efficacité des SVM est supérieure à celle de toutes les autres méthodes sur la classification de textes. Son efficacité est aussi très bonne pour la reconnaissance de formes. Un autre intérêt est la sélection de Vecteurs Supports qui représentent les vecteurs discriminant grâce auxquels est déterminé l'hyperplan. Les exemples utilisés lors de la recherche de l'hyperplan ne sont alors plus utiles et seuls ces vecteurs supports sont utilisés pour classer un nouveau cas. Cela en fait une méthode très rapide [Site4, 2008].

3.1.1.6. Programmation génétique :

C'est une méthode générale qui peut être utilisée après n'importe qu'elle méthode précédente, par exemple avec les arbres de décisions. En entrée, un algorithme génétique reçoit une population de classifieurs non optimaux. Le but du programme génétique est de produire un classifieur plus optimal que chacun de ceux de la population d'origine. D'une façon simple, cela consiste à extraire les meilleures parties de chaque classifieur d'origine et de les mettre ensemble pour produire un nouveau classifieur. Cela suppose de pouvoir comparer l'efficacité d'un classifieur. Un résultat important de la méthode est qu'après chaque itération on obtient un classifieur meilleur qu'avant. On peut donc arrêter les itérations à tout moment, même si le résultat n'est pas l'optimum [Site4, 2008].

3.2. Classification non supervisée :

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

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

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

3.2.1. Quelques algorithmes de classification non supervisée :

- Algorithme CURE (Clustering Using REpresentatives):

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

- Algorithme BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies):

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

- Algorithme de ROCK (RObust Clustering using linKs):

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

Deux individus sont voisins si leur similarité dépasse un seuil Θ : sim (p, q) > Θ

La similarité entre une paire d'individus peut être mesurée en se basant sur les distances métriques, le coefficient de Jaccard ou n'importe quelle fonction de similarité non métrique.

Le nombre de liens entre une paire d'individus est le nombre de leurs voisins communs. En général, les individus se trouvent dans une classe ont le nombreux voisins, par conséquent, de nombreux liens. En se basant sur le concept liens (links) entre deux individus, on définit le lien ou l'interconnexion entre deux classes par le nombre de liens croisés (cross links) entre eux. Ce nombre est calculé par la somme de liens entre tous les points situés dans deux classes.

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

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

3.3. Critères d'agrégation :

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

Les critères les plus classiques sont le plus proche voisin, le diamètre maximum, la distance moyenne et la distance entre les centres de gravités [Site3, 2008]

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

- **Le critère du saut minimal :**

La distance entre deux classes C_1 et C_2 est définie par la plus courte distance séparant un individu de C_1 et un individu de C_2: $D(C1, C2) = \min(\{d(x, y)\}), x \in C_1, y \in C_2)$.

- **Le critère de saut maximal :**

La distance entre deux classes C_1 et C_2 est définie par la plus grande distance séparant un individu de C_1 et un individu de C_2. $D(C1, C2) = \max(\{d(x, y)\}), x \in C_1, y \in C_2)$.

- **Le critère de la moyenne :**

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

$$D(C1, C2) = \frac{1}{nc_1 \times nc_2} \sum_{x \in C1} \sum_{y \in C2} d(x, y)$$

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

- **Le critère de Ward :**

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

$$D(C_1, C_2) = \frac{nc_1 \times nc_2}{nc_1 + nc_2} d^2(gc_1, gc_2)$$

Avec :
- gc_1 : Le centre de gravité de C_1
- gc_2 : Le centre de gravité de C_2

- **Le critère des centres de gravité :**

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

$$D(C_1, C_2) = d(gc_1, gc_2)$$

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

4. Classification d'opinions :

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

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

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

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

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

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

A. Extraction des informations à partir des textes :

Il y a deux techniques principales :

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

B. Identification des synonymes :

La tâche d'extraction des caractéristiques des produits peut extraire un nombre important des noms de caractéristiques, et beaucoup d'entre eux sont synonymes, par exemples les mots en anglais « picture » et « image », dans des critiques sur une « caméra » ont le même sens et doivent être regroupés ensemble. Pour construire un résumé basé sur la description des caractéristiques des produits, il est crucial d'identifier et grouper les synonymes ensemble. Les méthodes existantes en extraction des synonymes et similarité lexicale se basent sur l'utilisation des dictionnaires, et le calcul de l'information mutuelle des paires des mots [Turney, 2002].

4.1. Travaux de recherches :

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

Dans [Généreux et al, 2009], et dans les expériences montrées, les auteurs ont classifiés les opinions en deux grandes classes, positives et négatives. Pour faire cette classification, l'approche a été basée sur un classifieur binaire, reposant sur un apprentissage à partir de documents représentatifs préalablement annotés. Le classifieur essaie de repérer automatiquement les éléments pertinents pour une prise de décision (texte véhiculant une opinion positive *vs* texte véhiculant une opinion négative). Ils ont appuyés sur une machine à vecteur support (SVM) binaire entraînée sur des données typiques pour la tâche. Le but de cette étape est de outre la reconnaissance des opinions en elles-mêmes et de permettre ensuite le regroupement des phrases extraites en fonction de l'opinion exprimée et de tenir compte de la proportion d'opinions allant dans un sens ou dans l'autre, afin d'améliorer le rendu et la lisibilité du résumé.

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

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

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

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

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

A. Classification basée sur un apprentissage supervisé :

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

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

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

B. Classification basée sur un apprentissage non supervisée :

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

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

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

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

4.2. Méthodes de classification :

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

4.2.1. Méthode symbolique :

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

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

4.2.2. Méthode statistique :

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

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

4.2.3. Méthode hybride :

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

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

5. Travaux connexe sur la langue arabe :

Au cours des dernières années, l'analyse du sentiment a suscité une attention considérable et ses applications se sont répandues dans presque tous les domaines possibles. De nombreux systèmes et applications sont conçus pour l'anglais et d'autres langues indo-européennes. Cependant, peu d'études se sont concentrées sur la langue arabe.

Nous présentons dans cette section, quelques travaux en classification d'opinions des contextes arabe :

[Abdul Majeed et Diab, 2011] ont présenté un corpus manuellement annoté récemment développé de Modern Standard Arabic (MSA) avec un nouveau lexique de polarité. Ils ont mené leurs expériences sur trois paramètres de prétraitement différents basés sur le texte tokenisé de la Penn Arabic Treebank (PATB). Ils ont adopté une approche de classification en deux étapes, dans la première étape, ils ont construit un classificateur binaire pour trier l'objectif à partir de cas subjectifs. Pour la deuxième étape, ils ont appliqué une classification binaire qui distingue les cas positifs et négatifs.

[El-Halees, 2011] a présenté une approche combinée qui extrait les opinions des documents arabes. Il a utilisé une approche combinée qui se compose de trois méthodes: d'abord, la méthode basé sur le lexique, qui classe certains documents. Deuxièmement, les documents classifiés sont utilisés comme un ensemble de formation pour le modèle Maximum Entropy, finalement, le classificateur le plus proche voisin (ang. the K-nearest neighbor) est utilisé pour classer le reste des documents.

[Rushdi-Saleh et al., 2011] ont présenté un Corpus d'opinion pour l'arabe (OCA). Il est composé d'évaluations arabes extraites de pages Web spécialisées liées aux films utilisant

la langue arabe. Ils ont utilisé deux classificateurs : Support Vector Machine et Naïve Bayes.

[Al-Subaihin et al., 2011] ont proposé un outil d'analyse des sentiments pour l'arabe moderne utilisant l'informatique humaine. Cet outil aidera à construire, à développer et à maintenir dynamiquement les lexiques de l'outil. Ils ont également inspecté le problème de l'analyse des sentiments sur le texte arabe dans le World Wide Web. La solution du problème qu'ils proposent est une approche basé sur le lexique (ang. a lexicon based approach).

[Lazhar, 2014] dans a proposé une approche ONTOMART basée sur une exploration ontologique pour identifier les opinions exprimées dans les textes arabes, Elle se compose de trois modules principaux : extraction les cibles de passage d'opinions explicites et implicites à partir des textes évaluatifs, extraction des expressions d'opinion en se basant sur le lexique de sentiments, puis l'association des caractéristiques identifiées avec les expressions d'opinions correspondantes en formant une liste de couples caractéristique-opinion. La tâche de classification d'opinions en positive et négative est guidée par les Machines à Vecteurs de Support (SVM) comme une technique d'apprentissage supervisé.

6. Conclusion :

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

Contrairement à la classification thématique des documents, et quel que soit la méthode utilisée, la classification d'opinions classifie les opinions analysées en trois grandes catégories : positive, négative et (probablement) neutre.

CHAPITRE III :
L’approche proposée

1. Introduction :

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

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

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

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

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

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

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

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

Nous présentons dans ce chapitre, notre approche proposée pour la classification d'opinions, et l'architecture proposée pour notre système, en détaillant chaque processus à part.

2. Description :

Les travaux de recherche, qui ont travaillé sur l'analyse des sentiments ont approuvé une amélioration remarquable de la performance de leurs systèmes d'extraction d'opinion. Nous croyons que l'utilisation adéquate des traits linguistiques peut aider à améliorer à la fois le processus de classification d'opinion.

Notre approche est la définition d'une méthode de construction d'une base lexicale arabe à partir des tweets du domaine étudié, permettant d'appréhender de gros volumes de données qui seront utilisées dans la phase de classification d'opinions à l'aide d'une ressource externe qui comporte des connaissances. Ces connaissances sont des mots ou des expressions à caractère purement subjectif qui reflètent des sentiments, des opinions, des émotions ou des états psychologiques.

La ressource comporte des milliers de mots de sentiments, des émotions et des états psychologiques, tels que des noms comme بح (amour) ou فوخ (peur), des verbes comme يآب (aimer), بعرِ (effrayer), et des adjectifs comme مTآب (amoureux) ou Ωسح (jaloux). Dans ce contexte nous proposons une classification dans laquelle ces mots sont en plusieurs classes homogènes, chaque classe est nommée par le sentiment ou l'état psychologique décrit, comme la classe فوخ (peur), qui contient les mots relatifs à un sentiment de peur (فوخ - peur,ﺵشﺥ - crainte, رعذ - frayeur, بعرِ - effrayer, بعرم - effrayant, etc.). Nous distinguons deux catégories de mots :

a) Les mots de polarité négative qui décrivent un sentiment ou un état psychologique plutôt désagréable, comme la peur ou la colère.
b) Les mots de polarité positive qui décrivent un sentiment ou un état psychologique plutôt agréable, comme l'amour ou l'amusement.

La conception de l'approche proposé permettant l'analyse d'opinions à partir des tweets collectés dans la première phase de notre travail et les prétraitements que nous avons appliqués dans la deuxième phase, puis la classification des opinions prétraités en positive ou négative, tout en utilisant des techniques symboliques pour la fouille de la subjectivité et la ressource externe.

3. Notre système de classifier une opinion :

Dans cette section, nous allons présenter les différentes étapes que nous avons suivi pour concevoir notre système proposé, en expliquant le détail de chaque phase, en commençant par la collecte des opinions jusqu'à la classification des opinions prétraités.

3.1. Collecte des opinions :

C'est à partir de Twitter que nous avons récupéré les opinions, en utilisant Streaming APIs[1] , à l'aide du programme de Python (Voir Chapitre IV, 3.1), cette récupération de donnés de Twitter est sous forme des tweets composé de Username (détenteur d'opinion), heure et date de création et tweet lui-même, le tous sont stockés dans une base de données (Voir Annexe 1).

3.1.1. Normalisation des textes :

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

• Suppression des caractères spéciaux et les chiffres ;

• Remplacement de ﺇ, ﺃ et ﺁ avec ﺍ;

• Remplacement de la lettre finale ﻯ avec ﻱ;

• Remplacement de la lettre finale ﺓ avec ﻩ;

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

3.1.2. Encodage uniques des textes :

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

Tous les textes de notre corpus sont représentés avec un encodage UTF-8[2].

3.2. Architecture générale :

La mise en œuvre fonctionnelle de notre approche repose sur plusieurs modules pour classifier les opinions.

Les modules de notre approche sont représentés en série, Nous décrivons dans ici toutes les tâches de chaque module dont l'approche est composée et les différentes interactions entre les constituants de chaque module.

[1] Les API en streaming donnent aux développeurs un faible accès latence au flux global de données Tweet de Twitter.

[2] UTF-8 (UCS Transformation Format 8 bits) est un format de codage de caractères défini pour les caractères Unicode (UCS).

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

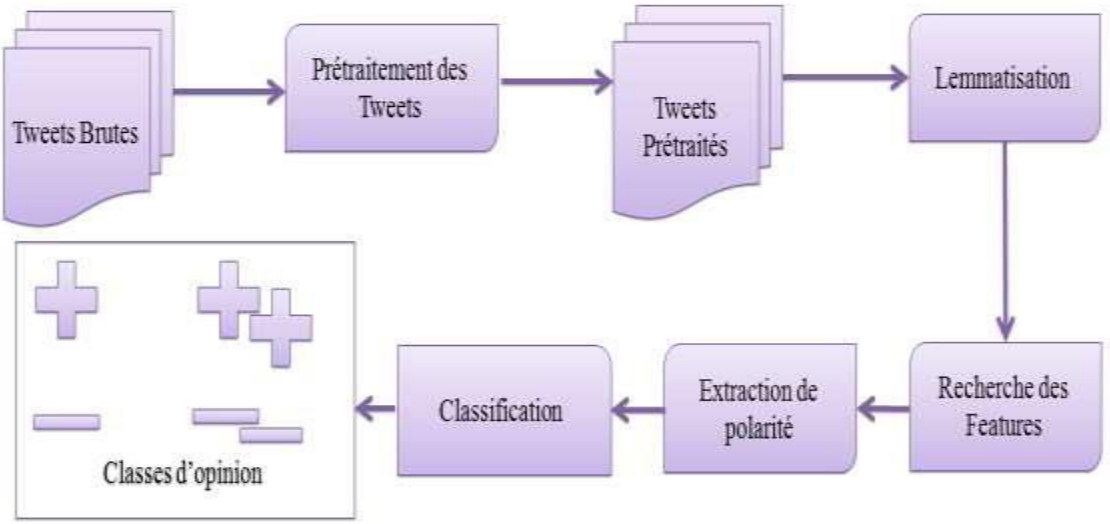


FIGURE 3.1 : *Architecture générale de notre système proposé*

Les modules suivants entre dans la composition de notre système :

Récupération des tweets ;
Prétraitement des tweets ;
Recherche des Features ;
Extraire des polarités ;
Classification.

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

3.2.1. Prétraitement des données :

Le prétraitement des textes nécessite des ressources externes comme des langages de programmation pour la phase de normalisation, des déclencheurs pour la suppression des répétitions, des expressions régulières pour la suppression des expressions supplémentaires …etc.

Nous détaillons dans ce paragraphe les tâches du module de prétraitement représenté par la figure suivante :

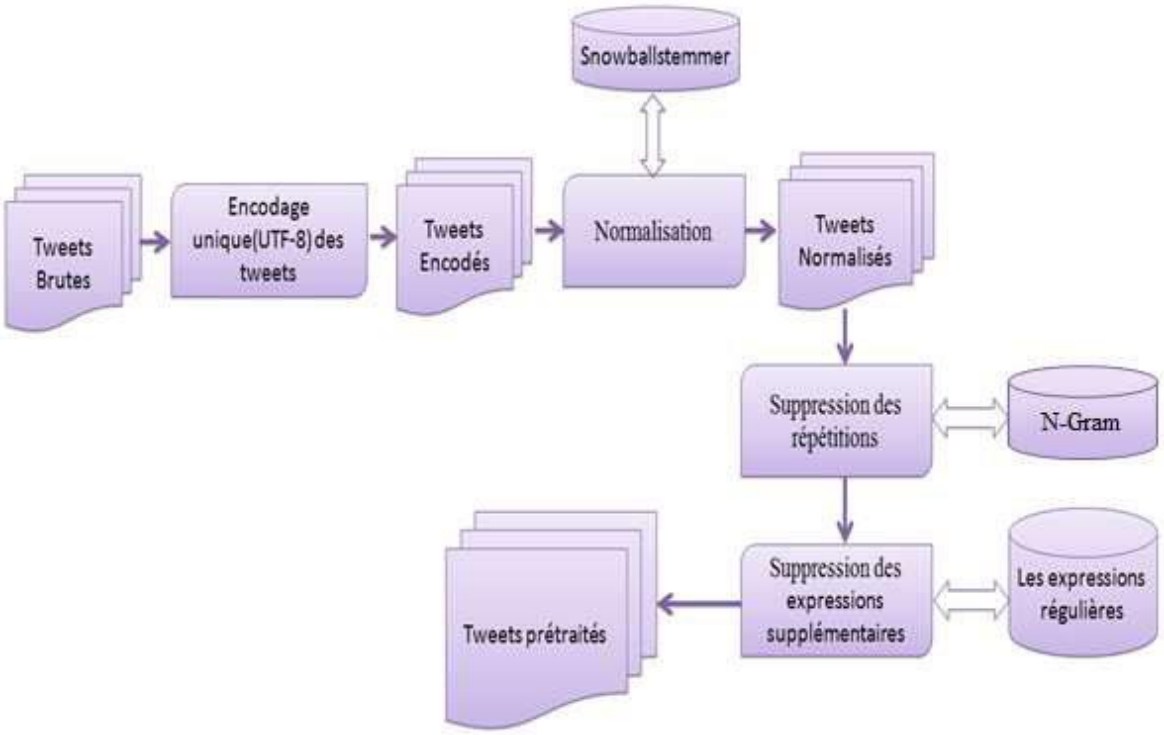


FIGURE 3.2 : *Processus de prétraitement proposé*

3.2.1.1. Suppression de la répétition :

Nous entendons de la répétition les tweets répéter ou retweeté par les utilisateurs.

Consiste à éliminer tous les tweets répéter. Pour chaque tweet reconnu, nous le détectons en utilisant N-Gram qui fait la comparaison entre deux tweets en utilisant cette formule :

$$= \frac{' *}{\sum \; ? + \sum} '$$

Où :

- ? Et ′: Les tweets qui on va comparer
- ∑ ? ET ∑ ′: La somme de nombre de caractères de ? et ′
- : Le pourcentage de la similarité entre ? et ′
- : Le pourcentage de la différentiation entre ? et ′

Nous comparons ? avec chaque tweets de Base de donné non traités, si nous trouvons un tweet similaire ($> 7?\%$) nous le n'ajoutons pas.

3.2.1.2. Suppression des expressions supplémentaires :

Les expressions supplémentaires distinguent les mots non significatifs par exemple : l'URL, username, les Hashtags.

Consiste à supprimer tous les mots non significatifs .Pour chaque mot reconnu, nous le vérifier avec l'utilisation des expressions régulières (Voir Annexe 1), Si un mot en fait partie, il ne sera éliminé. Par exemple le tweet suivant :

''الان#http://ara.tv/6wkbk ايسورو qعلى ارنإ ةديدج Elأ يكريم رفي ابوئع Yسلجم ويشلا''

Apres la suppression de ces expressions supplémentaires (les mots soulignent) nous obtenons comme suit :

''ويشلا سلجم Y ابوئع رفي يكريم Elأ ةديدج ارنإ على qايسورو''

4. Traitement des données :

Le traitement des données nécessite une étude approfondie, nous se limitons dans notre travail à quelques techniques telles que la radicalisation et classifications selon la polarité.

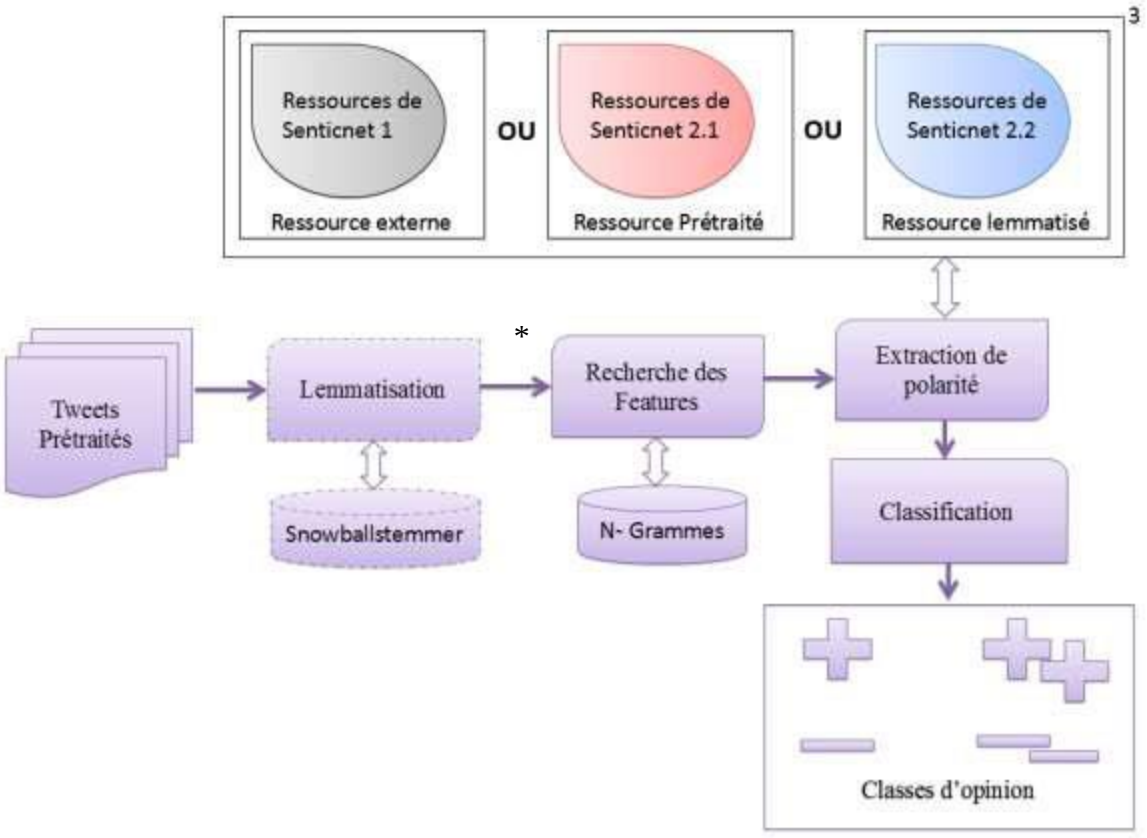


FIGURE 3.3 : *Processus de traitement proposé*

* La lemmatisation des tweets prétraités sera appliqué lorsque d'utilisation de ressource Senticnet 2.2 ;ressource lewŵatisé pour l'extractioŶ de polarité.

[3] Voir 4.3

4.1. Lemmatisation (radicalisation) :

La lemmatisation (Voir Annexe 1) est L'extractioŶ du radicale d'uŶ ŵot ŶousélimiŶoŶs les suffixes et les préfixes

Dans notre approche nous avons utilisé le dictionnaire Snowballstemmer pour extraire les racines des mots ou les lemmes (Qui utilise les dictionnaires).

4.2. Recherche des Features :

A l'aide de N-gramme (Voir Annexe 1) nous détectons les Features dans chaque tweet de tweets prétraités pour extraire la polarité de ces Features et classifier le tweet.

L'utilisation de n-gramme pour la recherche de features est causé par l'existence des concepts sous forme d'un mot composé dans la ressource externe Senticnet, si on trouve un mot composé dans un tweet et il se trouve dans la ressource Senticnet alors on prend sa polarité lors de l'extraction de polarité et nous le ajoutons à la somme des polarités après nous le supprimons pour éviter la répétition de Feature. Nous illustrons cette explication par la figure suivante :

```
La phrase est : إن هذه المشكلة هي مشكلة صعبة
Le stemming phrase est : ان هذه مشكل هي مشكل صعب

Debut de programme de NGramme :

Le concepte detecté est : مشكل صعب et la polarity de ce concept est : -0.934
Le concepte detecté est : مشكل et la polarity de ce concept est : -0.74
Fin..

la polarity globale est : -0.837
```

FIGURE 3.4 : *Résultat d'utilisation de n-gramme*

4.3. Extraction de polarité :

Nous avons mentionné précédemment l'utilisation de ressource de SenticNet (Voir Annexe 1) en trois type ressource externe, ressource traités et ressource lemmatisé pour l'extraction de polarité de Features détectés et avant de traitée la ressource externe nous avons éliminé les mots vides existantes de cette ressource, On appelle mots-vides les termes qui ne sont pas indexés, soit parce qu'ils ne sont pratiquement pas porteurs d'information, soit parce qu'ils sont d'un usage si courant qu'on est certain de les retrouver dans tous les documents(exemple : …في، على، الى), soit parce qu'ils

appartiennent au langage argotique ou familier (exemple : ouais). L'exemple ci-dessous contient quelque mot vide dans la langue arabe :

{ أ نظلأ فـلـأ اماأ امأإ أ امأ سمأ سمأ دمأ كمامأq متزأ تـزأ أبزأ لانلأ نزنزأ ا... }

Nous allons expliquer chaque ressource et le traitement applique :

- Pour la première ressource est une ressource externe de SenticNet sans aucun traitement.

```
senticnet["أصبح_متحمس"] = [0.707, -0.833, 0, 0, "#فرحة",
"#مفاجأة", 0.513, "ضحك", "قهقه", "سعد", "تريد_الانضمام_إلى", "تصبح_سعيدة"]
senticnet["أصبح_مناسباً"] = [0.071, 0, 0, 0, "#فرحة", "#فرحة"]
senticnet["أصبح_مرهق"] = [-0.098, -0.041, 0.013, -0.118,
"#حزن", "#قرف", 0.063, "استنفاد","ضيق_التنفس", "تعب", "تشنج_الساق"]
senticnet["أصبحت_حزينة"] = [-0.993, 0, 0.673, -0.928, "#حزن", "#قرف", -0.865, "تعس",
"الملل", "افتقد_شخص", "فقدان_الوظيفة"]
```

FIGURE 3.6 : *présentation des concepts de ressource Senticnet*

- Pour la deuxième ressource est une ressource externe traitée ou nous ajoutons les synonymes et les Antonyme des concepts de ressource externe par l'importation depuis un dictionnaire arabe virtuelle[4].

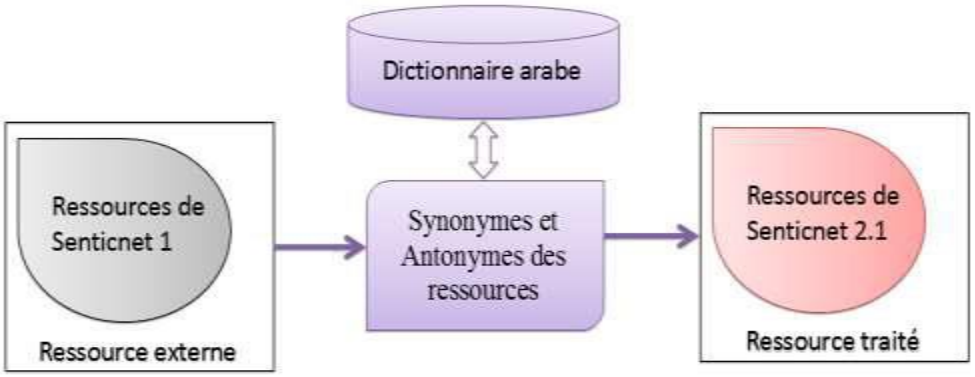


FIGURE 3.7 : *présentation de traitement de ressource*

La ressource de Senticnet 1 il existés 25187 concepts après le l'ajout des synonymes et l'antonyme nous avons obtenu 42966 concepts.

[4] Le site utilisé c'est http://www.almaany.com/

Remarque :

Il y a un problème des mots complexe (mot + suffixe ou préfixe) quand nous travaillons avec Senticnet 2.1 par exemple 'لاآب' existe dans la ressource mais le mot 'بح' n'existe pas malgré ce sont le même mot, pour cela nous avons construire la ressource Senticnet 2.2.

- Pour la dernière ressource est une ressource traitée lemmatisé c'est-à-dire après l'ajout des synonymes et les antonymes nous faisons la lemmatisation de ces nouveaux concepts.

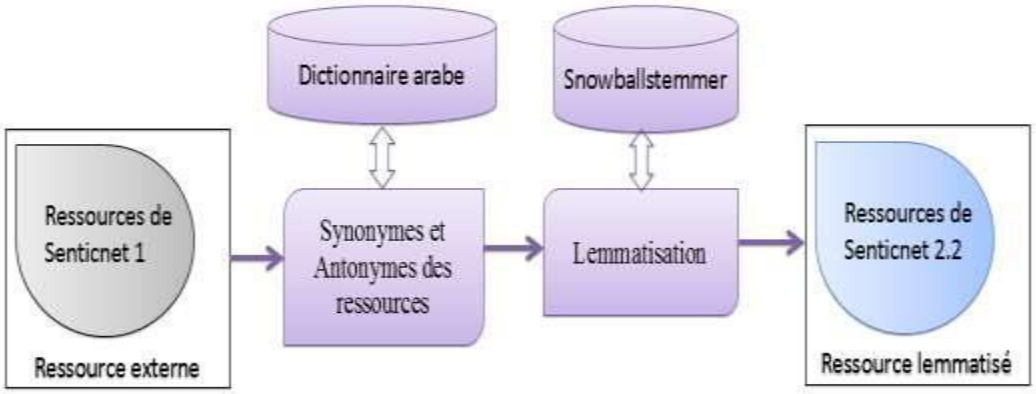


FIGURE 3.8 : *présentation de lemmatisation de ressource*

La ressource de Senticnet 1 il existés 25187 concepts après le l'ajout des synonymes et l'antonyme et lemmatisation nous avons obtenu 28404 concepts.

Remarque :

Nous rencontrons le problème d'ambiguïté dans le ressource Senticnet 2.2 par exemple le mot 'حطس' et le mot 'طسآبلا' ont des polarités différentes (0 .724 et -0.733 successivement) mais ont le même lemme 'حطس'

A l'aide de ses 3 ressources de Senticnet nous extrayons la polarité des Features détecté dans la phase de la recherche des Features pour pouvoir classifier les tweets en catégories selon notre approche proposée.

4.4. Classification d'opinion :

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

Dans la classification thématique des textes, les marqueurs linguistiques jouent un rôle très important pour déterminer le thème, tandis qu'en classification d'opinions ces marqueurs ne jouent aucun rôle, cependant, les mots qui expriment les polarités des opinions sont importants.

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

(1) Positive forte ((ديج
2)) Positivé faible (نسج)

(3) Négative forte (عيس)
(4) Négative faible ((ادج عيس

5)) Neutre (يدايج)

5. Conclusion :

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

CHAPITRE IV :

Implémentation et Evaluation

1. Introduction :

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

2. Constitution de la base de données :

La base de données c'est le modèle pour représenter les opinions, cette représentation basée sur l'utilisation du langage SQL (Voir Annexe 2) pour stocker les tweets qui seront utilisées dans des étapes ultérieures de notre travail. L'utilité d'élaborer ce modèle, est pour faciliter la représentation des opinions extraites, et de séparer les différents constituants d'une opinion, en vue de faciliter l'identification de l'orientation sémantique de l'opinion.

Nous avons créés deux bases de donnée pour stockes les tweets télécharger ou :

- La première base ''France'' stocke 6895 tweets pour le domaine de ∆ا∆ر∆ئ∆س∆ ∆E∆ا∆ا∆ا∆ا∆ا∆ت∆ا∆" ∆ا∆ر∆ز∆ن∆س∆ " et contient deux tables une table ''twtfrance'' représente les tweets téléchargé et la deuxième table ''twtfrancetraite'' représente les tweets prétraités.
- La deuxième base ''galaxy'' stocke 1370 tweets pour le domaine de ∆ه∆ا∆ـ∆ف∆ ∆س∆ا∆م∆س∆ا∆و∆غ∆" "+8∆س∆ا8 ∆س∆ا∆و et contient deux tables une table ''twtgalaxy'' représente les tweets téléchargé et la deuxième table ''twtgalaxytraite'' représente les tweets prétraités.
- Table brute : la table qui contient les tweets brutes sans aucun prétraitement.

  Exemple de tweets brute :

  ∆ا∆ق∆ل∆م ∆ر∆م∆أ ∆ا∆ا∆ف∆ ∆ا∆ر∆ب∆ا∆ ∆ا∆ل∆ ∆ر∆ئ∆ ∆ح∆س∆ ∆ا∆ى∆ ∆ع∆ ∆غ∆أ ∆ر∆ئ∆ **∆ا∆ا∆ن∆ي∆ ∆ا∆م∆ ∆ك∆ر∆ا∆و∆** ∆ا∆ر∆ن∆ز∆س∆ ∆ا∆و∆ ∆ك∆و∆ب∆ه∆: ∆ح∆ ∆ص∆ ∆و∆ ∆ل∆ ∆ا∆ر∆ئ∆ ∆ي∆س∆ ∆ا∆ ∆ف∆ن∆ر∆س∆ ∆ا∆ ∆ي∆ ∆ا∆ج

  ∆ا∆ر∆ض∆ ∆ا∆ع∆ ∆ا∆ا∆ ∆ا∆ر∆ق∆ ∆ا∆و∆ ∆ا∆ل∆ ∆ا∆ب∆ ∆ق∆ ∆ا∆ي∆غ∆ ∆ر∆ Ω ∆ا∆ا∆ن∆ي∆ ∆و∆ @France24_ar

- Table prétraités : la table qui contient les tweets prétraités dans ce dernier est résultat de l'application des diffèrent prétraitement sur les tweets brutes qui sont stockés dans la table brute.

  Exemple de tweet prétraité :

  ∆ا∆ق∆ل∆م ∆ر∆م∆أ ∆ا∆ا∆ف∆ ∆ا∆ر∆ب∆ا∆ ∆ا∆ل∆ ∆ر∆ئ∆ ∆ح∆س∆ ∆ا∆ى∆ ∆ع∆ ∆غ∆أ ∆ر∆ئ∆ **∆ا∆ا∆ن∆ي∆ ∆ا∆م∆ ∆ك∆ر∆ا∆و∆** ∆ا∆ر∆ن∆ز∆س∆ ∆ا∆و∆ ∆ك∆و∆ب∆ه∆: ∆ح∆ ∆ص∆ ∆و∆ ∆ل∆ ∆ا∆ر∆ئ∆ ∆ي∆س∆ ∆ا∆ ∆ف∆ن∆ر∆س∆ ∆ا∆ ∆ي∆ ∆ا∆ج

  ∆ا∆ر∆ض∆ ∆ا∆ع∆ ∆ا∆ا∆ ∆ا∆ر∆ق∆ ∆ا∆و∆ ∆ا∆ل∆ ∆ا∆ب∆ ∆ق∆ ∆ا∆ي∆غ∆ ∆ر∆ Ω ∆ا∆ا∆ن∆ي∆ ∆و∆.

3. Environnement de développement :

Nous présentons dans cette section, le langage de programmation Python utilisé, et l'environnement de développement Html et feuille de style en cascade (Voir Annexe 2).

3.1. Python :

Python[1] est un puissant langage de programmation informatique moderne. Il porte des similitudes avec Fortran[2], l'un des premiers langages de programmation, mais il est beaucoup plus puissant que le Fortran.

Il permet d'utiliser des variables sans les déclarer (c.-à-d., il détermine implicitement les types), Vous n'êtes pas obligé de définir des classes en Python (contrairement à Java) mais vous êtes libre de le faire lorsque cela vous convient.

Il a été développé par Guido van Rossum, et c'est un logiciel gratuit. Mais Python est également gratuit d'autres manières importantes, par exemple, vous êtes libre de le copier autant de fois que vous le souhaitez, et d'étudier librement le code source, et de faire des modifications [Doty, 2008].

Python est un langage de programmation facile à apprendre et puissant. Il dispose de structures de données efficaces de haut niveau et d'une approche simple mais efficace de la programmation orientée objet. La syntaxe élégante de Python et le typage dynamique, ainsi que sa nature interprétée, en font un langage idéal pour le script et le développement rapide d'applications dans de nombreux domaines sur la plupart des plates-formes [Site5, 2001].

3.1.1. Flask :

Flask[3] est un framework open-source de développement web en Python, il est basé sur *Werkzeug*[4] pour la partie serveur / WSGI[5] et sur *Jinja 2* (Voir 3.1.2) pour le *templating*. Puisqu'il s'appuie sur *Werkzeug*, il est en mesure de fonctionner de façon autonome, sans l'aide d'un serveur *WSGI* tiers. Son but principal est d'être léger, afin de garder la souplesse de la programmation Python, associé à un système de templates. Il est distribué sous licence BSD[6].

3.1.2. Les modules utilisés :

Pour la collecte des données :

Tweepy :

[1] https://www.python.org/.
[2] Fortran (*FORmula TRANslator*) est un langage de programmation utilisé principalement pour le calcul scientifique.

[3] http://flask.pocoo.org/.
[4] Werkzeug est une bibliothèque d'utilité WSGI pour Python. Il est largement utilisé et autorisé par BSD.

[5] WSGI (Web Server Gateway Interface) est une spécification qui définit une interface entre des serveurs et des applications web pour le langage Python.

[6] Licence BSD (Berkeley Software Distribution License) est une licence libre utilisée pour la distribution de logiciels.

Tweepy prend en charge l'accès à Twitter via l'authentification de base et la nouvelle méthode, OAuth. Twitter a cessé d'accepter l'authentification de base, donc OAuth est maintenant la seule façon d'utiliser l'API de Twitter [Site6, 2011].

- OAuth un outil que Twitter utilise pour fournir un accès autorisé à son API.

JSON :

JavaScript Object Notation – Notation Objet issue de JavaScript est un format léger d'échange de données. Il est facile à lire ou à écrire pour des humains. Il est aisément analysable ou générale par des machines. Il est basé sur un sous-ensemble du langage de programmation JavaScript [Site7, 2005].

Pandas :

Pandas est une bibliothèque sous licence BSD ouverte, fournissant des structures de données et des outils d'analyse de données performants et faciles à utiliser pour le langage de programmation Python [Site8, 2016].

Snowballstemmer :

Snowball est un langage dans lequel les algorithmes de dérivation peuvent être facilement représentés.

L'algorithme Arabic Light Stemming fait pour Snowball, c'est rapide et peut être généré dans de nombreuses langues de programmation (à travers Snowball).

Pour le stockage des données :

Mysql :

Il répond à une logique client/serveur, c'est à dire que plusieurs clients (ordinateurs distants) peuvent se connecter sur un seul serveur qui héberge les données [Site9, 2003].

Pour le traitement des données :

Jinja2 :

Jinja2[7] est un moteur de templates pour le langage de programmation Python. Il dispose d'un support unicode complet, un environnement d'exécution en boîte en bloc intégré optionnel, largement utilisé et autorisé par BSD.

Render-template :

Retourner un modèle à partir du dossier de modèle avec le contexte donné.

Template_folder le dossier qui contient les modèles qui doivent être utilisés par l'application.

[7] http://jinja.pocoo.org/.

3.2. HTML 5 :

HTML5 est une évolution de la norme HTML (Voir Annexe 2) regroupant un ensemble de technologies : SVG[8], CSS3, WebGL[9], File API, MathML[10]… Le but avoué de cette nouvelle version d'HTML est de faciliter le développement d'interfaces utilisateur riches. HTML5 est beaucoup plus orienté applicatif que ses prédécesseurs et veut permettre de s'affranchir de plugin pour utiliser au maximum les technologies web natives pour construire une application riche. On peut facilement grâce à HTML5 visionné des vidéos sans plugin, ajouter des effets visuels aux textes, images, vidéos… On peut aussi utiliser des polices non-standards, dessiné en SVG, faire de la 3D [Site10, 2011].

4. Présentation de l'application :

Pour l'implémentation de notre approche nous avons développé un site web en Python à l'aide de l'environnement HTML, il est muni d'une interface graphique conçue à l'aide de la bibliothèque Flask et Jinja2, et à l'aide de la bibliothèque Render-template qui est basée sur Jinja2 elle-même.

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

- **Package CollectionCorpus :**

Il contient les 3 classes suivantes :

Classes CollectTweets pour récupérer les tweets de "S8 + و S8 فـتـاه غـروسـمـاس" à partir de twitter et les stockés dans la table brute de la base ''france'' et ''galaxy'' successivement.

La class BDD traitee pour les prétraitements des tweets, la suppression des expressions supplémentaires a laides des expressions régulières et la suppression des répétions à l'aide de trigger.

- **Package NewSenticNet :**

Contient 2 classes :

Classe arabicDict sert a connecté au site de dictionnaire arabe pour la récupération des synonymes et les antonymes des concepts du ressource senticnet1.

[8] SVG est un langage aux possibilités semblables au Flash.

[9] **WebGL** est une spécification d'interface de programmation de 3D dynamique pour les pages et applications HTML5.

[10] **Mathematical Markup Language** est un langage permettant l'affichage de symboles mathématiques sur Internet.

Classe newSenticNet pour éliminer les mots vides et crée les 2 ressource traité de Senticnet.

Plus les 2 ressources traitées :

NSN : (Ressource senticnet2.1) Ressource 1 plus l'ajout des synonymes et antonymes des concepts.

StemNSN : (Resource senticnet2.2) ressource2.1 lemmatisé.

- **Package SentimentAnalysis :**

Contient 3 classes :

Classe SQL permet de faire les opérations sql connecter et déconnecté de la base de données, calculé la polarité et retourné les tweets nécessaires.

Classe tweetsPolarity contient des méthodes qui calcules la polarité.

Classe SentimentAnalysis cette classe a 2 méthodes :

- La première méthode ResultCorpus pour crée un objet de la classe SQL et retourné un dictionnaire des tweets positives et négatives.
- La deuxième méthode ResultSearch crée un objet de tweetsPolarity et retourné un résultat de type réel : polarité.

- **Package WebSite :**

Pour l'affichage des résultats et l'exécution des opérations cités précédemment.

Nous modélisons notre développement avec un diagramme de classe

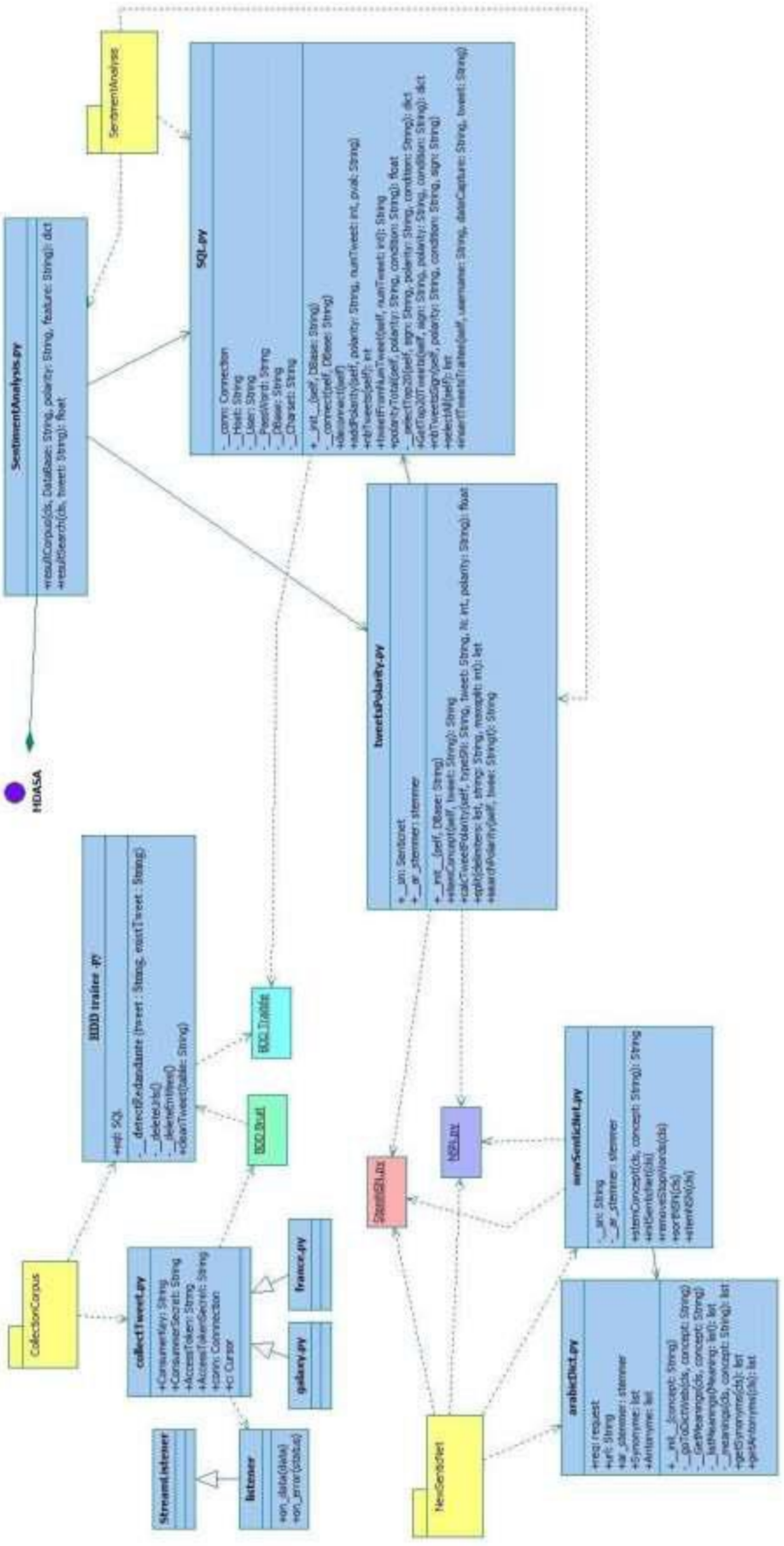


FIGURE 4.1 : *Diagramme de classe*

4.1. Déroulement :

Nous présentons dans cette section les différentes étapes de déroulement de notre système, dès la sélection de tweets jusqu'à la classification illustré par des capture d'écran de site web.

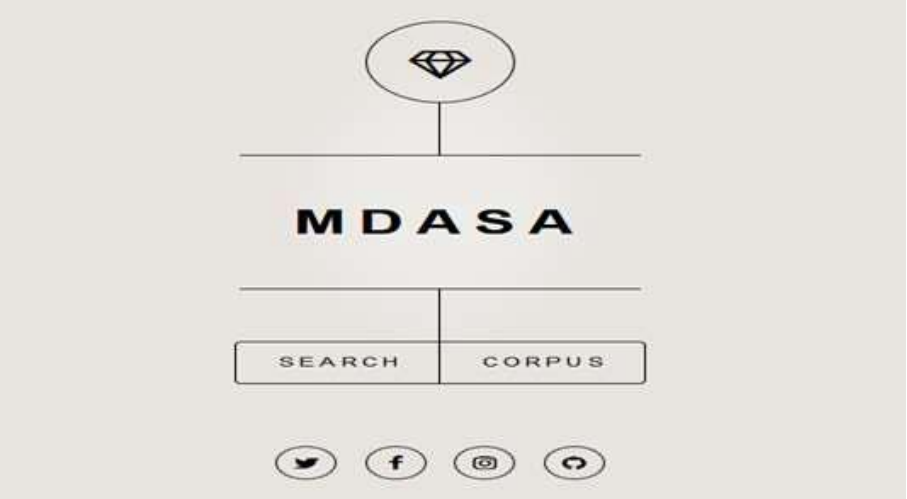


FIGURE 4.2 : *Fenêtre d'accueil de notre site web*

La fenêtre présente deux choix :

SEARCH pour analyse une opinion donner par l'utilisateur.
CORPUS pour chercher dans la base de donnes les opinions de deux domaines étudie.

- **SEARCH**



Figure 4.3 : *Fenêtre de SEARCH*

L'utilisateur écriera son opinion dans la case ''Write your opinion'' et cliquera sur le button ''SEE POLARITY''.



FIGURE 4.4 : *écrire une opinion*



FIGURE 4.5 : *résultat de SAERCH*

- **CORPUS**



FIGURE 4.6 : *fenêtre de CORPUS*

Pour la fenêtre CORPUS nous avons 3 champs :

- Champ CHOOSE A RESOURCE ou l'utilisateur choisira une ressource des 3 ressources avant.
- Champ CHOOSE A DOMAIN ou l'utilisateur choisira un domaine des 2 domaines étudié.
- Champ FEATURE champ optionnel ou l'utilisateur écrira un feature souhaité.

Apres la saisie des champs nécessaires nous cliquons sur SEE CCORPUS pour avoir les résultats.

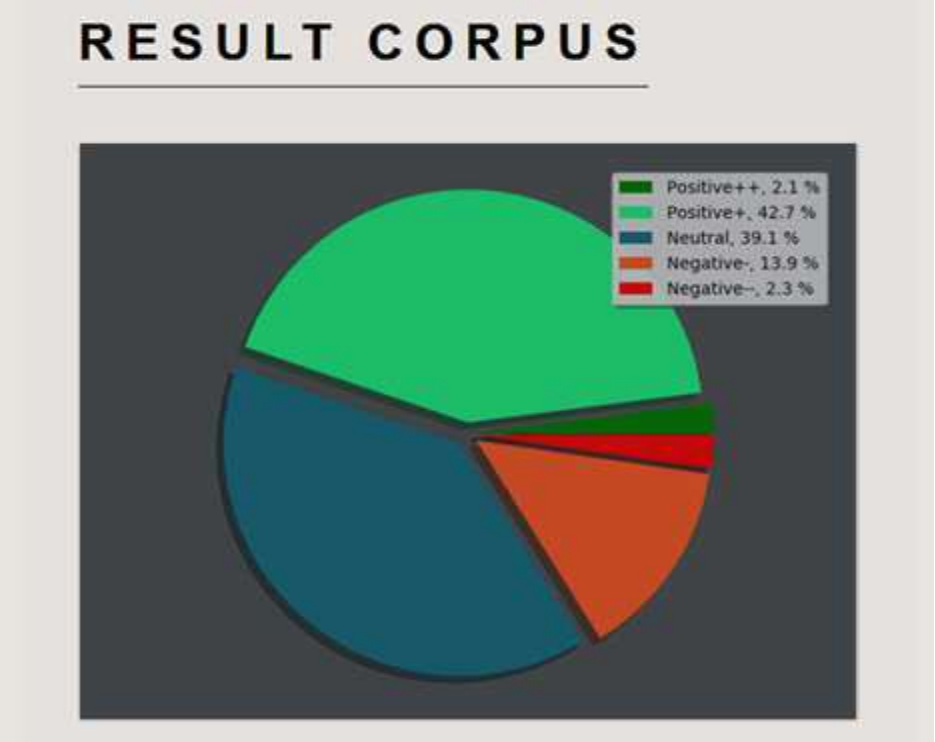


Figure 4.7 : *Résultat de CORPUS*

POLARITY TOTALE :

0.15288856589147265

TOP 20 POSITIVE :

num	Username	Tweet	Date capture	Polarity
1	alatrozy	إيمانويل ماكرون شخصية مصنوعة بمهارة من قبل أباطرة رجال المال	Sun May 07 16:33	0.947
2	TaghridFayad	إيمانويل ماكرون شخصية مصنوعة بمهارة من قبل أباطرة رجال المال	Sun May 07 16:41	0.947
3	bintalwatan1414	خادم الحرمين الشريفين يبحث برقية تهنئة للمخلصة السيد إيمانويل ماكرون بمناسبة انتخابه رئيساً للجمهورية الفرنسية...	Sun May 07 22:27	0.932
4	qy4pc	خادم الحرمين الشريفين يبحث برقية تهنئة للمخلصة السيد إيمانويل ماكرون بمناسبة انتخابه رئيساً للجمهورية الفرنسية.....	Sun May 07 22:28	0.932
5	AliAppleseed	تفتح صناديق الاقتراع بفرنسا أبوابها صباح الأحد للإدلاء بأصواتهم واختيار رئيسهم المقبل بين إيمانويل ماكرون ومنافسته مارين لوبان	Sat May 06 15:59	0.921

Figure 4.8 : *Top 20 opinions positives*

TOP 20 NEGATIVE :

Num	Username	Tweet	Date capture	Polarity
1	wattar93	إيمانويل ماكرون رئيساً لفرنسا ⬅الإنتخابات صارت مثل مايسير بالعالم العربي بالغلط حتى عنا بتصير احسن من هيك بس نحنا مانمن...	Sun May 07 18:04	-0.919
2	ibmbader	مواكبة مستمرة على قناة أزيد الفعل الدولية والداخلية على فوز إيمانويل ماكرون بالانتخابات الرئاسية الفرنسية	Sun May 07 18:49	-0.899
3	Democratic_AC	"حليف لوبان : إيمانويل ماكرون يريد إبقاء فرنسا "داخل سجن الاتحاد الأوروبي	Mon May 01 16:08	-0.871
4	news_ger_chekel	إيمانويل ماكرون رئيساً لفرنسا ⬅الإنتخابات صارت مثل مايسير بالعالم العربي بالغلط حتى عنا بتصير احسن من هيك بس نحنا مانمدح إلا الاجانب	Sun May 07 18:03	-0.826
5	fullife	زمان م إيمانويل ماكرون عازمة الجيران و فاتحة صندوق بيبس	Sun May 07 18:49	-0.822
6	v_Dandy	لأن مارين لوبان عنصرية ، أما ماكرون ضد العنصرية _ فالتزم يز علوا على خسارة الحزب العنصري	Sun May 07 19:00	-0.811

Figure 4.9: *Top 20 opinions négatives*

5. Classification et Évaluation :

5.1. RapidMiner :

RapidMiner[11] est un logiciel open source et gratuit dédié au data mining. Il contient de nombreux outils pour traiter des données : lecture de différents formats d'entrée, préparation et nettoyage des données, statistiques, tous les algorithmes de data mining, évaluation des performances et visualisations diverses. C'est un logiciel puissant, il n'est pas facile à manipuler au premier abord, mais avec un peu de pratique, il permet de mettre en place rapidement une chaîne complète de traitement de données, de la saisie des données à leur classification.

5.2. Classification :

Dans le domaine de fouille d'opinions, 4 mesures de performance sont généralement utilisées pour évaluer la tâche de classification ; La précision, le rappel, la pertinence et F1 (aussi appelé F-score ou F-mesure) :

$$Précision = \frac{TP}{TP + FP}$$

$$Rappel = \frac{TP}{TP + FN}$$

$$Pertinence = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1 = \frac{2 \times Précision \times Rappel}{Précision + Rappel}$$

Ou : **TP** : True Positive, **FN**: False Negative, **FP**: False Positive, **TN**: True Positive

En comparant le résultat du classificateur avec une prédiction humaine, la Précision est la portion des occurrences classées correctement comme positive contre toutes les occurrences prédites comme positives (TP+FP). Le Rappel est la portion des occurrences classées correctement comme positive contre toutes les occurrences attribuées à la classe Positive. Polarité de Senticnet1, polarité de Senticnet2.1 et polarité de Senticnet2.2 sont les trois catégories essentiellement couvertes par notre corpus sur lequel nous avons évalué le classificateur à l'aide de l'outil RapidMiner.

[11] http://www.rapidminer.com

La Pertinence est un taux mesurant la performance du classificateur. F1 combine la Précision et le Rappel, qui est une moyenne harmonique de ces deux derniers.

Pour évaluer notre classificateur Naïve bayes en utilisant l'outil RapidMiner, nous devrions étendre la notion de Précision et de Rappel en définissant des valeurs distinctes pour les 5 classes (Very good, Good, Neutre, Bad et Very bad). Pour cette tâche, nous avons utilisé un corpus d'apprentissage de 500 tweets et 400 textes pour tester le classificateur.

Les étapes évaluées en utilisant RapidMiner :

- Préparation de la base de données (sous forme d'un fichier Excel).
- Application de classificateur Naïve bayes pour chaque teste.

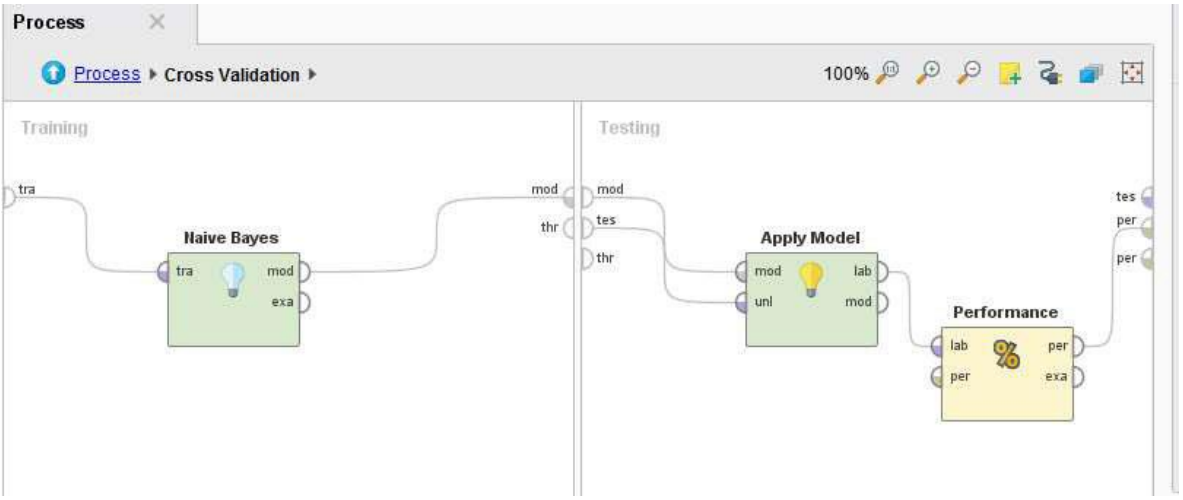


FIGURE 4.10: *design d'application de Naïve bayes*

- Validation de modèle afin de calculer les mesures de performance.

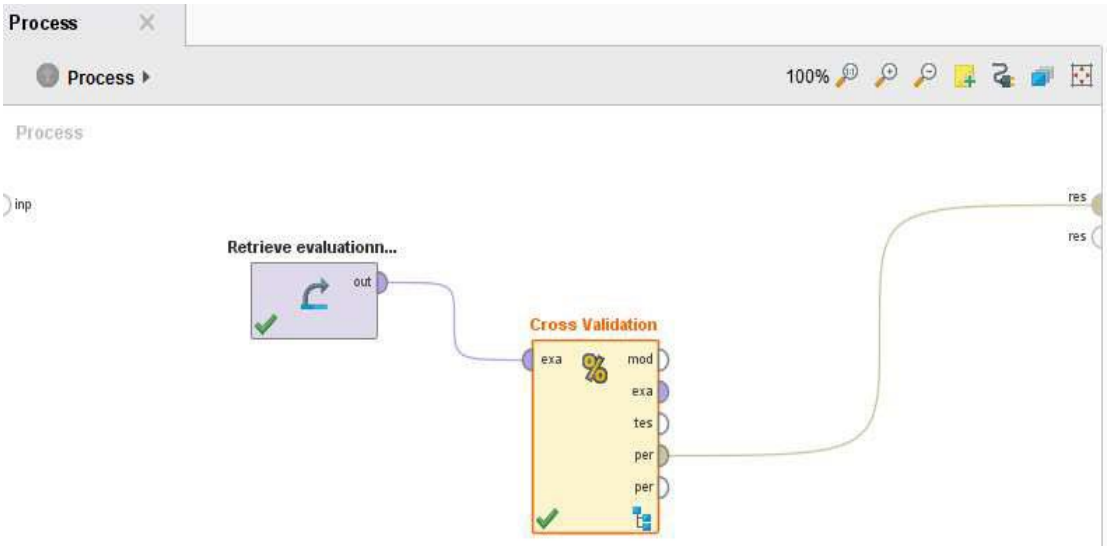


FIGURE 4.11 : *Validation de modèle*

Les résultats du classificateur Naïve bayes sont montrés dans les figures suivantes, où la précision et le rappel sont les valeurs moyennes de précision et de rappel des classes séparées :

- Classification selon Senticnet 1 :

accuracy: 58.52% +/- 3.43% (mikro: 58.52%)

	true Good	true Neutral	true Bad	true Very Good	true Very Bad	class precision
pred. Good	0	0	0	0	0	0.00%
pred. Neutral	49	133	17	13	11	59.64%
pred. Bad	0	4	1	0	1	16.67%
pred. Very Good	0	0	0	0	0	0.00%
pred. Very Bad	0	0	0	0	0	0.00%
class recall	0.00%	97.08%	5.56%	0.00%	0.00%	

FIGURE 4.12 : *Résultats de classification selon Senticnet1*

- Classification selon Senticnet 2.1 :

accuracy: 58.95% +/- 2.83% (mikro: 58.95%)

	true Good	true Neutral	true Bad	true Very Good	true Very Bad	class precision
pred. Good	0	0	0	0	0	0.00%
pred. Neutral	49	135	18	13	12	59.47%
pred. Bad	0	2	0	0	0	0.00%
pred. Very Good	0	0	0	0	0	0.00%
pred. Very Bad	0	0	0	0	0	0.00%
class recall	0.00%	98.54%	0.00%	0.00%	0.00%	

FIGURE 4.13 : *Résultats de classification selon Senticnet2.1*

- Classification selon Senticnet 2.2 :

accuracy: 59.80% +/- 4.01% (mikro: 59.83%)

	true Good	true Neutral	true Bad	true Very Good	true Very Bad	class precision
pred. Good	0	0	0	0	0	0.00%
pred. Neutral	49	133	17	13	8	60.45%
pred. Bad	0	0	0	0	0	0.00%
pred. Very Good	0	3	1	0	0	0.00%
pred. Very Bad	0	1	0	0	4	80.00%
class recall	0.00%	97.08%	0.00%	0.00%	33.33%	

FIGURE 4.14 : *Résultats de classification selon Senticnet2.2*

- Classification générale :

accuracy: 59.39% +/- 9.53% (mikro: 59.39%)

	true Good	true Neutral	true Bad	true Very Good	true Very Bad	class precision
pred. Good	10	7	0	3	1	47.62%
pred. Neutral	39	120	13	9	6	64.17%
pred. Bad	0	7	4	1	3	26.67%
pred. Very Good	0	2	0	0	0	0.00%
pred. Very Bad	0	1	1	0	2	50.00%
class recall	20.41%	87.59%	22.22%	0.00%	16.67%	

FIGURE 4.15 : *Résultats générale de classification*

- Les résultats de classificateurs Naïve bayésien

	Précision	**Rappel**	**Pertinence**	**F-mesure**
Senticnet1	59.64%	97.08%	58.52%	71.6%
Senticnet2.1	59.47%	98.54%	58.95%	74.17%
Senticnet2.2	60.45%	97.08%	59.80%	74.51%
Total	64.17%	87.59%	59.39%	73.89%

TABLEAU 4.1 : *Résultats générale de classification d'opinion*

5.3. Evaluation :

Pour démontrer l'efficacité d'utilisation la lemmatisation de la tâche de classification d'opinion, nous avons effectué des épreuves séparés, le premier épreuve concerne l'utilisation de la ressource brute le deuxième concerne l'utilisation de la ressource traité par l'ajout des synonymes et antonymes qui donne un résultat bon par rapport au première, le troisième épreuve concerne l'utilisation de la ressource lemmatisé qui améliore le résultat malgré la complexité morphosyntaxique de la langue arabe qui rend difficile à lemmatisé les mots complexe.

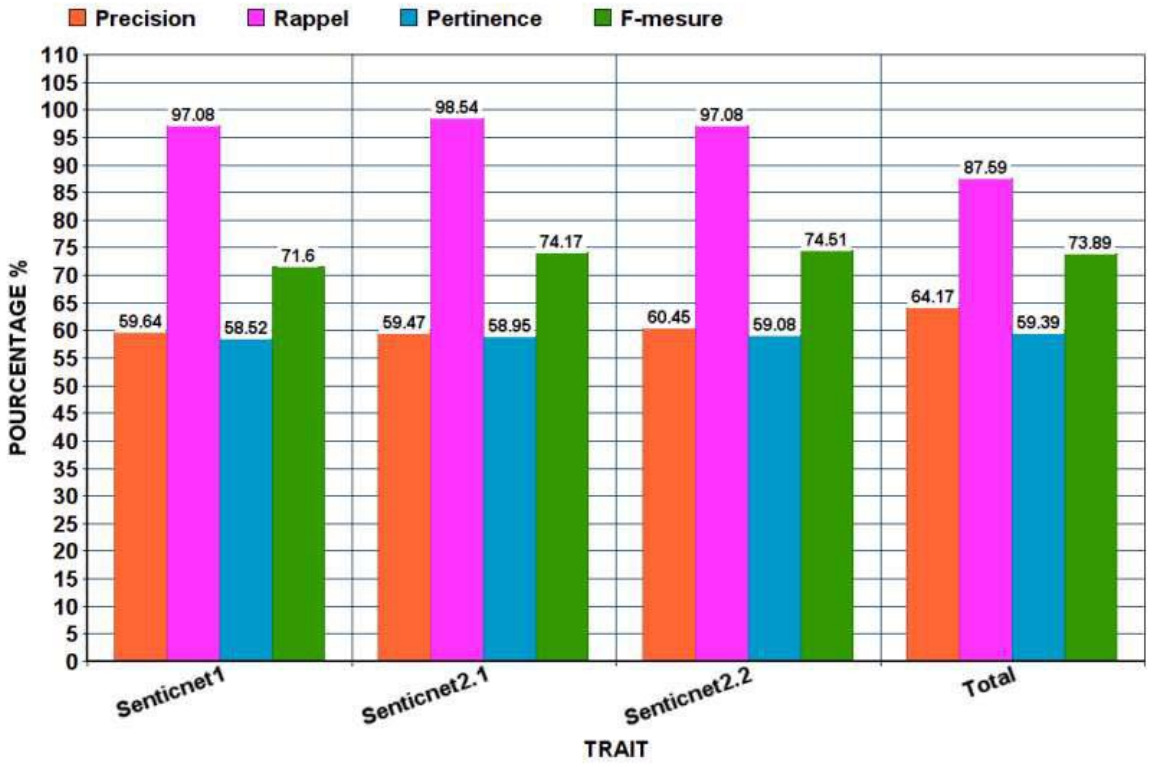


FIGURE 4.16 : *Évaluation de différentes ressources pour la classification d'opinion*

6. Conclusion

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

Les résultats de notre système ont été liés au choix des techniques utilisées et d'attributs d'opinions que nous avons proposé dans notre approche. La bonne préparation des textes (prétraitement), l'utilisation les différentes méthodes, sont des facteurs indispensables, qui ont participé dans l'amélioration de la performance de notre système.

Pendant l'évaluation de notre système, il était possible de prévoir un système plus performant lorsque les ressources linguistiques pour la langue arabe sont offertes d'une façon plus élaborée. C'est le cas de notre ressource de sentiments traité qui ne couvre pas toutes les expressions d'opinion contrairement aux autres lexiques disponibles pour l'anglais comme SentiWordnet qui est plus riche et plus structuré, ainsi que les expressions d'opinion synonymes qui ont la même polarité et intensité sont regroupés..
L'évaluation a également montré que le système peut être plus performant lorsque l'ontologie du domaine et le lexique des sentiments sont plus riches. Un tel système peut donc combiner plusieurs approches pour améliorer sa performance.
La complexité morphologique et syntaxique de certaines expressions arabes a diminué considérablement la performance d'identification des processus d'identification des expressions d'opinion et des caractéristiques.

Conclusion et perspectives

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

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

L'approche que nous avons proposée est basée sur l'extraction des features et leurs polarités à partir des tweets. Ces dernier éléments expriment les points de vue des détenteurs (opinions holders) et leurs orientations sémantiques : positives, négatives et neutres. Pour cette raison nous avons reposés sur une ressource externe Senticnet. Les opinions analysées, sont ensuite classées selon leurs orientations sémantiques en cinq catégories : positive forte, positive faible, négative forte, négative faible et neutre.

Notre proposition a rencontré quelques difficultés dues à la complexité syntaxique de la langue arabe, qui a posé des problèmes au niveau d'analyse des tweets sur laquelle se base notre approche. Ce qui a nous demandé d'appliquer certaines méthodes bien définies. Le prétraitement des textes d'opinions et l'approche pour l'analyse sémantique des opinions ont démontré leur nécessité. La difficulté de détection de feature et la polarité d'opinion est fortement liée à l'absence d'application des techniques d'analyse, qui déborde le cadre de notre travail.

Comme perspectives, nous envisageons d'améliorer la qualité du lexique. L'objectif est de vérifier que la polarité attribuée à chaque mot correspond bien au contexte le plus fréquent dans lequel il est utilisé. Finalement, nous envisageons de traiter le problème d'ambiguïté des mots d'opinion susceptibles d'être affectés à plusieurs classes de polarité. Ceci peut se faire à notre avis à travers l'ajout d'étiquettes de POS dans le lexique et à travers l'attribution de score à chaque mot semblablement au lexique SentiWordnet.

ANNEXE 1

Base de données :

Une base de données informatique est un ensemble de données qui ont été stockées sur un support informatique, et organisées et structurées de manière à pouvoir facilement consulter et modifier leur contenu [Site11].

Trigger :

Les triggers (ou déclencheurs) sont des objets de la base de données. Attachés à une table, ils vont déclencher l'exécution d'une instruction, ou d'un bloc d'instructions, lorsqu'une, ou plusieurs lignes sont insérées, supprimées ou modifiées dans la table à laquelle ils sont attachés [Site14].

Tout comme les procédures stockées, les triggers servent à exécuter une ou plusieurs instructions. Mais à la différence des procédures, il n'est pas possible d'appeler un trigger : un trigger doit être déclenché par un événement.

Un trigger est attaché à une table, et peut être déclenché par :

une insertion dans la table (requête INSERT) ;
la suppression d'une partie des données de la table (requête DELETE) ;
la modification d'une partie des données de la table (requête UPDATE).

Par ailleurs, une fois le trigger déclenché, ses instructions peuvent être exécutées soit juste avant l'exécution de l'événement déclencheur, soit juste après.

Les expressions régulières :

Les expressions régulières constituent un système très puissant et très rapide pour faire des recherches dans des chaînes de caractères (des phrases, par exemple). C'est une sorte de fonctionnalité ***Rechercher / Remplacer*** très poussée, dont vous ne pourrez plus vous passer une fois que vous saurez vous en servir.

Les expressions régulières vont nous permettre d'effectuer des recherches et des remplacements poussés dans des textes. Voici quelques exemples pratiques de ce que vous serez en mesure de faire [Site15, 2008].

✓ Vérifier la validité des chaînes de caractères (email, entrée chiffrée, entrée texte, etc...).
✓ Extraire des parties bien précises d'une chaîne, d'un texte, d'une page, etc...
✓ Rendre des emails et url cliquables dans vos textes.

✓ Gérer des éléments spécifiques placés entre vos balises perso, mise en forme, style spécial, etc...

Lemmatisation (radicalisation) :

Il y a une autre façon pour rendre les termes d'un document plus utiles pour une recherche efficace. Cette voie consiste en l'élimination des différences morphologiques non significatives sémantiquement. L'idée est de fusionner les termes ayant un sens similaire avec de petites différences sur la forme morphologique. On peut remarquer facilement que beaucoup de mots ont des formes légèrement différentes, mais leur sens reste le même ou très similaire. C'est notamment le cas des mots conjugués ou dérivés. Par exemple, les mots suivants ont des sens très similaires : informer, informés, informent, information, informateur. Si tous ces mots sont traités séparément, le rappel de recherche souffrira parce que cette différence de forme empêche le système de retrouver un texte dans lequel un mot similaire apparaît. Ainsi, l'idéal est d'éliminer toutes ces différences non significatives et ramener tous ces mots à une forme identique qu'on appellera le lemme (stem) ou la racine. L'idée de lemmatisation est d'éliminer ces indices de forme ou terminaisons à partir des termes et de ne garder que la racine ou le lemme [Kadri, 2008]. Il y a plusieurs méthodes de lemmatisation/troncature des mots. Nous présentons ici unes de ses méthodes :

Méthode de troncature de Porter :

Cette méthode consiste à examiner seulement la forme de mot, et selon la forme, on déduit la racine. Cet algorithme élimine les terminaisons d'un mot anglais en cinq étapes : la première étape essaie de transformer le pluriel en singulier. Les étapes suivantes éliminent au fur et à mesure les dérivations comme par exemple le suffixe (ness) qu'on ajoute derrière certains adjectifs comme _happiness_ [M.Porter, 1980]. Ainsi, _happiness_ et _happy_ seront tous transformés à _happi_.

Utilisation des dictionnaires :

Pour savoir si une séquence de lettres à la n d'un mot correspond à une terminaison, il suffit de faire une élimination ou une transformation, et de voir si la forme obtenue existe dans le dictionnaire. Dans le cas contraire, d'autres possibilités sont ensuite envisagées. L'utilisation d'une telle méthode requiert un dictionnaire électronique et des tables renfermant tous les affixes possibles d'une langue [Kadri, 2008].

Lemmatisation basée sur des étiqueteurs (taggeurs) :

L'idée de cette technique part du principe que pour trouver correctement le lemme d'un mot, il faut reconnaître sa catégorie grammaticale. Pour ce faire, il faut intégrer un étiqueteur automatique (analyseur de catégorie) dans un processus de lemmatisation. Avec ce mécanisme de reconnaissance de catégorie, on peut se

permettre de transformer une forme de mot en une forme standard (dite aussi forme de citation) [Kadri, 2008].

N-Gramme :

Un N-gram est une tranche de caractères N d'une chaîne plus longue. Bien que dans la littérature, le terme peut inclure la notion d'un ensemble de caractères co-occurrent dans une chaîne (par exemple, un N-gramme composé du premier et du troisième caractère d'un mot), [Cavnar et Trenkle, 1994].

La technique des n-grammes consiste à observer les collocations contiguës sur une fenêtre de n tokens[1] consécutifs d'un flux, et à essayer de tirer de ces observations des régularités relatives à un aspect particulier de ce flux [Stubbs & Barth 2003].

SenticNet :

SenticNet est une ressource lexicale pour l'analyse du sentiment au niveau du concept. Elle repose sur le Sentic Computing un nouveau paradigme multidisciplinaire pour l'analyse des sentiments.

SenticNet est une base de connaissances qui peut être utilisée pour le développement d'applications dans divers domaines tels que la grande analyse des données sociales, l'interaction homme-ordinateur et la cybersanté. Il est disponible en tant que référentiel XML autonome ou comme API [Cambria et al., 2014].

Il est en mesure d'associer la polarité et l'information affective à des concepts complexes tels que l'accomplissement d'objectifs, de célébrer une occasion spéciale et ainsi de suite [Cambria et Hussain, 2012]. À l'heure actuelle, SenticNet fournit des scores de sentiment (dans une plage comprise entre -1 et 1) pour 30.000 concepts de bon sens, instanciés soit par des mots simples, soit par des expressions multi-mots [Cambria et Hussain, 2015]. Le sentiment est véhiculé par chaque terme est défini sur la base de l'intensité de seize émotions de base, définies dans un modèle appelé Hourglass of Emotions ou Sablier d'Emotions.

[1] Le lexème est le morphème lexical d'un lemme, c'est-à-dire une unité de sens et de son qui n'est pas fonctionnelle ou dérivationnelle.



FIGURE: *Hourglass de SenticNet*

ANNEXE 2

Langage SQL :

Il a été créé dans les années 1970 et c'est devenu standard en 1986 (pour la norme ANSI - 1987 en ce qui concerne la norme ISO). Il est encore régulièrement amélioré [Site11].

Le SQL (Structured Query Language) est un langage permettant de communiquer avec une base de données. Ce langage informatique est notamment très utilisé par les développeurs web pour communiquer avec les données d'un site web [Site12].

Grâce au langage SQL, vous pourrez rechercher certains enregistrements afin de les extraire, dans l'ordre dans lequel vous souhaitez les extraire. Par exemple, vous pouvez créer une instruction SQL qui extrait les informations de plusieurs tables simultanément, ou alors un enregistrement spécifique [Site13].

Ces fonctionnalités sont comme suit :

- Définition des objets de la base de données (LLD).
- Manipulation de données (LMD).
- Contrôle des accès aux données.
- Gestion de transaction.

Feuille de style en cascade CSS (ang. Cascading Style Sheets) :

Une feuille de style CSS est un langage informatique (langage CSS) qui décrit la présentation des documents HTML, XHTML et XML. Les standards définissant le code CSS sont publiés par le World Wide Web Consortium (W3C).

L'utilisation du CSS est indispensable pour le développement web afin de rendre le site esthétique et responsive design. Dernière version CSS : CSS3.

HTML :

HTML, ou « **HyperText Markup Language** », est un langage de balisage (c.-à-d. décrit un document grâce à des balises qui sont des valeurs entourées de deux caractères, < et >, qui décrivent le contenu sémantique de la donnée) permettant d'écrire de l'hypertexte pour afficher des pages web.

HTML permet également de structurer sémantiquement et logiquement et de mettre en forme le contenu des pages, d'inclure des ressources multimédias dont des images, des formulaires de saisie, et des programmes informatiques. Il permet de créer des documents interopérables avec des équipements très variés de manière conforme aux exigences de l'accessibilité du web. Il est souvent utilisé conjointement avec le langage de programmation JavaScript et des feuilles de style en cascade (CSS).

Références bibliographiques

[Abbès, 2004]

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

[Abdul Majeed et Diab, 2011]

Abdul-Majeed. M.,Diab. M., et Korayem M., « Subjectivity and sentiment analysis of modern standard Arabic. » Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Vol.2.2011.

[Al-Subaihin et al., 2011]

Al-Subaihin A., Al-Khalifa H., et Al-Salman A.M., «A proposed sentiment analysis tool for modern arabic using human-based computing. » Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services. ACM, 2011.

[Boubou, 2007]

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

[Boullier et Lohard, 2012]

Dominique Boullier et Audrey Lohard «Opinion mining et Sentiment analysis Chapitre5. Détecter les tonalités : opinion mining et sentiment analysis »2012.

[Breck, 2008]

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

[Cadilhac, 2010]

CADILHAC Anaïs « Rôle d'une ontologie de domaine dans l'extraction d'opinion à partir de textes » Institut de recherche en informatique de Toulouse, 2010.

[Cambria et al., 2014]

Cambria, E., Olsher, D. et Rajagopal, «D.SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis». In: AAAI, pp. 1515–1521. Quebec City, (2014)

[Cambria et Hussain, 2012]

Erik Cambria et Amir Hussain. «Sentic computing. » Springer, 2012.

[Cambria et Hussain, 2015]

Erik Cambria (School of Computer Engineering Nanyang Technological University Singapore, Singapore) Amir Hussain (Computing Science and Mathematics University of Stirling Stirling, UK) «Sentic Computing A Common-Sense-Based Framework for Concept-Level Sentiment Analysis» 2012.

[Candilier, 2006]

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

[Cavnar et Trenkle, 1994]

William B. Cavnar et John M. Trenkle «N-Gram-Based Text Categorization» Environmental Research Institute of Michigan P.O. Box 134001 Ann Arbor MI 48113-4001

[Dhokrat et al., 2015]

Asmita Dhokrat Sunil Khillare C. et Namrata Mahender «Review on Techniques and Tools used for Opinion Mining» Dept. of CS & IT Dr. Babasaheb Ambedkar Marathwada University Aurangabad, India 2015

[Ding et al., 2007]

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

[Doty, 2008]

Python Basics S.R. Doty August 27, 2008

[Douzidia, 2004]

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

[El-Halees, 2011]

El-Halees A., « Arabic Opinion Mining Using Combined Classification Approach. » Proceedings of the International Arab Conference on Information Technology, ACIT. 2011.

[Généreux et al, 2009]

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

[Gillot ,2010]

Sébastien Gillot «Rapport de stage Fouille d'opinions » (MRI parcours P4) 4 juin 2010

[Hadrich et Chaâben, 2006]

Lamia Hadrich Belguith et Nouha Chaâben, « Analyse et désambiguïsation morphologiques de textes arabes non voyellés », Faculté des Sciences Économiques et de Gestion de Sfax – Laboratoire LARIS, 2006.

[Hu et Liu, 2004]

Hu Minqing et Liu Bing, «Mining opinion features in customer reviews », In Deborah L. Mcguinness, George Ferguson, Deborah L. Mcguinness, and George Ferguson, editors,AAAI,. AAAI Press, pp. 755–760, 2004

[Hu et al., 2005]

Bing Liu, Minqing Hu et Junsheng Cheng «Opinion Observer: Analyzing and Comparing Opinions on the Web» Proceedings of the 14th international World Wide Web conference (WWW-2005), May 10-14, 2005, in Chiba, Japan.

[Hu et al., 2006]

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

[Jindal et Liu, 2006]

Nitin Jindal et Bing Liu. «Mining Comprative Sentences and Relations » Proceedings of 21st National Conference on Artificial Intellgience (AAAI-2006), July 16.20, 2006, Boston, Massachusetts, USA.

[J.Wiebe et al., 1999]

Wiebe Janyce, Rebecca F. Bruce, et Thomas P. O'Hara «Development and use of a gold-standard data set for subjectivity classifications», In Proceedings of the Association for Computational Linguistics (AC L-1999). 1999.

[J. Wiebe et al., 2005]

J. Wiebe, T. Wilson et C. Cardie «Annotating expressions of opinions and emotions in language.Language Resources and Evaluation», 2005.

[Kadri et Benyamina, 1992]

Y. Kadri et A. Benyamina, « Un système d'analyse syntaxicosémantique du langage arabe non voyellé », Mémoire d'ingénieur, Université d'Oran, (1992).

[Kadri, 2008]

Youssef Kadri « Recherche d'Information Translinguistique sur les Documents en Arabe » Université de Montréal 2008.

[Kelaiaia, 2008]

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

[Liu, 2006]

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

[Liu, 2009]

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

[Liu, 2010]

Bing Liu «Sentiment analysis and subjectivity», In Nitin Indurkhya and Fred J. Damerau, editors, Handbook of Natural Language Processing 2010.

[Liu, 2012]

Bing Liu «Sentiment Analysis and Opinion Mining», Graeme Hirst University of Toronto 2012.

[Lazhar, 2014]

Lazhar FAREK « Identification d'opinions dans les textes arabes en utilisant les ontologies » université Badji Mokhtar Annaba, 2014.

[Maurel et al., 2007]

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

[Mesfar, 2008]

Slim MESFAR « Analyse morphosyntaxique automatique et reconnaissance des entités nommées en arabe standard » Université de Franche-Comté, 2008

[M.Porter, 1980]

M. Porter «An algorithm for suffix stripping. Automated Library and Information Systems», 1980.

[Pang et al., 2002]

Pang, Bo, Lillian Lee, et Shivakumar Vaithyanathan «Thumbs up? : sentiment classification using machine learning techniques», In Proceedings of Conference on Empirical Methods in Natural Language Processing, 2002.

[Pang et Lee, 2008]

Pang, Bo et Lillian Lee, «Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval», 2008.

[Rushdi-Saleh et al., 2011]

Rushdi-Saleh M., Martín-Valdivia M., Ureña-López L., Perea-Ortega J.M.et Bilingual «Experiments with an Arabic-English Corpus for Opinion Mining», 2011.

[Savoy, 1993]

J. Savoy «Stemming of FrenchWords Based on Grammatical Category»; Journal of the American Society for Information Science, 1993.

[Souissi, 1997]

SOUISSI E « Étiquetage grammatical de l'arabe voyellé ou non » ; Thèse de doctorat, Université de Paris III, 1997.

[Stubbs & Barth 2003]

Stubbs M. et Barth I. «Using recurrent phrases as text-type discriminators: a quantitative method and some findings», Functions of Language, 2003.

[Turney, 2002]

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

[Yvon, 2007]

François Yvon « Une petite introduction au Traitement Automatique des Langues Naturelles » , 2007.

Références Webographiques :

[Site1, 2009] Lien: http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html.

[Site2, 2015] Lien: https://www.atala.org/Fouille-d-opinions-et-analyse-de-sentiments

[Site3, 2008] Lien: http://sis.univ-tln.fr/~tollari/ARTICLES/DEA2003

[Site4, 2008] Lien: http://www.dbmsmag.com/9807m05.html

[Site5, 2001] Lien: https://docs.python.org/3/tutorial/index.html

[Site6, 2011] Lien: http://pythoncentral.io/introduction-to-tweepy-twitter-for-python/

[Site7, 2005] Lien: http://www.json.org/json-fr.html

[Site8, 2016] Lien: http://pandas.pydata.org/

[Site9, 2003] Lien: http://apprendre-python.com/page-database-data-base-donnees-query-sql-mysql-postgre-sqlite

[Site10, 2011] Lien : https://www.synbioz.com/blog/introduction_a_html5

[Site11] Lien : https://openclassrooms.com/courses/administrez-vos-bases-de-donnees-avec-mysql/introduction-14

[Site12] Lien: http://sql.sh/

[Site13]Lien: https://openclassrooms.com/courses/apprenez-a-programmer-en-vb net/introduction-au-langage-sql

[Site14] Lien: https://openclassrooms.com/courses/administrez-vos-bases-de-donnees-avec-mysql/triggers