MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DR. YAHIA FARES DE MEDEA

FACULTE DES SCIENCES ET DE LA TECHNOLOGIE

Département du Génie Électrique et de l'Informatique



Projet de fin d'études

Pour l'obtention du Diplôme de Master en Informatique

Option : Ingénierie des Systèmes et Technologies du Web

THÈME

CLASSIFICATION D'OPINIONS PAR GENRE DANS LES RÉSEAUX SOCIAUX : ARABIC TWITTER

Proposé et dirigé par :

M. MAZARI Ahmed Cherif

Préparé et présenté par :

M^{elle}. OULD KEBAILI Hadjer

Année universitaire 2015 - 2016

Remerciements

Merci Allah, le Tout Puissant qui m'ont donné la force, le courage et la volonté pour réaliser ce travail.

Je tiens à exprimer ma profonde reconnaissance à mon encadreur, Mr. MAZARI Ahmed Cherif, pour m'avoir dirigée, guidée, conseillée, pendant le déroulement de ce travail, dont j'espère être à la hauteur de ses attentes.

Je remercie du fond du cœur mes chers parents de tout ce qu'ils ont faisaient pour moi.

Mes remerciements vont également à ma famille pour le soutien et l'encouragement.

Je tiens à remercier aussi tous les membres du jury qui me font l'honneur d'examiner mon travail.

Je tiens enfin à remercier tous ceux qui m'ont aidé et contribué à réaliser ce travail de près ou de loin.

Dédicace

Je dédie ce travail :

À mes parents,

À mes sœurs et mon frère,

À tous les membres de ma famille,

À mes amies et mes collègues de la section ISTW,

À tous ceux qui, de près ou de loin, ont soutenu mon travail,

par leurs encouragements et conseils.

Résumé

Avec la croissance rapide des technologies des réseaux sociaux sur le web au cours des dernières années, l'analyse des sentiments et l'identification du genre sont révélés de plus en plus utile. La prédiction exacte des attributs démographiques des média sociaux et d'autres contenu en ligne informel est précieux pour le marketing, la personnalisation, la sociologie et les enquêtes juridiques. L'analyse de paternité fournit des informations sur les détenteurs d'une opinion y compris souvent le genre du détenteur. Les hommes et les femmes expriment des opinions d'une manière nettement différentes.

Ce mémoire traite la tâche de classification de l'utilisateur arabe dans les médias sociaux, avec une application à Twitter. Nous en déduisons automatiquement les valeurs des attributs de l'utilisateur tels que l'orientation politique ou sportive en tirant parti des informations observables telles que les coordonnées du profil utilisateur, le comportement et le contenu linguistique de l'utilisateur Twitter. Nous employons une approche d'apprentissage automatique qui repose sur un ensemble complet de traits dérivés de ces informations qui montrent que les traits linguistiques riches sont très prometteurs pour les besoins de classification du genre.

Ce mémoire étudie aussi le problème de la classification d'opinions dans les textes arabes standard et dialectique. Notre approche introduit aussi une classification des tweets d'un hashtag en positif et négatif en se basant sur l'apprentissage automatique. Nous avons utilisé un ensemble de différents traits numérique (standard) et linguistique en appliquant le classificateur naïf bayésien.

Le présent mémoire décrit également un ensemble de statistiques qui démontre la différence entre les opinions mâle et femelle.

Mots clés : analyse des sentiments, identification du genre, classification d'opinion, apprentissage automatique, langue arabe.

Table des matières

Introduction **1**

I État de l'art 6

1 Fouille d'opinion 7

1 Introduction . . . 9
2 Enjeux du Web 2.0 . . . 9
3 Médias sociaux . . . 11
4 Sentiment . . . 12
5 Opinion . . . 12
5.1 Opinion et fait . . . 12
5.2 Complexité de la notion d'opinion . . . 13
6 Fouille d'opinions et analyse des sentiments . . . 14
7 Processus de la fouille d'opinions . . . 15
7.1 Acquisition et prétraitement du corpus . . . 16
7.2 La pertinence par rapport au sujet . . . 16
7.3 La détection d'opinions . . . 16
8 Domaines de la fouille d'opinions . . . 16
8.1 Identification des textes d'opinions . . . 16
8.2 Résumé d'opinions . . . 17
8.3 Classification d'opinions . . . 17
9 Facteurs de difficulté de la fouille d'opinions . . . 17
9.1 Hétérogénéité des données . . . 17
9.2 Dépendance du contexte . . . 19
9.3 Dépendance du domaine . . . 19
9.4 Ambiguïté subjective . . . 19
10 Domaines d'application de la fouille d'opinions . . . 19
10.1 Marketing . . . 19
10.1.1 Coté entreprises . . . 19
10.1.2 Coté clients . . . 20

10.2 Politique 20
11 Terminologie 20
11.1 Objet 21
11.2 Modèle d'un objet 22
11.3 Opinion 22
11.4 Détenteur de l'opinion *(ang. Opinion holder)* 22
11.5 Orientation d'une opinion 23
11.6 Opinion exprimée sur un trait 23
11.7 Trait explicite et trait implicite 23
11.8 Modèle d'un document d'opinions 23
11.9 Émotions 24
11.10 Subjectivité 25
11.11 Phrase subjective 26
11.12 Opinion explicite et opinion implicite 26
11.13 Phrase d'opinions 26
12 Conclusion 27

2 Classification d'opinions **28**
1 Introduction 30
2 Classification d'opinions 30
2.1 Objectifs et définition 31
2.2 Complexité et caractéristiques de la tâche 32
3 Classification d'opinions au niveau des documents 32
3.1 Classification d'opinions non supervisé 33
3.1.1 Approches basée sur les patrons syntaxiques 33
3.1.2 Approches linguistiques basées sur les lexiques 34
3.2 Classification d'opinion supervisée 39
3.2.1 Approches statistiques basée sur l'apprentissage automatique 39
3.2.2 Classification d'opinion utilisant la fonction de score personnalisé 51
3.3 Classification d'opinion semi supervisée (approches hybrides) 52
3.3.1 Linguistique au service de l'apprentissage automatique . . . 53
3.3.2 Apprentissage automatique au service de la linguistique . . . 53
3.3.3 Fusion a posteriori des résultats des deux approches 53
4 Classification d'opinions au niveau des phrases 54
5 Classification d'opinions au niveau des traits linguistiques 55
6 Classification d'opinions avec croisement de langues 56
7 Classification d'opinions avec croisement de domaines 58
8 Mesures d'évaluations 59
9 Fouille d'opinions en langue arabe 62
10 Travaux connexes à l'identification du genre 63
11 Conclusion 63

II Contribution **65**

3 GenderSentiment : Approche proposée **66**
1 Introduction . . . 68
2 Motivation . . . 69
3 Description . . . 70
4 Entrées de notre système . . . 70
4.1 Profils utilisateurs . . . 70
4.1.1 Traits du profils . . . 71
4.1.2 Traits de comportement de messagegerie . . . 71
4.1.3 Traits linguistiques . . . 71
4.2 Lexique des sentiments . . . 71
4.2.1 Initialisation du dictionnaire . . . 73
4.2.2 Enrichissement du dictionnaire . . . 74
4.3 Textes évaluatifs . . . 74
4.3.1 Segmentation . . . 74
4.3.2 Étiquetage . . . 75
5 Architecture . . . 76
5.1 Détection du genre . . . 76
5.1.1 Extraction des traits du profils . . . 76
5.1.2 Extraction des traits de comportement de messagerie . . . 76
5.1.3 Extraction des traits linguistiques . . . 77
5.1.4 Association des traits avec les utilisateurs . . . 78
5.1.5 Classification . . . 78
5.1.6 Classificateur . . . 78
5.2 analyse des sentiments . . . 79
5.2.1 Identification des textes subjectifs . . . 79
5.2.2 Extraction des traits explicites . . . 80
5.2.3 Association des traits avec les textes subjectifs . . . 81
5.2.4 Classification . . . 81
5.2.5 Classificateur . . . 82
5.3 Évaluation des méthodes de classification . . . 83
5.4 Architecture générale . . . 83
6 Conclusion . . . 84

4 Implémentation et évaluation **85**
1 Introduction . . . 87
2 Choix du domaine étudié . . . 87
3 Lexiques de sentiments . . . 87
4 Corpus . . . 88
5 Environnement de développement . . . 89
5.1 Langage de programmation JAVA . . . 89

5.2 Environnement de développement NetBeans 89
6 Outils utilisés 89
6.1 Twitter et Twitter4j 90
6.2 MADAMIRA 90
6.3 RapidMiner 91
7 Expérimentation et évaluation 91
7.1 Expérimentation 91
7.1.1 Interface de l'application 92
7.1.2 Composantes de l'application 94
7.2 Classification et Évaluation 96
7.2.1 Classification 96
7.2.2 Évaluation 101
8 Conclusion 103

Conclusion et perspectives **104**

Table des figures

1.1 Commentaire collecté du site web restomontreal.ca . . . 13
1.2 Commentaire collecté du site web restomontreal.ca . . . 13
1.3 Commentaire collecté du site web restomontreal.ca . . . 13
1.4 Commentaire collecté du site web restomontreal.ca . . . 14
1.5 Processus de la fouille d'opinions . . . 15
1.6 Ensemble de commentaires sur une vidéo retrouvés sur Youtube . . . 18
1.7 Ensemble de commentaires retrouvés sur Facebook, exprimés en sarcasme . . . 18
1.8 Exemple des émotions de base et secondaires . . . 25

2.1 Exemple d'arbre de synonymes et d'antonymes présents dans WordNet . . . 37
2.2 Exemple d'un clasificateur SVM . . . 45
2.3 Utilisation d'une fonction noyau . . . 46
2.4 Exemple d'un arbre de décision . . . 49
2.5 Structure d'un neurone artificiel - Le neurone calcule la somme de ses entrées puis cette valeur passe à travers la fonction d'activation pour produire sa sortie. 51
2.6 Représentation du bruit et du silence en recherche d'information. . . . 60

3.1 Exemple d'un texte arabe segmenté . . . 75
3.2 Exemple d'un texte arabe étiqueté . . . 75
3.3 Architecture de la détection du genre . . . 78
3.4 Processus d'identification des textes subjectifs avec enrichissement de lexique des sentiments . . . 80
3.5 Architecture d'analyse des sentiments . . . 82
3.6 Architecture d'analyse des sentiments . . . 84

4.1 Structure de base d'un tweet . . . 89
4.2 diagramme de classe pour notre application . . . 92
4.3 Interface principale de notre application . . . 93
4.4 Application du classificateur Naïf Bayes . . . 97
4.5 Validation de test . . . 98
4.6 Résultat de classification selon les traits de profil . . . 98
4.7 Résultat de classification selon les traits de comportement de messagerie . . . 98

4.8 Résultat de classification selon les traits linguistiques 99
4.9 Résultat de classification générale 99
4.10 Résultat de classification des tous les textes évaluatifs 100
4.11 Résultat de classification des textes subjectifs 100
4.12 Résultat de classification des opinions des hommes 101
4.13 Résultat de classification des opinions des femmes 101
4.14 Évaluation des différents traits de détection du genre 102
4.15 Évaluation des différents traits de classification d'opinions 102
4.16 statistiques d'opinions 103

Liste des tableaux

3.1 Patrons POS tags pour extraire des phrases de deux mots. 33
3.2 Exemple de catégories d'adverbes 38
5.3 Polarité des traits 56
8.4 Exemple d'une matrice de confusion pour une classification binaire 60

4.1 Structure simplifiée d'un dictionnaire des sentiments 73

7.1 Résultats des traits de détection du genre 99
7.2 Résultats des traits de classification d'opinions 100
7.3 Résultats de classification des opinions des hommes et des femmes 101

Liste des acronymes

API Application Programming Interface

BSWE Bilingual Sentiment Word Embeddings

CLSC Cross Language Sentiment Classification

EM Expectation Maximization

EO Expression d'Opinion

HTML HyperText Markup Language

IBM International Business Machines

IDE Integrated Development Environment

IDF Inverse Document Frequency

MaxEnt Maximum Entropy

MSA Modern Standard Arabic

NB Naïf Bayes

NLP Natural Language Processing

NN Neutral Networks

NTUSD NTU Sentiment Dictionary

OM Opinion-Mining

POS Part Of Speech

PMI Pointwise Mutual Information)

RI Recherche d'Informations

SA Sentiment Analysis

SO Sentiment Orientation

SS Segment Subjectif

SVM Support Vector Machine

TAL Traitement Automatique des Langues Naturels

TF Term Frequency

TS Texte Subjectif

URL Uniform Resource Locator

UGC User Generated Content

WWW World Wide Web

XML eXtensible Markup Language

- **Motivation**
  De nos jours, l'Internet est un outil incontournable d'échange d'information tant au niveau personnel que professionnel. Le Web nous offre un monde de l'information prodigieux et a évolué des simples ensembles de pages statiques vers des services de plus en plus complexes. Ces services nous offrent l'achat de tous les produits, la lecture du journal préféré en ligne, la rencontre de l'âme soeur, la discussion sur de multiples forums ou la possibilité de s'exprimer sur les blogs.

  Du point de vue des clients, donner une considération aux opinions des autres avant d'acheter un produit est un comportement commun, même avant l'existence de l'Internet. Dans l'ère numérique, la différence est que le client a un accès à des milliers d'opinions, ce qui améliore grandement la prise de décision. Fondamentalement, les clients veulent trouver le meilleur produit avec le prix le plus bas. En d'autres termes, ils cherchent des produits qui satisfont leurs besoins avec des prix dont ils sont capables de payer.

  Ces besoins se varient entre les gens sachant que les besoins d'un homme ne sont pas les mêmes besoins d'une femme. Il est important de souligner que le but d'analyser les opinions des autres vient de leur caractère neutre, qui ne sont habituellement pas liés à une organisation ou une entreprise parce qu'elles représentent la voix des consommateurs ordinaires, et qu'elles diffèrent grandement des annonces (les annonces sont généralement biaisées et ont une tendance à favoriser le produit, en soulignant les aspects positifs et en dissimulant les négatifs) [1].

  Du point de vue du commerce électronique *(ang. e-commerce)*, recevoir les commentaires des consommateurs peut grandement améliorer ses stratégies afin d'augmenter les bénéfices du secteur. Par exemple, une boutique en ligne peut placer des annonces en mesurant le niveau de satisfaction des consommateurs pour un produit donné. Il est fréquent de trouver des produits avec des milliers d'opinions, donc, il pourrait être une tâche difficile pour un client d'analyser chacune d'eux. En outre, il pourrait être un travail très fastidieux de trouver des opinions sur quelques caractéristiques d'un produit, généralement, lorsqu'il s'agit d'une exigence d'un client expérimenté.

  Une différence importante rend les techniques actuelles de classification pas assez efficaces pour décrire les informations représentées par les opinions. Cette différence est principalement due à la nature de l'information textuelle dans le monde. Ces informations sont soient des faits ou des opinions. Les systèmes de recherche actuels sont axés sur les faits (par exemple, des mécanismes de classement utilisés par les moteurs de recherche). Un fait est habituellement égal à tous les mêmes autres faits. Cependant, une opinion est une croyance ou un jugement sur un sujet ou un objet. Par conséquent, une opinion exprimée sur un objet est généralement différente des autres opinions exprimées sur le même objet.

- **Problématique de recherche**
  Ce mémoire présente une contribution à deux domaines : l'identification du genre, et la fouille d'opinion *(ang. Opinion-Mining)*, dont l'objectif est de concevoir et de développer une approche pour la classification des opinions par genre dans les réseaux sociaux et plus précisément dans les textes arabes de la source Internet la plus connue Twitter sur un domaine étudié.

  Le genre est une question de vie importante. Surtout quand nous ferons face aux défis de la vie, nous avons tendance à prendre en compte des solutions spécifiques qui correspondent le mieux que le genre. Qaund la discussion va de la mode, à la santé, à la richesse ou à l'emploi, la distinction de genre fait partie intégrante. Dans chacun de ces sujets, la discussion diffère en ce qui concerne le genre, afin de comprendre et d'offrir la meilleure solution unique à chaque genre. Les médias sociaux, tels que Twitter, offre aux utilisateurs un moyen de communiquer et de partager avec des amis, les familles, les organisations et le gouvernement. . Et, inhérent à ces communications volontaires sont des opinions, des styles, des préférences et des idées qui peuvent être propres à chaque genre. Comprendre le contenu généré par l'utilisateur en fonction du genre serait utile aux particuliers, aux entreprises, et même les gouvernements pour la recommandation personnelle, la personnalisation, la publicité ciblée, et la formulation des politiques.

  Le contenu généré sur la base d'utilisateurs de Twitter et sa couverture fait Twitter une source riche pour des buts de recherche. Pourtant, les caractéristiques latentes de Twitter font les défis actuels. Ces défis pourraient être traitées en utilisant des approches d'apprentissage automatique et en concevant des systèmes qui peuvent automatiquement et avec précision prédire les caractéristiques latentes de Twitter pour aider à construire l'intelligence d'affaires et de marketing, avis des clients, la planification des achats, de la gestion d'opinion publique, l'élaboration des politiques, des systèmes de recherche de sentiment, et ainsi de suite.

  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'identification d'opinions à partir des textes 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 défis 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 clients sous forme de commentaires et critiques sur les produits commercialisés, cette augmentation a été remarquée surtout après l'apparence du Web 2.0. Les opinions des clients représentent une source d'information qui ne devraient 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.

En Opinion-Mining, les travaux de recherche peuvent être catégorisés en deux principaux sous-thèmes : la classification d'opinions au niveau de la caractéristique et la classification d'opinion au niveau du document. La différence entre ces deux niveaux est expliquée dans [2] comme suit : la classification d'un document en positive ou négative est basée sur le sentiment global exprimé par le détenteur de l'opinion *(ang. Opinion holder)*. Cependant, l'objectif de la classification au niveau de la caractéristique est de produire un résumé plus ou moins détaillé sur les caractéristiques de l'objet sur lequel une opinion a été exprimée. En effet, l'opinion globale prévue exprimée dans un document dépend sur les polarités positives et négatives (orientations sémantiques) des expressions d'opinion utilisées dans le document. Dans un commentaire ou une critique, les opinions peuvent être exprimées de manière explicite comme dans les exemples suivants : « Un bon travail » (1), « Une mauvaise habitude » (2), tandis que dans d'autres, il est difficile de localiser l'entité sur laquelle l'opinion est exprimée comme dans les exemples suivants : « Ahmed a une bonne note » (3), « Le mauvais service nous a obligé de quitter le complexe touristique » (4). Dans les exemples (1) et (2), il est clair de constater que les mots d'opinion « bon » et « mauvais » sont respectivement exprimés sur les deux entités « travail » et « habitude » qui donnent respectivement une opinion positive envers « travail » et une opinion négative envers « habitude ». Dans l'exemple (3), deux entités sont présentes « Ahmed » et « note » et un seul mot d'opinion « bon », il est facile de trouver en utilisant les outils de TALN que l'adjectif « bon » est lié à l'entité « note », si une opinion positive est exprimée envers « note », intuitivement, nous pouvons déduire que l'opinion positive est exprimée implicitement sur l'entité « Ahmed » parce que c'est lui qui a obtenu la « note ». La même observation pour l'exemple (4) où l'opinion négative exprimée sur l'entité « service » en utilisant l'adjectif « mauvais », une opinion est implicitement exprimée sur l'entité « complexe touristique », parce que les deux entités « complexe touristique » et « service » sont sémantiquement dépendantes.

Dans ce mémoire, nous présentons une approche de classification d'opinions par genre sur des textes arabes. Cette approche vise à extraire les traits *(ang. features)* sur lesquelles des utilisateurs ont été classé et des opinions ont été exprimées. Des informations d'un profil utilisateur ont été utilisé pour classer les utilisateurs en mâle et femelle. Un lexique des sentiments spécifique à un domaine étudié est utilisé pour identifier les expressions

d'opinion et les textes subjectifs, où chaque mot de ce lexique peut être un adjectif, un verbe ou un adverbe éventuellement combiné avec certains modificateurs tels que la négation et la confirmation, et pour chaque mot de ce lexique, une polarité positive ou négative est associée.

- **Contexte du travail**
  Le premier axe de notre recherche est la construction d'un corpus d'apprentissage arabe sur un domaine donné et qui regroupe les opinions et ses détenteurs.

  Le deuxième axe est la détection du genre des détenteurs d'opinion arabophones basée sur l'apprentissage automatique et qui néglige un ensemble de trait de profil utilisateur (numéro de téléphone, instagram, URL...), un ensemble de traits de comportement de messagerie et enfin un ensemble qui est le plus riche composé des traits linguistiques afin de classer les détenteurs en mâle et femelle.

  Le troisième axe est l'identification des textes subjectifs à partir de l'ensemble des textes évaluatifs en se basant sur un lexique de sentiments.

  — Les mots de polarité négative qui décrivent un sentiment ou un état psychologique plutôt désagréable, comme la peur ou la colère.

  — Les mots de polarité positive qui décrivent un sentiment ou un état psychologique plutôt agréable, comme l'amour ou l'amusement.

  Puis la classification des opinions en positives et négatives.

  Le quatrième axe consiste à faire des statistiques qui évaluent les résultats des opinions selon le genre des détenteurs.

- **Organisation du mémoire**

Le mémoire est structuré en deux parties.

Dans la partie I - État de l'art - nous introduisons le problème de la fouille d'opinion et puis nous donnons un bref aperçu de l'état de l'art de la classification d'opinion. Nous sondons et discutons la littérature pertinente. La vue d'ensemble est structuré dans les chapitres suivants :

chapitre 1 : définit les termes de base de la fouille d'opinion : subjectivité, polarité, intensité d'opinion, et des opinions elle mêmes : opinion, sentiment, émotion, opinions implicites et explicites et le lien entre eux, etc. Nous présentons aussi, les applications de la fouille d'opinions, et sa contribution efficace dans la prise de décision.

chapitre 2 : Dans ce chapitre, nous présentons le problème de classification d'opinions et les différentes techniques et algorithmes utilisés ainsi que les travaux connexes dans le domaine de la fouille d'opinions qui nous mènent à découvrir les points faibles et les points forts de chaque approche utilisée, les difficultés rencontrées, en vue de donner plus de motivation envers notre approche.

Dans la Partie II - Contributions - détaille les principales contributions et l'implémentation du cas d'étude avec une évaluation finale et nous discutons des orientations pertinentes pour la recherche future.

chapitre 3 : Dans le cadre de ce chapitre, nous proposons une approche GenderSentiment, pour la classification des opinions par genre à partir des textes arabes, nous présentons l'architecture générale de notre approche en détaillant chaque module à part, en partant de la sélection des données jusqu'à la phase de classification des détenteurs d'opinion et des opinons identifiées et des statistiques concernant ces derniers.

chapitre 4 : C'est dans ce chapitre, que nous présentons l'implémentation et l'évaluation de notre approche, en concluant par un résumé regroupant les difficultés rencontrées, les points faibles et les points forts de notre approche.

Enfin, une conclusion donnant un aperçu synthétique sur notre approche et un ensemble de perspectives sur les facteurs qui ont influencé la performance de notre système, et qui seront sujets d'éventuels futurs travaux de recherche.

Première partie

État de l’art

Chapitre 1

Fouille d'opinion

Sommaire

1 Introduction . . . 9
2 Enjeux du Web 2.0 . . . 9
3 Médias sociaux . . . 11
4 Sentiment . . . 12
5 Opinion . . . 12
5.1 Opinion et fait . . . 12
5.2 Complexité de la notion d'opinion . . . 13
6 Fouille d'opinions et analyse des sentiments . . . 14
7 Processus de la fouille d'opinions . . . 15
7.1 Acquisition et prétraitement du corpus . . . 16
7.2 La pertinence par rapport au sujet . . . 16
7.3 La détection d'opinions . . . 16
8 Domaines de la fouille d'opinions . . . 16
8.1 Identification des textes d'opinions . . . 16
8.2 Résumé d'opinions . . . 17
8.3 Classification d'opinions . . . 17
9 Facteurs de difficulté de la fouille d'opinions . . . 17
9.1 Hétérogénéité des données . . . 17
9.2 Dépendance du contexte . . . 19
9.3 Dépendance du domaine . . . 19
9.4 Ambiguïté subjective . . . 19
10 Domaines d'application de la fouille d'opinions . . . 19
10.1 Marketing . . . 19
10.2 Politique . . . 20
11 Terminologie . . . 20
11.1 Objet . . . 21
11.2 Modèle d'un objet . . . 22
11.3 Opinion . . . 22
11.4 Détenteur de l'opinion *(ang. Opinion holder)* . . . 22
11.5 Orientation d'une opinion . . . 23
11.6 Opinion exprimée sur un trait . . . 23
11.7 Trait explicite et trait implicite . . . 23
11.8 Modèle d'un document d'opinions . . . 23
11.9 Émotions . . . 24
11.10 Subjectivité . . . 25
11.11 Phrase subjective . . . 26
11.12 Opinion explicite et opinion implicite . . . 26
11.13 Phrase d'opinions . . . 26
12 Conclusion . . . 27

1 Introduction

« Que pensent les autres ? » a toujours été un élément d'information important pour la plupart d'entre nous pendant le processus de prise de décision. Pang et Lee (2008) exposent dans leur livre [3] les intérêts socio-économiques d'un moteur de recherche hypothétique qui répondrait à la requête « Que pensent les gens de ... ?».. Avant l'apparence des réseaux sociaux et les sites de commerce électroniques sur le web, beaucoup d'entre nous demande à ses amis ou sa famille de recommander un produit ou un service avant la prise de décision. Mais avec le développement du Web 2.0, la prise de décision est devenue plus ou moins facile qu'avant, l'information est devenue à la portée de tout le monde où les utilisateurs disposent de nombreuses et nouvelles facilités (forums, blogs, groupes de discussions, etc.) pour exprimer leurs opinions sur des différents sujets, produits ou services.

La quantité énorme d'informations portant des opinions et des critiques disponibles pose un problème du coût et du temps pour l'explorer, d'où il est devenu indispensable de proposer de nouvelles approches afin d'extraire automatiquement les opinions exprimées. Pour répondre à ce besoin, de nombreux travaux de recherche sont apparus. Ils sont issus de différents domaines : fouille de données, aide à la décision, modélisation des connaissances, traitement automatique des langues ou linguistique.

Ce chapitre donne un aperçu général sur la fouille d'opinions, son origine et ses problèmes, en présentant les différents concepts utilisés, les axes de recherche, et le problème de classification d'opinions. Des concepts comme les opinions implicites et explicites exprimées sur des traits linguistiques *(ang. features)* implicites et explicites sont exposés dans ce chapitre, illustrés par des exemples.

2 Enjeux du Web 2.0

Aujourd'hui, l'internaute n'est plus simplement spectateur, il peut s'il le souhaite devenir acteur. Et tout est mis en oeuvre pour que le simple spectateur devienne acteur du Web afin, entre autres choses, de le fidéliser. Cette révolution a donnée naissance à ce que l'on nomme aujourd'hui le Web 2.0 appelée encore Web Social ou Web Participatif [4]. Bien que la définition du Web 2.0 ne soit pas encore parfaitement établie, un grand aspect caractérisant cette nouvelle génération du Web peut être mis en avant : le rôle central de l'utilisateur.

Une caractéristique principale du Web 2.0 qui le distingue nettement du Web 1.0 est donc la prise de contrôle de l'information par les utilisateurs. N'importe quel internaute peut aujourd'hui apporter sa pierre à l'édifice. Il peut se faire une place sur la toile, collaborer, partager des informations, des outils, des fichiers multimédias, donner ses opinions, commenter, réagir, etc. et tout ceci sans connaissances spécifiques. En effet, quand auparavant il fallait un minimum de savoir-faire en informatique et en programmation pour créer son espace sur Internet, aujourd'hui il suffit de savoir cliquer car de plus en plus d'outils sont mis à disposition de tout un

chacun afin de faciliter toutes ces interactions. Parmi ces outils, nous pouvons citer les réseaux sociaux, les blogs, les wikis, les sites de partage de vidéos, de photos, de musiques, etc. [4].

De nombreux sites présents sur Internet aujourd'hui offrent la possibilité à tous leurs visiteurs de laisser, au minimum, une trace textuelle et ainsi s'exprimer publiquement. Tout ce contenu, qu'il soit textuel ou autre, est appelé Contenu Généré par les Utilisateurs ou UGC *(ang.User Generated Content)*. Il représente une quantité de données de plus en plus importante sur la toile et est composé, en très grande partie, de données textuelles. Ce nouvel espace d'expression représente une grosse quantité d'informations, notamment en termes d'avis et d'opinions, susceptibles d'être exploitées à des fins diverses. Les données textuelles, notamment, peuvent être analysées dans différents buts. Par exemple, dans le domaine de la fouille d'opinion *(ang.Opinion Mining)*, les textes sont utilisés afin de permettre à des entreprises de connaître automatiquement l'image que les consommateurs ont d'eux, de même pour faire de la comparaison d'articles de vente, réaliser des sondages, détecter des rumeurs, etc.

Les textes rédigés par les internautes sont en général beaucoup plus subjectifs que les articles rédigés par des professionnels et donc beaucoup plus porteurs d'opinion. De plus, ils contiennent un éventail d'avis et de jugements souvent plus représentatifs du consommateur ordinaire, étant rédigés par un panel d'individus variés parmi lesquels on retrouve tout autant de profanes que de connaisseurs du sujet abordé. Par exemples, pour ce qui est du cas des films, il est possible de trouver des avis de spectateurs de tous âges et de tous horizons. Ce que n'offrent pas les critiques journalistiques qui sont généralement rédigés par des personnes du même milieu, c'est-à-dire dans ce cas, par des journalistes cinéphiles appartenant à la population active et ayant fait un certain nombre d'années d'études [4].

Généralement, le Contenu Généré par les Utilisateurs est donc composé de données textuelles qui sont porteuses d'opinions et de sentiments. L'accès au contenu sémantique de ces données, préalable à la connaissance des opinions qu'elles véhiculent, représente un enjeu pour de nombreux acteurs :

- le consommateur, c'est-à-dire chacun de nous, qui veut s'informer avant toute décision qu'elle soit d'achat ou autre ;
- les fournisseurs de biens et de services qui cherchent à se positionner les uns par rapport aux autres dans un univers hautement compétitif et face à une demande de plus en plus complexe à identifier ;
- les chercheurs : économistes, sociologues,... ou simplement les responsables publics qui cherchent à comprendre le comportement individuel ou collectif pour anticiper, réguler ou ajuster les rapports entre les différents agents socioéconomiques.

C'est dans ce contexte que s'introduit le domaine de la fouille d'opinion [5].

3 Médias sociaux

De nos jours, l'Internet se révèle plus que jamais un outil indispensable d'échange d'informations. Il nous offre une quantité considérable d'informations à une vitesse inédite et ses services s'adaptent de plus en plus aux besoins des internautes. Ceux-ci peuvent consulter l'Internet pour trouver de l'information, envoyer des e-mails, acheter des produits, lire des journaux en ligne, etc.

Pendant les dernières années, l'Internet a connu encore une plus vaste portée grâce au développement des médias sociaux. Basés sur des techniques de communication faciles et accessibles pour tous, ces médias favorisent les interactions sociales à travers l'Internet. Les médias sociaux se distinguent des médias traditionnels tels que les journaux, la télévision et la radio parce que leur utilisation est peu coûteuse et libre, de façon à permettre à tout le monde d'accéder à ou de publier de l'information. Ils subviennent aux besoins des individus d'échanger des opinions, de demander des conseils et de communiquer de façon rapide et facile. Les possibilités sont nombreuses : envoyer des messages et des photos, réaliser des vidéoconférences, télécharger des documents, etc. Offrant un accès libre et gratuit, les médias sociaux ont considérablement favorisé la communication de masse et ils ont déclenché le débat public sur Internet.

Les médias sociaux qui ont récemment pu bénéficier d'un considérable essor sont les réseaux sociaux tels que Facebook, LinkedIn et Twitter. Ce sont des sites web qui rassemblent des identités sociales telles que des individus, des entreprises et des organisations qui peuvent échanger de l'information à travers des interactions sociales. Grâce à leur caractère maniable et leur accès libre, les réseaux sociaux bénéficient d'un succès croissant auprès du grand public.

La popularité des nouveaux médias est d'autant plus grande que la demande d'informations est devenue plus importante dans notre société. En général, les gens aiment consulter les avis d'autres personnes avant de passer à l'action ou de se faire une opinion. Autrefois, ces avis provenaient surtout de leur environnement social direct. Aujourd'hui, les gens aiment également consulter les opinions objectives de personnes qui leur sont inconnues.

Ce besoin pressant d'informations dans la société n'a pas échappé aux nombreux commerçants qui se réfèrent au Web afin de savoir ce qui est dit à l'égard de leurs produits. De la même façon, la vie politique est de plus en plus dominée par le flux informationnel sur Internet et plus concrètement sur les réseaux sociaux. Depuis quelques années, les hommes politiques étendent leurs campagnes électorales jusqu'à Facebook et Twitter. Tel est le cas de Barack Obama, l'actuel président des États-Unis. Celui-ci a, en vue des élections présidentielles de 2012, utilisé Twitter comme l'un de ses principaux outils pour faire campagne.

Tous les récents développements dans le domaine d'échange d'informations et d'opinions ont donné un coup de volant aux applications informatiques conçues pour l'analyse et la détection de sentiments exprimés sur Internet. Présentée dans la littérature sous le nom de *« Opinion*

Mining » ou *« Sentiment Analysis »*, l'analyse des sentiments s'utilise entre autres pour la détection d'opinions sur des sites web et des réseaux sociaux, l'éclaircissement sur le comportement des consommateurs, la recommandation de produits et l'explication du résultat des élections. Elle consiste à rechercher des textes évaluatifs sur Internet tels que des critiques et des recommandations et à analyser de façon automatique ou manuelle les sentiments qui y sont exprimés afin de mieux comprendre l'opinion publique.

Il a déjà été démontré par des études antérieures que l'analyse des sentiments s'avère particulièrement intéressante pour ceux qui ont intérêt à connaître l'opinion publique, que ce soit pour des raisons personnelles, commerciales ou politiques. Ainsi, de nombreux systèmes autonomes ont déjà été développés pour l'analyse automatique des sentiments. Généralement, ces systèmes étaient entraînés aux textes évaluatifs traditionnels tels que les comptes rendus cinématographiques ou les critiques d'un livre. Toutefois, depuis quelques années se sont graduellement ajoutés à ces textes traditionnels les textes non traditionnels tels que les messages envoyés via les réseaux sociaux. Ceux-ci constituent une source précieuse d'opinions échangées parmi les multiples internautes.

4 Sentiment

C'est le jugement que porte un individu sur un objet ou un sujet, ce jugement étant caractérisé par une polarité et une intensité.

5 Opinion

Selon la définition citée dans [6], les opinions sont généralement des expressions subjectives qui décrivent les sentiments, les évaluations des gens vers des entités, des événements et leurs propriétés.

5.1 Opinion et fait

La différence entre *« fait »* et *« opinion »* est qu'un fait est quelque chose qui est empiriquement vrai et peut-être étayé par des preuves, tandis qu'une opinion est une croyance qui peut ou peut ne pas être sauvegardée avec un certain type de preuves. Une opinion est peut aussi être définie comme une déclaration subjective qui peut être le résultat d'une émotion ou une interprétation individuelle d'un fait. Par exemple, les différences biologiques entre mâles et femelles sont un fait, une préférence pour un sexe sur l'autre est une opinion.

5.2 Complexité de la notion d'opinion

Pour démontrer la complexité de la notion d'opinion, nous allons nous baser sur un ensemble de critiques [1] exprimées par des clients ont déjà visité un restaurant à la ville de Montréal(Canada). Les exemples sont les suivants, prises sous forme de captures écran :



FIGURE 1.1 – Commentaire collecté du site web restomontreal.ca

Dans les premières phrases *« Trés belle acceuil »*, *« nourriture excellente »* et *« Gens sympathique chaleureux »*, il est facile de déduire que l'orientation sémantique globale est positive parce que l'auteur a utilisé que des mots et des expressions positives *« trés belle »* , *« excellente »* et *« sympathique »*. Dans la dernière phrase l'expression *« un peu bruyant »* . Bien que l'adjectif *« bruyant »* ait une polarité négative parce que l'auteur a spécifié qu'il y a du bruit, la critique porte une opinion positive. Ce genre de phrases nécessite une analyse profonde pour extraire correctement leurs orientations globales.

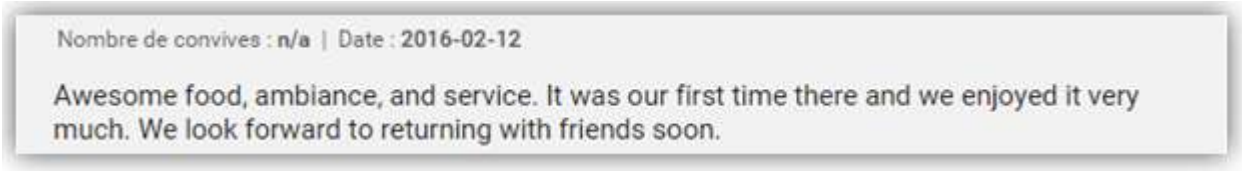


FIGURE 1.2 – Commentaire collecté du site web restomontreal.ca

Dans la deuxième critique, *« Awesome food, ambiance,and service.It was our first time here and we enjoyed it very much.We look forward to returning with friends soon »*. L'orientation sémantique globale est claire, toutes les expressions et les mots porteurs d'opinions sont positifs : *Awesome, ambiance, enjoyed* sont des porteurs d'opinions positives.



FIGURE 1.3 – Commentaire collecté du site web restomontreal.ca

La troisième critique, toutes les phrases *« Les sushis sont toujours aussi frais et délicieux »* , *« Le service un peu lent, mais à peine »* et *« Pour nous c'est notre favoris pour les sushis*

1. http://www.restomontreal.ca/restaurant-reviews/

» sont subjectives à cause de la présence des mots et d'expressions porteurs d'opinions : *frais, délicieux, un peu lent, favoris*. La présence des polarités opposées dans les deux dernières rend difficile à extraire l'orientation globale dans cette critique.



FIGURE 1.4 – Commentaire collecté du site web restomontreal.ca

Dans la quatrième critique, il est facile de déduire que l'orientation sémantique globale est positive parce que l'auteur a utilisé que des adjectifs positifs *« best »* et *« good »*

En général, les sentiments et la subjectivité sont très sensibles au contexte [7, 8] et dépendent du domaine. La dépendance du domaine est en partie une conséquence des changements de vocabulaire, par exemple, la même expression peut indiquer différents sentiments dans différents domaines. L'exemple suivant, cité dans [2], illustre la dépendance de domaine : Dans le domaine électronique, *« long »* peut avoir une polarité positive, comme dans le commentaire *« the battery life of Camera X is long»*. Tandis que dans le domaine informatique, il peut avoir une polarité négative, comme dans *« Program X takes a long time to complete »*. Dans [2], les auteurs ont mentionné aussi qu'il n'est pas facile de construire des lexiques de sentiments pour tous les domaines d'intérêt.

De plus sur Internet, chacun utilise son propre vocabulaire, ce qui rend la tâche plus difficile même s'il s'agit du même domaine. En plus, il est très difficile d'affecter correctement le poids pour des phrases de la critique. Très souvent nous avons une description très positive d'un restaurant, avec la meilleure nourriture, le bon accueil, la propreté, etc., mais une phrase comme *« Malgré tout ça j'ai quitté le restaurant avant de terminer mon repas »* peut changer toute l'opinion en négative bien que toutes les autres phrases soient positives.

Ces exemples montrent qu'il est encore impossible d'arriver à un cas idéal de notation des sentiments dans un texte écrit par les divers utilisateurs. Car cela ne respecte aucune règle et il est impossible de prévoir tous les cas possibles, en plus très souvent la même phrase peut être considérée comme positive pour une personne et négative pour une autre.

6 Fouille d'opinions et analyse des sentiments

Le terme fouille d'opinion *(ang.Opinion Mining)* apparaît dans un article de Dave [9] qui a été publié dans l'acte de conférence WWW 2003. Selon Dave, l'Opinion Mining devrait "traiter un ensemble de résultats de recherche pour un cas donné, générer une liste des attributs (qualité, caractéristiques, etc.) et agréger des avis sur chacun d'entre eux (mauvais, modéré, de bonne qualité). Toutefois, l'Opinion Mining a récemment été interprétée de manière plus générale pour

inclure de nombreux types d'analyse d'évaluation de texte [10].

Le terme "Analyse des Sentiments" est utilisé pour décrire l'analyse automatique de texte évaluatif et pour la recherche de valeur prédictive des jugements. Elle a été introduite dans les travaux de Das et Chen [11] en 2001 afin d'analyser des sentiments dans le cadre de l'économie de marché. Ensuite d'autres travaux sur l'analyse des sentiments ont été proposés par Turney [12] et Pang et al [13]. Depuis 2002, un nombre important d'articles citant l'Analyse des Sentiments ont vus le jour, ces travaux se concentrent sur la classification des commentaires et à leur polarité (positif ou négatif). Aujourd'hui, l'Opinion Mining et l'Analyse des Sentiments font partie du même domaine de recherche.

Donc, 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.

7 Processus de la fouille d'opinions

Le processus d'un système d'identification d'opinions comprend trois étapes : acquisition et analyse du corpus, étude de la pertinence des documents par rapport à un sujet, détection de l'opinion et ré-ordonnancement des documents.

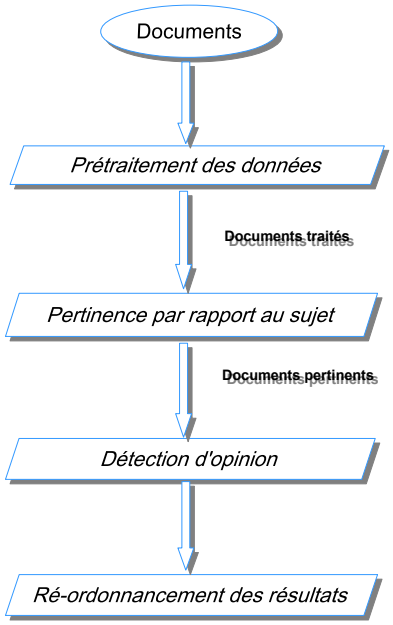


FIGURE 1.5 – Processus de la fouille d'opinions

7.1 Acquisition et prétraitement du corpus

Dans cette phase, les textes sont prétraités linguistiquement en éliminant les mots vides et les mots qui n'apportent aucune information importante, ainsi qu'une analyse lexicale pour enlever les mots qui ont un sens commun (redondant). Concernant les blogs la majorité des travaux se sont portés sur un nettoyage au niveau des balises HTML inutiles, ainsi que sur les documents qui ne sont pas écrits en langue traitée. Certaines techniques de fouille d'opinions [14] utilisent la structure des phrases afin d'identifier l'opinion. Dans cette étape, un étiquetage grammatical est fait (pour reconnaître l'adjectif, l'adverbe, le verbe, etc.), les grammaires de dépendances sont utilisées pour structurer la phrase de manière hiérarchique.

7.2 La pertinence par rapport au sujet

Cette étape consiste à étudier la pertinence des documents par rapport à un sujet donné. Parmi les méthodes les plus utilisées, nous pouvons citer la méthode probabiliste OKAPI et BM25 [15]. Les documents sont classés, et généralement les 1000 premiers documents les plus pertinents sont extraits, et sont utilisés pour l'étape suivante : la détection d'opinions.

7.3 La détection d'opinions

Plusieurs méthodes ont été utilisées pour la détection d'opinions. Leur but est de réordonner les documents pertinents selon un score d'opinion. Ainsi les documents qui contiennent le plus d'opinions sont classés parmi les premiers. Après ces étapes, une évaluation des résultats est faite. Les résultats sont souvent confrontés à la perception humaine de l'opinion. La comparaison est faite grâce à des mesures de similarité.

8 Domaines de la fouille d'opinions

Dans [4], l'auteur décrit la fouille d'opinion comme un sous-domaine de la fouille de textes qui consiste à analyser des textes afin d'en extraire des informations liées aux opinions et sentiments. L'une des tâches de la fouille d'opinions, appelée classification d'opinions, a pour objectif de classer les textes suivant l'opinion qu'ils expriment. Cette classification peut se faire sur deux classes (positif ou négatif), sur trois classes (positif, négatif ou neutre) ou sur plus de classes encore. Dans [3] et d'après les auteurs, la fouille d'opinion peut être divisée en trois sous-domaines.

8.1 Identification des textes d'opinions

Elle peut consister à identifier dans une collection textuelle les textes porteurs d'opinion, ou encore à localiser les passages porteurs d'opinion dans un texte. Plus précisément, nous parlons ici de classer les textes ou les parties de textes selon qu'ils sont objectifs ou subjectifs.

8.2 Résumé d'opinions

Consiste à rendre l'information rapidement et facilement accessible en mettant en avant les opinions exprimées et les cibles de ces opinions présentes dans un texte. Ce résumé peut être textuel (extraction des phrases ou expressions contenant les opinions), chiffré (pourcentage, note), graphique (histogramme) ou encore image (thermomètre, étoiles, pouce levé ou baissé...).

8.3 Classification d'opinions

A pour but d'attribuer une étiquette au texte selon l'opinion qu'il exprime. Nous considérons généralement les classes positive et négative, ou encore positive, négative et neutre.

9 Facteurs de difficulté de la fouille d'opinions

La fouille d'opinions est un problème de traitement automatique de textes. Dans cette section, nous distinguons ce problème avec la fouille de textes *« classique »* qui ne s'intéresse pas à l'opinion portée par les textes. Dans [16], l'auteur explique que le succès vienne à l'exploitation de plus en plus poussée de l'apprentissage artificiel dans le traitement de l'information, notamment la catégorisation de textes. Une autre explication est l'avènement de l'internet social et les intérêts socio-économique qu'il soulève.

Comme nous l'avons mentionné dans la définition , ce domaine est connu sous les noms de *opinion-mining*, *sentiment analysis* ou encore *subjectivity analysis*. Il s'intéresse au traitement automatique des opinions, des sentiments et de la subjectivité dans les textes. À titre de rappel, les informations textuelles sont, en général, soit des faits, soit des opinions. Tandis qu'un fait est une expression objective, une opinion est subjective, et décrit les sentiments ou l'évaluation d'une personne, à propos d'un objet d'intérêt ou de ses propriétés.

L'opinion-mining est certainement un domaine encore jeune mais son exploitation présente de nombreux problèmes techniques et malgré les améliorations depuis les débuts des recherches, ils n'ont toujours pas été résolus de manière convaincante. Parmi les facteurs qui rendent la fouille d'opinions difficile, nous trouvons :

9.1 Hétérogénéité des données

Les données textuelles d'opinion sont hétérogènes. D'une part, les opinions sont parfois dans des critiques formatées, parfois présentées dans un format libre. Dans les sites d'évaluation tels que Epinions [2] et Amazon [3], les critiques se présentent dans un format qui présente les avantages et les inconvénients de manière très synthétiques

2. Page web : www.epinions.com
3. Page web : www.amazon.com

D'autre part, dans un format libre, le langage utilisé pour exprimer une opinion rend parfois plus difficile l'extraction des opinions sur les blogs ou les forums que dans un article de journal en ligne [17], en effet un langage plus familier, des phrases grammaticalement incorrectes ou des expressions locales empêchent l'analyse correcte de ces opinions. Même dans un article ou un commentaire exprimé dans un langage soutenu, une faute d'orthographe ne va pas permettre la reconnaissance du mot et peut engendrer des résultats erronés surtout si la faute se situe sur un mot particulièrement porteur d'opinion. Le sarcasme [4], l'ironie [5], le cynisme ou le langage



FIGURE 1.6 – Ensemble de commentaires sur une vidéo retrouvés sur Youtube

figuré comme la métaphore qui influence le sens d'un texte sont complètement ignorés. Les figures 1.6 et 1.7 illustre un ensemble de commentaires exprimés dans un langage de sarcasme et d'ironie.



FIGURE 1.7 – Ensemble de commentaires retrouvés sur Facebook, exprimés en sarcasme

4. Le **sarcasme** désigne une moquerie ironique, une raillerie tournant en dérision une personne ou une situation.
5. L'**ironie** désigne un décalage entre le discours et la réalité, entre deux réalités ou plus généralement entre deux perspectives, qui produit de l'incongruité.

9.2 Dépendance du contexte

En général, les sentiments et la subjectivité sont très sensibles au contexte. Les techniques existantes utilisent des mots d'opinion tels que *« grand »*, *« étonnant »*, *« pauvre »*, *« bon »*, *« mauvais »*, etc., pour identifier l'orientation d'une opinion sur une caractéristique d'un produit.

Bien que les orientations de ces mots soient évidentes, les orientations de nombreux autres mots dépendent du contexte. Par exemple, le mot *« longue »* peut indiquer une opinion positive ou une opinion négative sur une caractéristique d'un produit dépendant de la caractéristique elle-même, comme indiqué dans l'exemple suivant :
« La durée de vie de la batterie de mon ordinateur portable est longue, c'est magnifique ».
« Mon ordinateur portable prend une longue durée pour démarrer, c'est un problème vraiment gênant ». Dans la première phrase, l'adjectif *« longue »* porte une polarité positive alors que dans la deuxième porte une polarité négative.

9.3 Dépendance du domaine

La dépendance du domaine est en partie une conséquence des changements de vocabulaire, par exemple, la même expression peut indiquer différents sentiments dans différents domaines [3, 18].

9.4 Ambiguïté subjective

La principale difficulté qu'apporte le langage est le caractère ambigu des mots. L'ambiguïté, en fouille d'opinion, se situe plus au niveau de la subjectivité, qu'au niveau sémantique. Dans [19], par exemple, selon son contexte, le mot *grand* peut tantôt être factuel (par exemple, pour désigner la taille d'un individu), tantôt exprime une opinion (par exemple, la taille d'un mobile).

10 Domaines d'application de la fouille d'opinions

L'importance de la détection d'opinion est présente dans plusieurs domaines, mais la plus grande application de l'Opinion Mining reste dans le monde du business, et du politique.

10.1 Marketing

10.1.1 Coté entreprises

L'opinion Mining permet au fournisseur d'un produit ou d'un service d'avoir plus de connaissances sur les consommateurs, pour anticiper leurs besoins et leurs attentes afin de tenter d'améliorer la qualité du produit/service et d'augmenter les profits. Les professionnels du marketing digital, du service client et de la communication cherchent aujourd'hui l'outil idéal leur permettant d'analyser automatiquement la teneur et la tonalité des conversations publiées sur leurs

marques sur le web [20]. Le domaine de la fouille d'opinion est devenu un enjeu majeur pour toute entreprise désireuse de mieux comprendre ce qui plait et déplait à ses clients [21].

10.1.2 Coté clients

Le client peut de son côté :

- Donner son opinion.
- S'inspirer des sentiments et opinions d'autres clients sur le produit auquel il s'intéresse et profiter ainsi d'une aide à la décision [17].
- Comparer les produits avant de les acquérir [21].
- Ne pas lire tous les commentaires concernant un produit donné, il suffit de voir le pourcentage positifs associé à ce produit.

10.2 Politique

Les documents portant des textes rapportant des débats politiques, (un corpus de réactions à des propositions de lois) sont de taille très importante. Ils contiennent environ dix fois plus de textes que celui des critiques d'un produit. Ce qui implique la variété énorme des jugements politiques dans ces articles. Ces derniers se retrouvent au coeur des débats avec l'avènement des médias sociaux (et plus spécialement sur le Web). Les acteurs politiques ont également suivi cette tendance, tel qu'avant de promulguer une nouvelle loi, les politiciens essayent de récolter l'avis des internautes sur cette loi. Il est intéressant de connaître aussi l'avis des internautes sur tel homme politique pour une élection présidentielle [22].

11 Terminologie

L'analyse des sentiments ou la fouille d'opinions est le traitement automatique des opinions, sentiments et émotions exprimées dans les textes. Nous utilisons le segment du texte suivant, il s'agit d'un commentaire sur une *« voiture »* pour comprendre les différents concepts utilisés (un nombre est associé à chaque phrase pour en faciliter la consultation) :

« (1) Hier, j'ai acheté une voiture (2) C'était vraiment cher (3) mais, elle était jolie (4) elle, m'a plait beaucoup (5) les chaises étaient confortables (6) Cependant, mon frère a dégouté ce modèle ». La question est : Qu'est-ce-que nous voulions fouiller ou extraire de ce texte ?

La première chose que nous pouvions noter est qu'il y a beaucoup d'opinions dans ce commentaire. Les phrases (3), (4) et (5) expriment des opinions positives, tandis que les phrases (2) et (6) expriment des opinions ou émotions négatives. Ensuite, nous remarquons que les opinions sont exprimées envers certaines caractéristiques ou objets. Les opinions dans les phrases (2), (3) et (5) sont sur la voiture comme un tout, et l'opinion exprimée dans la phrase (5) était sur les chaises. Finalement, nous pouvions aussi noter que le détenteur *(ang.holder)* de l'opinion dans

les phrases (2), (3) et (4) est l'auteur de commentaire *« moi »*, mais dans la phrase (6) c'est *« mon frère »*

Avec cet exemple, nous définissons d'une façon formelle l'analyse des sentiments et la fouille d'opinions. Généralement, les opinions peuvent être exprimées envers n'importe quoi : un produit, un service, un individu, une organisation, un événement, ou un sujet [1, 23].

Notons qu'un objet peut avoir un ensemble de composants (parties) et un ensemble d'attributs (propriétés), et chaque composant peut avoir à son tour, ses propres sous-composants et son ensemble d'attributs, et ainsi de suite. Ainsi, un objet peut-être décomposé hiérarchiquement selon la relation *« est une partie de »*.

11.1 Objet

Un objet O est une entité qui peut être un produit, une personne, un événement, une organisation ou un sujet. Il est associé à une paire $O\,(T, A)$, où T est une hiérarchie de composants (ou parties), sous-composants, et ainsi de suite, et A est un ensemble d'attributs de O. Chaque composant a son propre ensemble de sous-composants et attributs [3, 24, 25].

Exemple : Une marque particulière de téléphone cellulaire est un objet. Il a un ensemble de composants, par exemple, la batterie et l'écran, il a aussi un ensemble d'attributs, par exemple, la qualité de son, la taille et le poids. Le sous-composant « batterie » a aussi son ensemble d'attributs, par exemple, la durée de vie de la batterie, et la taille de la batterie, etc.

En se basant sur cette définition, un objet peut être représenté comme un arbre, une hiérarchie ou une taxonomie. La racine de l'arbre est l'objet lui-même. Chaque noeud non racine est un composant ou un sous-composant de l'objet. Chaque lien est une relation de type « est une partie de ». Chaque noeud est également associé à un ensemble d'attributs ou de propriétés. Une opinion peut être exprimée sur un noeud ou un attribut du noeud.

Suite à l'exemple précédent de la section 11, nous pouvons exprimer une opinion sur le téléphone cellulaire lui-même (le noeud racine), par exemple, *« Je n'aime pas ce téléphone »*, ou sur l'un de ses attributs, par exemple, *« La qualité de son de ce téléphone est mauvaise »*. De même, on peut aussi exprimer une opinion sur l'un des composants du téléphone ou tout autre attribut des composants.

En pratique, il est probablement trop complexe d'utiliser une représentation hiérarchique d'un objet et des opinions sur l'objet. Ainsi, pour des raisons de simplification et pour éviter toute éventuelle confusion, il est conseillé d'utiliser le terme *caractéristique* ou *trait* pour représenter à la fois les composants et les attributs [25].

Dans cette simplification, l'objet lui-même peut aussi être vu comme une caractéristique (mais une caractéristique spéciale), qui est la racine de l'arbre original. Un commentaire exprimé envers l'objet lui-même est appelée une opinion générale sur l'objet (par exemple, *« J'aime ma voiture »*). Une opinion sur une caractéristique spécifique est appelée opinion spécifique sur une caractéristique de l'objet, par exemple, *« L'écran tactile de l'iPhone est vraiment souple »*, où *« écran tactile »* est une caractéristique (composant) de l'iPhone.

L'utilisation du terme caractéristique est assez commune dans le domaine des produits parce que les gens utilisent souvent le terme caractéristique du produit. Toutefois, lorsque les objets sont des événements et des sujets, le terme caractéristique peut ne pas sembler naturel. En effet, dans certains autres domaines, les chercheurs utilisent également le terme *thème* [26] ou *aspect* [27, 28] pour signifier caractéristique. Dans notre mémoire nous utilisons le terme *trait*.

11.2 Modèle d'un objet

Un objet O est représenté par un ensemble fini de traits *(ang.features)*, $F = f_1, f_2, ..., f_n$, ce qui inclut l'objet lui-même comme un trait spécial. Chaque trait $f_i \in F$peut être exprimée avec un quelconque ensemble fini de mots ou de phrases $W_i = w_{i1}, w_{i2}, ..., w_{im}$, qui sont des synonymes du trait [25], ou indiqués par un quelconque ensemble fini d'indicateurs de traits $I_i = i_{i1}, i_{i2}, ..., i_{iq}$ du trait.

11.3 Opinion

Une opinion sur un trait f, est un point de vue positif ou négatif, une attitude, une émotion ou une évaluation sur f d'un détenteur d'opinion [23, 25].

11.4 Détenteur de l'opinion *(ang. Opinion holder)*

Le détenteur d'une opinion est la personne ou l'organisation qui exprime l'opinion.
Les détenteurs d'opinions sont également appelés sources d'opinion [29]. Dans le cas des critiques sur produits, les détenteurs d'opinions sont généralement les auteurs des commentaires postés. Les détenteurs d'opinions sont plus importants dans les articles des journaux parce qu'ils expriment souvent d'une manière explicite l'état de la personne ou l'organisation qui détient une opinion particulière [30, 31, 26]. Par exemple, le détenteur de l'opinion dans la phrase *« Ahmed a exprimé son désaccord sur le projet »* est *« Ahmed»*.

11.5 Orientation d'une opinion

L'orientation d'une opinion sur un trait f indique si l'opinion est positive, négative ou neutre [3, 23]. L'orientation de l'opinion est également connue sous le nom l'orientation du sentiment, la polarité de l'opinion, ou l'orientation sémantique [3, 25].

11.6 Opinion exprimée sur un trait

Une opinion exprimée sur un trait f d'un objet O, évaluée dans un document d, est un groupe de phrases consécutives en d, qui exprime une opinion positive ou négative sur f [3]. Il est possible qu'une séquence de phrases (au moins une) dans un document véhiculant une opinion exprime une opinion sur un objet ou un trait de l'objet. Il est également possible qu'une seule phrase exprime des opinions sur plus d'un trait, par exemple : *« La qualité de son de ce téléphone est bonne, mais la vie de la batterie est courte. »*
La plupart des recherches actuelles portent sur les phrases, c'est-à-dire, chaque passage est composé d'une seule phrase [3].

11.7 Trait explicite et trait implicite

Dans [6], si une 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, mais f, est implicite, alors f, est appelée trait implicite dans p.

Exemple :
« La durée de vie de la batterie » dans la phrase suivante le trait est explicite :
« La durée de vie de la batterie de ce téléphone est trop courte »
« La taille » est un trait implicite dans la phrase suivante comme il n'apparaît pas dans la phrase, mais il est implicite : *« Ce téléphone est trop grand »*.

Ici, *« grand »*, qui n'est pas un synonyme de *« taille »*, est appelé indicateur de trait. Beaucoup d'indicateurs de traits sont des adjectifs et des adverbes. Certains adjectifs et adverbes sont généraux et peuvent être utilisés pour modifier quoique ce soit, par exemple, *bon*, *mauvais*, et *grand*, mais beaucoup d'indicateurs indiquent les traits qu'ils sont susceptibles de modifier, par exemple, *belle* (apparence).

11.8 Modèle d'un document d'opinions

Un document général d'opinions d contient des opinions sur un ensemble d'objets $O_1, O_2, ..., O_Q$ à partir d'un ensemble de détenteurs d'opinions $h_1, h_2, ..., h_p$. Les opinions sur chaque objet o_j sont exprimées sur un sous-ensemble de traits de F_j de l'objet O_j. Une opinion peut être l'un des deux types suivants :

- Opinion directe
  Une opinion directe est un quintuple $(O_j, f_{jk}, oo_{ijkl}, h_i, t_l)$, où O_j est un objet, f_{jk} est un trait de l'objet O_j, oo_{ijkl} est l'orientation ou la polarité de l'opinion sur le trait f_{jk} de l'objet O_j, h_i est le détenteur de l'opinion et t_l est le temps où l'opinion est exprimée par h_i. L'orientation de l'opinion oo_{ijkl} peut être positive, négative ou neutre (ou mesurée sur une échelle plus fine pour exprimer des opinions différentes de différentes ensités (forces) [32]. Pour le trait f_{jk} commentée par le détenteur d'opinion h_i, il / elle choisit un mot ou une expression de l'ensemble correspondant à des synonymes W_{jk}, ou un mot ou une phrase de l'ensemble des indicateurs du trait I_{jk} pour décrire ce dernier, puis exprime une opinion positive, négative ou neutre sur le trait.
- Opinion comparative
  Une opinion comparative exprime une relation de similitude ou de différence entre deux ou plusieurs objets. Une opinion comparative est généralement exprimée en utilisant la forme comparative ou superlative d'un adjectif ou d'un adverbe, mais pas toujours. Ce modèle de texte d'opinions couvre l'essentiel mais pas toutes les informations intéressantes ou tous les cas possibles. Par exemple, il ne couvre pas la situation décrite dans la phrase suivante : *« Le viseur et les lentilles de cette caméra sont trop proches »*, qui exprime une opinion négative sur la distance entre les deux composantes.

  Dans les opinions directes, il y a en fait deux principaux sous-types :

  — Dans le premier sous-type, les opinions sont exprimées directement sur un objet ou sur les traits de l'objet, par exemple, *«La qualité de son de ce téléphone est bonne.»*.

  — Dans le deuxième sous-type, les opinions sont exprimées sur un objet en fonction de ses effets sur certains autres objets. Ce sous-type se produit souvent dans le domaine médical où les patients expriment des opinions sur les médicaments ou décrivent leurs effets secondaires. Par exemple, la phrase *« Après avoir pris ce médicament, mon genou gauche me sentait bien »*, décrit un effet souhaitable de médicament sur le genou, et implique donc une opinion positive sur le médicament.

11.9 Émotions

Les émotions sont nos sentiments subjectifs et les pensées [23, 25]. Les émotions ont été étudiées dans de nombreux domaines, par exemple, la psychologie, la philosophie, la sociologie, la biologie, etc. En se basant sur [33], les gens ont 6 types d'émotions primaires, i.e., l'amour, la joie, la surprise, la colère, la tristesse et la peur , qui peuvent être divisées en de nombreuses émotions secondaires et tertiaires comme indiqué dans la figure 1.8. Chaque émotion peut aussi avoir des intensités différentes. Les forces d'opinions sont étroitement liées à l'intensité de certaines émotions, par exemple, la joie et la colère. Toutefois, les concepts d'émotions et d'opinions ne sont pas équivalents, bien qu'ils aient une grande intersection.

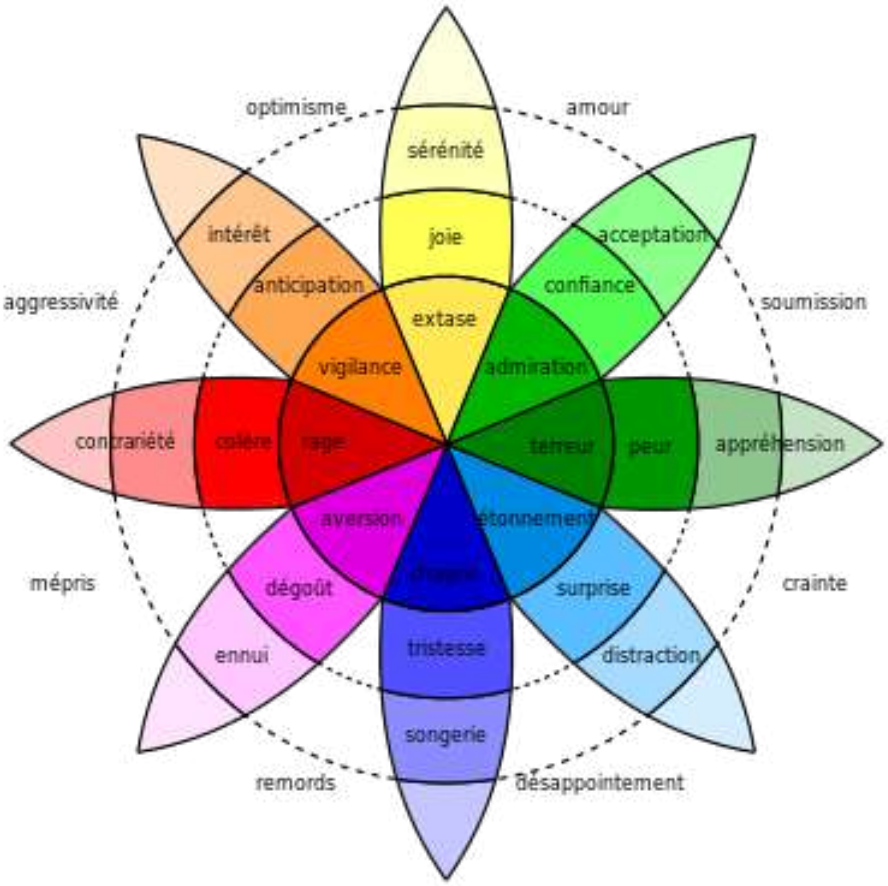


FIGURE 1.8 – Exemple des émotions de base et secondaires

11.10 Subjectivité

Soit l'exemple suivant :
« (1) Samedi dernier, j'ai acheté un téléphone Nokia et mon collègue a acheté un téléphone Samsung. (2) Nous nous sommes appelés quand nous sommes arrivés à la maison. (3) Le son sur mon téléphone n'était pas si clair, pire que mon ancien téléphone. (4) La caméra était bonne. (5) Mon collègue était très heureux avec son téléphone. (6) Je voulais un téléphone avec une bonne qualité de son. (7) Donc mon achat a été une vraie déception. (8) hier, j'ai retourné le téléphone. »

L'exemple ci-dessus a révélé un autre problème appelé la subjectivité. C'est, dans un document, quelques phrases expriment des opinions et d'autres pas. Par exemple, les phrases (1), (2), (6) et (8) n'expriment aucune opinion. Le problème de la subjectivité a été largement étudié dans [34, 29, 35, 32].

11.11 Phrase subjective

Une phrase objective exprime quelques informations factuelles sur le monde, alors qu'une phrase subjective exprime des sentiments ou des croyances personnelles. Par exemple, dans l'exemple cité dans la sous-section précédente, les phrases (1), (2) et (8) sont des phrases objectives, tandis que toutes les autres phrases sont des phrases subjectives.

Les expressions subjectives prennent plusieurs formes, par exemple, les opinions, les allégations, les désirs, les croyances, les soupçons et des spéculations [36, 37]. Ainsi, une phrase subjective peut ne pas contenir une opinion. Par exemple, la phrase (6) de l'exemple de la sous-section précédente, est subjective, mais elle n'exprime pas une opinion positive ou négative sur n'importe quel téléphone spécifique. De même, nous devons également noter que pas toutes les phrases objectives ne contiennent aucune opinion, la deuxième phrase de l'exemple cité dans la sous-section suivante montre ceci.

11.12 Opinion explicite et opinion implicite

Une opinion explicite sur un trait f, est une opinion exprimée explicitement sur f dans une phrase subjective. Une opinion implicite sur le trait f, est une opinion sur f implicite dans une phrase objective.

Exemple :

La phrase suivante exprime une opinion positive explicite :

« La qualité de son de ce téléphone est incroyable. »

La phrase suivante exprime une opinion négative implicite :

« L'afficheur s'est cassé en deux jours. »

Bien que cette phrase énonce un fait objectif, il indique implicitement une opinion négative sur l'afficheur. En fait, la phrase (8) dans l'exemple de la section précédente, peut aussi être déclarée à impliquer une opinion négative. En général, les phrases objectives qui impliquent des opinions positives ou négatives indiquent souvent les raisons de l'opinion.

11.13 Phrase d'opinions

Une phrase d'opinions est une phrase qui exprime explicitement ou implicitement des opinions positives ou négatives. Elle peut être une phrase subjective ou objective. Comme nous pouvons le voir, les concepts phrase d'opinions et phrase subjective ne sont pas les mêmes, bien que les phrases d'opinions sont souvent un sous-ensemble de phrases subjectives. Les approches pour les identifier sont similaires. Ainsi, pour la simplicité de présentation, nous utilisons les deux termes indifféremment. La tâche de déterminer si une phrase est objective ou subjective est appelé classification de subjectivité.

12 Conclusion

Ce chapitre présente brièvement le domaine de la fouille d'opinions, ses problématiques, et la terminologie utilisée. Les facteurs qui rendent difficile la fouille d'opinions comme l'hétérogénéité des données, la dépendance du domaine et du contexte, ainsi l'ambiguïté subjective, offrent un défi de traitement automatique et des approches qui s'adaptent au niveau de granularité pour gagner en précision.

La classification des opinions est l'un des sujets les plus largement étudiés dans le domaine de fouille d'opinions, dont une grande importance a été accordée à l'étude des techniques de classification qui seront présentées dans le chapitre suivant du présent mémoire.

Chapitre 2

Classification d'opinions

Sommaire

1 Introduction 30

2 Classification d'opinions 30

2.1 Objectifs et définition 31

2.2 Complexité et caractéristiques de la tâche 32

3 Classification d'opinions au niveau des documents 32

3.1 Classification d'opinions non supervisé 33

3.2 Classification d'opinion supervisée 39

3.3 Classification d'opinion semi supervisée (approches hybrides) 52

4 Classification d'opinions au niveau des phrases 54

5 Classification d'opinions au niveau des traits linguistiques 55

6 Classification d'opinions avec croisement de langues 56

7 Classification d'opinions avec croisement de domaines 58

8 Mesures d'évaluations 59

9 Fouille d'opinions en langue arabe 62

10 Travaux connexes à l'identification du genre 63

11 Conclusion 63

1 Introduction

L'analyse des textes en matière d'étude des sentiments ou d'opinions n'est pas récente[38, 39]. La première apparition du terme Opinion-Mining, signifiant littéralement, fouille d'opinions, était en 2003 [9]. Depuis cette année-là, le domaine est devenu très actif avec différents défis qui se sont créés et des centaines d'articles sur le sujet ont été publiés.
Dans [3], et d'après l'auteur, la fouille d'opinions peut être divisée en trois sous-domaines :

- l'identification des textes d'opinion, qui peut consister à identifier dans une collection textuelle les textes porteurs d'opinion, ou encore à localiser les passages porteurs d'opinion dans un texte. Plus précisément, on parle ici de classer les textes ou les parties de textes selon qu'ils sont objectifs ou subjectifs ;
- le résumé d'opinion, qui consiste à rendre l'information rapidement et facilement accessible en mettant en avant les opinions exprimées et les cibles de ces opinions présentes dans un texte ;
- la classification d'opinions, qui a pour but d'attribuer une étiquette au texte selon l'opinion qu'il exprime. On considère généralement les classes positive et négative, ou encore positive, négative et neutre.

Ce sont clairement l'identification des textes d'opinion et la classification d'opinions qui sont les plus étudiées dans la littérature et sont également sujets à différents défis [40, 41].

Dans ce chapitre, nous nous intéresserons en premier axe à la tâche de classification d'opinions, et en deuxième axe, aux travaux de recherche qui focalisent sur la langue arabe, dont beaucoup de travaux ont été effectués sur le sujet, et trois grandes catégories de méthodes peuvent être mises en avant : les approches basées sur la linguistique, les approches basées sur l'apprentissage automatique et les approches hybrides.

2 Classification d'opinions

La façon dont l'analyse des sentiments est fait est grandement tributaire du niveau de granularité que nous voulons analyser. Différents niveaux de granularité nécessitent différents outils et méthodes, et quelques granularités sont plus réalisables dans certains contextes que d'autres.

Le niveau de granularité le plus élevé est le niveau du document *(ang. Document level)*. A ce niveau, nous nous préoccupons de déterminer le sentiment de chaque document dans son ensemble [13]. Pour avoir une valeur applicable à ce niveau de granularité, nous voulons généralement supposer que chaque document exprime le sentiment sur un seul sujet. Les corpus avec des documents tels que les évaluations des clients sont très appropriés pour l'analyse à ce niveau de granularité.

Un niveau plus détaillé de granularité est le niveau de la phrase *(ang. Sentence level)*, où les méthodes d'interprétation qui analyse des sentiments à ce niveau tente de déterminer le sentiment de phrases isolées.

Enfin, le meilleur niveau de granularité est le niveau de traits *(ang. Aspect level ou Feature level)*. Afin d'analyser les sentiments au niveau d'un trait, il faut créer un modèle plus holistique du sentiment où y compris la cible de sentiment exprimé est nécessaire. Ceci, bien sûr, nécessite le calcul et la modélisation des informations linguistiques plus avancées. Les systèmes performants d'analyse à ce niveau sont des outils très utiles pour effectuer des résumés de sentiment structurés sur des traits, transformant le texte non structuré en données structurées. Cet aspect de ce niveau est très précieux pour l'analyse qualitative et quantitative différente [42].

Dans cette section, nous décrivons les objectifs de la classification d'opinions, ainsi que ses intérêts, et les grandes approches de classification utilisées en fouille d'opinions.

2.1 Objectifs et définition

L'opinion peut être définie comme l'expression des sentiments d'une personne envers une entité [6]. L'opinion est subjective et peut être décrite avec certains attributs. L'attribut d'opinion le plus étudié est la polarité (positive, négative, et éventuellement neutre) qui définit si l'opinion est favorable ou défavorable. D'autres attributs sont l'intensité de l'opinion et le degré de subjectivité.

La classification de polarité d'opinion consiste donc à déterminer si un texte exprime des opinions positives ou négatives, quelque soit le sujet du texte ou les traits qui y sont discutées. Cette tâche porte sur des textes subjectifs, c'est à dire des textes contenant des opinions sur un sujet ou un objet tel que un film, une voiture, un téléphone portable ou encore une personnalité. Une définition plus formelle citée dans [6], de la classification d'opinion est donnée comme suit :
Soit C un ensemble de classes ordonnées représentant chacune un degré d'opinion et D un ensemble de textes subjectifs. La classification d'opinions consiste alors à trouver, pour tout d $\in$ *D, les couples (d, c) tel que c* $\in$ *C.*

Ce type de classification se fait généralement sur deux classes, "contient une opinion positive" ou "contient une opinion négative". Nous parlons alors de classification binaire. Elle peut également se faire sur trois classes en ajoutant une classe de neutralité destinée aux textes porteurs des opinions neutres. Nous pouvons également envisager une classification sur un nombre de classes supérieur à trois afin de mieux préciser l'intensité de l'opinion, mais cela reste assez rare dans le domaine.

2.2 Complexité et caractéristiques de la tâche

La classification d'opinions pourrait se comparer à une classification de textes comme les autres : *« étant donné les mots présents dans le texte, j'en déduis une classe »*. Par exemple, pour la classification par thème, il suffit de trouver un certain nombre de mots en lien, par exemple, au sport, à la politique ou encore au cinéma pour découvrir la classe du document. Seulement, en ce qui concerne les opinions, elles sont rarement exprimées par un seul mot. De plus, les discours contenant des opinions sont par définition des discours subjectifs et l'interprétation de la subjectivité est une tâche délicate. En effet, les opinions sont généralement exprimées par des verbes et des adjectifs qualificatifs qui possèdent une certaine hiérarchie suivant le degré de l'opinion qu'ils expriment. De plus, il existe la possibilité d'ajouter des adverbes afin de modifier la *« force »* de l'adjectif ou du verbe.

Outre les problèmes d'intensité d'opinion liés aux adjectifs et verbes, les opinions et les sentiments peuvent être exprimés de plusieurs façons. Ils peuvent par exemple être exprimés à l'aide des termes relevant de l'émotion. Ces termes liés à l'émotion telle que *« triste »*, ou *« effrayant »* sont difficiles à interpréter du point de vue de l'opinion si l'on ne connaît pas les goûts de l'auteur. Les goûts de l'auteur ont donc une importance mais le sujet également.

Dans le cas d'un film par exemple, le commentaire *« Ce film est à mourir de rire »* peut exprimer un avis très positif si l'objet visé est une comédie mais pas nécessairement s'il s'agit d'un film d'horreur ou d'un drame. Les avis peuvent également être exprimés sans employer un seul mot directement lié à l'opinion comme dans le commentaire *« courrez vite, acheter le livre »* [43]. Le traitement des négations semble également être une grosse problématique de la classification d'opinion [3].

3 Classification d'opinions au niveau des documents

La fouille de textes est l'extraction de connaissances dans des textes en langage naturel. La classification d'opinions est une tâche particulière de la fouille de textes et répond à un problème de traitement automatique de textes qui s'intéresse au traitement des opinions ou des sentiments et de la subjectivité dans les textes. Cette tâche peut être le sujet le plus largement étudié [34, 29, 35, 32, 36, 37], qui classifie un document d'opinions (par exemple, un commentaire sur un produit) pour définir s'il exprime une opinion positive ou négative.

La tâche est aussi connue sous le nom de classification de sentiments au niveau des documents *(ang. Document-level sentiment classification)*, car il considère l'ensemble des documents comme l'unité d'information de base [3]. Naturellement, elle peut également être appliquée à des phrases individuelles.

3.1 Classification d'opinions non supervisé

Puisque les mots de sentiment et les phrases sont souvent le facteur dominant pour la classification de sentiment, il n'est difficile d'imaginer de les utiliser pour la classification des sentiments d'une manière non supervisée. Nous discutons deux méthodes ici. L'une, basée sur la méthode de Turney (2002), effectue la classification en utilisant des patrons *(ang.Patterns)* syntaxiques fixes qui sont susceptibles d'exprimer des opinions. L'autre est basée sur un lexique de sentiment, qui est une liste de mots et expressions de sentiment positif et négatif.

3.1.1 Approches basée sur les patrons syntaxiques

Dans Turney (2002) [12], chaque patron syntaxique est une séquence d'étiquettes de POS tags avec certaines contraintes, l'algorithme consiste en trois étapes :

Etape 1 : Deux mots consécutifs sont extraites si leurs POS tags sont conformes à l'un des patrons dans le tableau 3.1. Par exemple, le patron 2 signifie que deux mots consécutifs sont extraits si le premier mot est un adverbe, le second mot est un adjectif, et le troisième mot (non extrait) n'est pas un nom. A titre d'exemple, dans la phrase «*This piano produces beautiful sounds*», «*beautiful sounds*» est extrait comme il satisfait le premier patron. La raison pour laquelle ces patronss sont utilisés est que JJ, RB, RBR et RBS [1] sont des mots expriment souvent des sentiments. Les noms ou les verbes agissent comme des contextes parce que dans différents contextes JJ, RB, RBR et RBS peut exprimer des sentiments différents. Par exemple, l'adjectif (JJ) «*unpredictable*» peut avoir un sentiment négatif dans une critique de voiture comme dans «*unpredictable plot*» mais il pourrait avoir un sentiment positif dans une critique de film comme dans «*unpredictable steering*»

	Premier mot	**Deuxième mot**	**Troisième mot(non extrait)**
1	JJ	NN ou NNS	Rien
2	RB, RBR, ou RBS	JJ	ni NN ni NNS
3	JJ	JJ	ni NN ni NNS
4	NN ou NNS	JJ	ni NN ni NNS
5	RB, RBR, ou RBS	VB, VBD, VBN, ou VBG	Rien

TABLE 3.1 – Patrons POS tags pour extraire des phrases de deux mots.

Etape 2 : Il estime l'orientation du sentiment *(ang. Sentiment Orientation)* des phrases extraites en utilisant la mesure PMI *(Pointwise Mutual Information)* :

$$PMI\left(terme_1, terme_2\right) = \log_2\left(\frac{Pr\left(terme_1 \wedge terme_2\right)}{Pr\left(terme_1\right)\left(terme_2\right)}\right) \tag{2.1}$$

PMI mesure le degré de dépendance statistique entre les deux termes. Ici, $Pr\left(terme_1 \wedge terme_2\right)$

1. **JJ :**Adjectif, **RB :**adverbe, **RBR :** adverbe comparatif, **RBS :**adverbe superlatif.

est la probabilité de co-occurrence réelle, et $Pr\,(terme_1)\,(terme_2)$ est la probabilité de co-occurrence des deux termes si elles sont statistiquement indépendantes. L'orientation de sentiment (SO) d'une phrase est calculée en fonction de son association avec le mot de référence positive *"excellent"* et le mot de référence négative *"poor"* :

$$SO\,(phrase) = PMI\,(phrase, "excellent") - PMI\,(phrase, "poor") \tag{2.2}$$

Etape 3 : Etant donné une critique, l'algorithme calcule la moyenne SO de toutes les phrases dans la critique et classe la critique comme positive si la moyenne SO est positive et négative sinon.
Une autre approche non supervisée est la méthode basée sur le lexique, qui utilise un dictionnaire de mots de sentiment et des phrases avec leurs orientations associés, et incorpore l'intensification et la négation pour calculer un score de sentiment pour chaque document [44]. Cette méthode a été utilisée à l'origine dans la classification d'opinion au niveau de phrase et de trait *(ang. feature)* [45, 46, 26].

3.1.2 Approches linguistiques basées sur les lexiques

Les mots de sentiment positif sont utilisés pour exprimer certains états ou qualités désirées tandis que les mots de sentiment négatifs sont utilisés pour exprimer certains états ou qualités indésirables. Exemples de mots de sentiment positifs sont ,*beautiful*, *wonderful*, et *amazing*. Exemples de mots de sentiment négatifs sont *bad*,*awful*, et *poor*.

En dehors des mots individuels, il y a aussi des expressions et des idiomes de sentiment par exemple, *cost someone an arm and a leg*. Collectivement, ils sont appelés lexiques de sentiment (ou lexiques d'opinion). Pour faciliter la présentation, à partir de maintenant, quand nous disons mots de sentiment, nous entendons les mots et expressions individuelles.

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

A titre d'exemple [47], les auteurs décrivent un système appelé Opinion Observer permet de comparer des produits concurrents en utilisant les commentaires écrits par les internautes.

Pour cela, ils ont une liste prédéfinie de termes désignant des caractéristiques de produits. Lorsque l'une de ces caractéristiques est présente dans un texte, le système extrait les adjectifs proches dans la phrase. Ces adjectifs sont ensuite comparés aux adjectifs présents dans leur lexique d'opinions et ainsi, une polarité est attribuée à la caractéristique du produit.

- Construction des lexiques d'opinion
  Les chercheurs ont proposé de nombreuses approches pour construire des lexiques d'opinions (sentiments). Trois approches principales sont : l'approche manuelle, l'approche basée sur les dictionnaires, et l'approche basée sur les corpus [48, 49] :

  — La méthode manuelle
  Consiste à remplir le lexique de mots d'opinions sans l'aide d'outils particuliers, la sélection des mots porteurs d'opinion et le choix de leur polarité se font donc uniquement par l'expertise humaine. Cette dernière demande un effort important en terme du temps. Généralement, elle n'est pas utilisé seule mais combinée avec des approches automatisées comme une vérification finale.

  Nous pouvons supposer qu'une part non négligeable de subjectivité peut alors entrer en jeu et peut entraîner certaines erreurs de classification. Quelque soit la méthode de construction des lexiques, une première étape de classification manuelle est nécessaire. En effet, une première liste de mots et d'expressions doit être identifiée. Ces mots sont appelés les germes (ang. seed words). Cet ensemble de mots est ensuite utilisé afin de découvrir, répertorier et classer d'autres mots et expressions porteurs d'opinions.

  — La méthode basée sur les corpus
  L'une des solutions permettant d'agrémenter cet ensemble de mots est donc l'utilisation de corpus de textes qui contribue à résoudre le problème de trouver des mots de sentiment avec des orientations spécifiques au contexte. Ses méthodes dépendent de schémas syntaxiques ou des schémas qui se produisent ensemble avec une liste de germes de mots de sentiment pour trouver d'autres mots de sentiment dans un grand corpus.

  L'approche basée sur les corpus a été appliquée à deux scénarios principaux : (1) donner une liste de germes des mots de sentiment*(ang. seed words)* connus (souvent à usage général), découvrir d'autres termes de sentiment et leurs orientations à partir d'un corpus de domaine, et (2) adapter un lexique de sentiment à usage général à un nouveau en utilisant un corpus de domaine pour les applications d'analyse de sentiment dans le domaine. Cependant, la question est plus complexe que la simple construction d'un lexique de sentiment spécifique d'un domaine parce que, dans le même domaine le même mot peut être positif dans un contexte, mais négatif dans un autre.

Dans [12], l'auteur propose la méthode suivante : afin de déterminer la polarité de mots ou expressions non classés, il compte le nombre de fois où ces mots ou expressions apparaissent dans le corpus à côté de mots ou expressions présents dans la liste de germes. C'est ce qu'on appelle, chercher les "cooccurrences" de mots. Un mot apparaissant plus souvent à côté de mots positifs sera donc classé dans la catégorie positif et inversement.

Dans [50], les auteurs ont proposé une méthode similaire, mis à part qu'ils utilisent la probabilité qu'un mot non classé soit proche d'un mot classé afin de mesurer la force de l'orientation du premier nommé. D'autres méthodes [51, 52] utilisent également cette hypothèse dans le but de compléter les lexiques d'opinion : deux mots ou groupes de mots ayant un fort taux d'apparitions communes sont supposés posséder une forte proximité sémantique.

Parmi ces méthodes qui ont été représentés par Hatzivassiloglou et McKeown [53] et [54]. Ils ont commencé avec une liste de germes d'adjectifs d'opinion, et les ont utilisés avec un ensemble de contraintes linguistiques pour identifier les adjectifs d'opinion supplémentaires et leurs orientations. Les contraintes sont des conjonctions de coordination présentes entre un mot déjà classé et un mot non classé comme AND, OR, BUT, EITHER-OR, et NEITHER-NOR . Par exemple la conjonction AND dit que les adjectifs conjoints ont généralement la même orientation. Cette idée est appelée cohérence de sentiment, qui n'est pas toujours cohérente en pratique.Il y a aussi des expressions adversative tels que, BUT, HOWEVER, qui sont indiqués comme des changements d'opinion. Par exemple, si la conjonction BUT sépare un mot classé positif et un mot non classé, alors le mot non classé sera considéré comme étant négatif.

Cette approche peut également être utilisé pour construire un lexique de sentiment à un usage général si un corpus très grand et très diversifié est disponible, l'approche basée sur les dictionnaires est généralement plus efficace pour cela, parce qu'un dictionnaire a tous les mots.

— La méthode basée sur les dictionnaires
C'est la solution la plus utilisée. Elle consiste à utiliser des dictionnaires de synonymes et antonymes existants tels que WordNet [55, 56] afin de déterminer l'orientation sémantique de nouveaux mots. Par exemple, Liu [46, 49] utilisent ce dictionnaire afin de prédire l'orientation sémantique des adjectifs. Dans WordNet, les mots sont organisés sous forme d'arbres, comme indiqué dans la figure 2.1 :

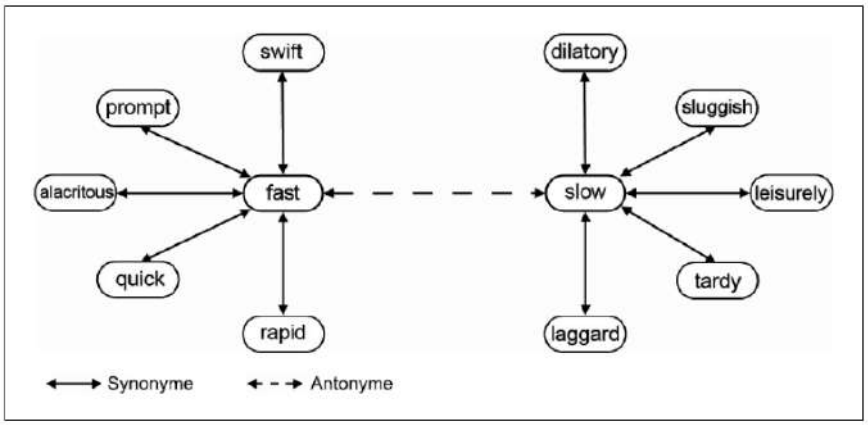


FIGURE 2.1 – Exemple d'arbre de synonymes et d'antonymes présents dans WordNet

Afin de déterminer la polarité d'un mot, ils traversent les arbres de synonymes et d'antonymes du mot et, s'ils trouvent un mot déjà classé parmi les synonymes, ils affectent la même polarité au mot étudié, ou bien la polarité opposée s'ils trouvent un mot déjà classé parmi les antonymes. S'ils ne croisent aucun mot déjà classé, ils réitèrent l'expérience en partant de tous les synonymes et antonymes, et ce jusqu'à rencontrer un mot d'orientation sémantique connue. Cette méthode peut toutefois entrainer un certain nombre d'erreurs car un grand nombre de mots peuvent avoir différentes significations.

[46, 26] a présenté la stratégie principale de l'approche basée sur un dictionnaire. Un petit ensemble de mots d'opinion sont collectées manuellement avec des orientations connues. Ensuite, cet ensemble est cultivé par la recherche dans le corpus bien connu WordNet [56] ou thesaurus [57] pour leurs synonymes et antonymes. Les mots nouvellement trouvés sont ajoutés à la liste des germes puis la prochaine itération commence. Le processus itératif arrête quand aucun de nouveaux mots sont trouvés. Après la terminaison du processus , une inspection manuelle peut être effectuée pour supprimer ou corriger les erreurs.

Des analyses plus poussées sont également faites. Afin de mesurer plus précisément la force de l'opinion exprimée dans une phrase, un moyen utilisé est l'extraction des adverbes associés aux adjectifs. Pour ce faire, Benamara et al. [58] proposent une classification des adverbes en cinq catégories : les adverbes d'affirmation, de doute, de forte intensité, de faible intensité et les adverbes de négation et minimiseurs. Un système d'attribution de points en fonction de la catégorie de l'adverbe permet de calculer la force exprimée par le couple adverbe-adjectif. Le tableau 3.2 présente quelques exemples de mots classés dans les catégories définies par Benamara et al. [58].

Classe	Mots
Adverbes d'affirmation	Absolutely, entirely, fully, certainly, fairly, exactly, totally, enough...
Adverbes de doute	Even, possibly, roughly, apparently, seemingly...
Adverbes de forte intensité	So, really, very, pretty, highly, extremely, much, well, too, quite...
Adverbes de faible intensité	Only, a little, almost, a bit, little, rather, nearly, barely, scarcely, weakly, slightly...
Négation et minimiseurs	Not, never, less, no...

TABLE 3.2 – Exemple de catégories d'adverbes

Toutes ces catégories d'adverbes ne sont pas toujours prises en compte car elles n'ont pas la même importance au niveau de la prédiction de note. Les négations paraissent logiquement être des termes importants à détecter, en plus des adjectifs et des verbes, car ils permettent d'inverser la polarité d'une phrase. Pour traiter cet aspect, Das et Chen [11] proposent par exemple d'ajouter des mots dans le dictionnaire d'opinion comme "like-NOT" qui sont utilisés lors de la détection d'un couple like-négation. Les négations peuvent alors être *not* , *don't*, *didn't* ou *never* pour le cas de l'anglais et *ne pas* pour le français. Mais le problème de la détection de la négation reste un problème très ouvert, les méthodes existantes n'étant pas réellement convaincantes.

En effet, l'expression de la négation peut être faite sans l'utilisation de ce type de mots. Les expressions telles que « *je suis contre* » ou encore « *je m'oppose* » peuvent également permettre d'inverser la polarité du reste de la phrase. La difficulté est également due aux différentes façons d'utiliser la négation comme le sarcasme ou l'ironie. L'interprétation de la négation nécessite alors une analyse syntaxique qui est un traitement très coûteux en temps de calcul et pas forcément très efficace suivant la qualité du texte analysé.

L'approche basée sur les dictionnaire a un inconvénient majeur qui est l'incapacité de trouver des mots de sentiment avec des orientations des domaines et des contextes spécifiques.

Qiu et il [59] ont utilisés l'approche basée sur les dictionnaires pour identifier les phrases de sentiment dans la publicité contextuelle.Ils ont proposé une stratégie publicitaire pour améliorer la pertinence des annonces et l'expérience utilisateur.Ils ont utilisé l'analyse syntaxique et les dictionnaires de sentiment et proposé une approche à base de règles pour lutter contre l'extraction de mots sujet et l'identification de l'attitude des consommateurs dans l'extraction de mot-clé de la publicité.Ils ont travaillé sur les forums web à partir de [2]. Leurs résultats ont démontré l'effica-

2. Page web : automotiveforums.com

cité de l'approche proposée sur l'extraction de mot-clé de la publicité et la sélection des annonces.

- Classification des textes grâce aux lexiques
  C'est la dernière étape qui consiste à déterminer la polarité d'une phrase à l'aide des mots porteurs d'opinion qui sont répertoriés dans les lexiques. La solution la plus simple consiste à compter le nombre de mots positifs et le nombre de mots négatifs présents. S'il y a une majorité de termes positifs, la phrase est déclarée positive. A l'inverse, si les mots négatifs sont les plus nombreux, la phrase est déclarée négative. Les phrases possédant autant de mots négatifs que de mots positifs peuvent être déclarées neutres [50], ou encore, la polarité de la phrase peut dépendre du dernier mot d'opinion parcouru [46]. On peut encore extraire plusieurs opinions dans une même phrase et les associer aux caractéristiques discutées [46].

3.2 Classification d'opinion supervisée

3.2.1 Approches statistiques basée sur l'apprentissage automatique

Dans ce type d'approches, les mots sont généralement considérés comme des variables équivalentes. L'aspect sémantique n'est alors pas pris en compte. Les méthodes les plus utilisées pour la classification d'opinions sont les méthodes de classification supervisée. Ce type de méthodes consiste à construire un modèle de classification à l'aide d'exemples. Des exemples sont des données dont nous connaissons déjà la classe. Nous parlons dans ce cas de données classées ou étiquetées. Voici une définition plus formelle du problème de la classification supervisée :

Soit X un ensemble de données, Y un ensemble d'étiquettes (ou classes) et D un ensemble des représentations des données. Soit $d : X \rightarrow D$*, une fonction qui associe à chaque donnée* $x \in X$ *une représentation* $d(x) \in D$ *et* $S \subset D \times Y$ *un ensemble de données étiquetées* $(d(x), y)$ *avec* $y \in Y$*. La classification supervisée consiste à construire un classificateur en s'appuyant sur l'ensemble S, qui permette de prédire la classe de toute nouvelle donnée* $x \in X$*, représentée par* $d(x) \in D$*.*

D'après cette définition, quatre éléments distincts entrent en jeu dans la classification supervisée :

- La représentation des données (l'ensemble D) ;
- Les étiquettes ou classes de prédiction (l'ensemble Y) ;
- Les données étiquetées, qui constituent le corpus d'apprentissage (l'ensemble S) ;
- Le classificateur ou le prédicteur.

- Corpus d'apprentissage
  Tout d'abord, il est nécessaire de posséder des exemples (données étiquetées) afin de

construire le "corpus d'apprentissage". Ce corpus ayant un impact direct sur l'apprentissage du modèle et par conséquent, sur les résultats de la classification, il est nécessaire que les exemples soient les plus représentatifs possibles de l'ensemble des données. En classification d'opinions, ou plus généralement en classification de textes, les corpus étudiés sont assez restreints car l'étiquetage des exemples est souvent effectué à la main. Ceci entraîne un coût élevé et ne permet donc pas l'obtention d'un gros corpus d'apprentissage. Cependant, les données présentes sur les sites Web 2.0, qui peuvent souvent être étiquetées par les auteurs eux-mêmes ou par d'autres internautes, permettent aujourd'hui de traiter des corpus beaucoup plus conséquents.

- Classes de prédiction
  Concernant le choix des classes, il est généralement imposé par le corpus d'apprentissage utilisé et la tâche visée. On parle de classification binaire lorsque le nombre de classes |Y| est égal à 2. Il peut naturellement être supérieur, mais il ne faut pas oublier qu'un nombre de classe élevé augmente le taux d'erreurs. En classification d'opinions, le nombre de classes de prédiction choisi est généralement de deux (aime, n'aime pas) ou trois (aime, avis neutre, n'aime pas) [40].
- Représentation
  Les documents textuels sont des suites de caractères. Afin d'en permettre l'exploitation, il est nécessaire de représenter les textes numériquement. En fouille de textes, les documents sont généralement représentés sous leur forme vectorielle dite en "*sac de mots*". Les mots sont alors considérés comme des objets distincts non ordonnés. Cette représentation consiste tout d'abord à sélectionner des variables représentant les dimensions d'un vecteur puis à représenter chaque document sur ce vecteur.

  En classification de texte, la représentation implique donc une étape préliminaire appelée segmentation, qui consiste à découper le texte afin de sélectionner les variables. Une fois les variables acquises, plusieurs documents peuvent être représentés sur le même espace vectoriel. On obtient alors une matrice dite de Salton [60]. Nous présentons tout d'abord en quoi consiste exactement la segmentation, puis nous énumérons les types de représentations les plus utilisées en recherche d'information qui sont également celles utilisées en classification d'opinions.
  - Segmentation
    Un document est un ensemble de caractères. La segmentation consiste à découper cet espace de caractères afin d'obtenir un espace de variables. Cet espace de variables est appelé vocabulaire et différents choix concernant sa construction sont possibles suivant le délimiteur choisi. On peut considérer les marques de ponctuation, afin de découper le document en phrases, ou encore les espaces associés à la ponctuation afin d'obtenir des mots. Ce sont d'ailleurs les mots qui sont les variables les plus utilisées en classification d'opinions, mais d'autres choix existent. On peut par exemple limiter la taille des variables à un nombre n de caractères, on parle alors de n-grammes de lettres. Les n-grammes peuvent également être formés de mots. Ces n-grammes de mots peuvent être construits selon leurs ordre dans

la phrase, afin de conserver un sens sémantique. Par exemple, dans la phrase «*je n'aime pas ce film* », le bigrammes (ou 2-grammes) *« aime pas »* sera considéré. On peut également construire les n-grammes de mots selon qu'ils apparaissent dans la même phrase par exemple. Une fois le vocabulaire sélectionné, différents choix sont possibles concernant la représentation du document sur le vecteur.

— Représentation binaire
Cette représentation est la moins coûteuse en temps de calcul. Elle consiste à indiquer, pour un document, quels mots du vocabulaire sont présents (valeur égale à 1) et quels mots sont absents (valeur égale à 0) [61, 62].

— Représentation fréquentielle
Cette représentation est une extension de la représentation binaire qui prend en compte le nombre d'occurrences des variables dans chaque document. Un texte est donc représenté par un vecteur dont chaque composante correspond à la fréquence des variables dans le texte [63].

— Représentation fréquentielle normalisée
Cette représentation consiste à normaliser les vecteurs de représentation des textes par la longueur des textes. C'est-à-dire que les fréquences des variables obtenues avec la représentation fréquentielle sont remplacées par la proportion des variables dans chaque document. La proportion s'obtient en divisant la fréquence de la variable par la taille du document.

— La représentation TF-IDF
TF-IDF est l'un des systèmes de pondération des traits les plus populaires. La méthode est composée d'utilisation de la fréquence de terme *(ang.Term Frequency)* avec la fréquence inverse de document*(ang.Inverse Document Frequency)* afin de mesurer l'importance d'un terme en tant que fonction de classification. Le TF est souvent calculé comme la fréquence brute du terme divisé par le nombre total de termes dans le document :

$$tf\,(t,d) = \frac{f\,(t,d)}{|d|} \tag{2.3}$$

où $f\,(t,d)$ est la fréquence brute de terme t dans le document d, et $|d|$ est la longueur du document.
Le IDF est alors calculé comme :

$$idf\,(t,d,D) = \log \frac{|D|}{|\,(d \in D : t \in d)\,|} \tag{2.4}$$

où D est le corpus. En clair, l'IDF est le $\log$ du nombre total de documents dans le corpus divisé par le nombre de documents dans lesquels le terme se produit. Le total TF-IDF est finalement calculée de telle sorte :

$$tfidf\,(t,d,D) = tf\,(t,d) \times idf\,(t,D) \tag{2.5}$$

L'idée derrière TF-IDF est qu'un terme est généralement considéré comme moins

informationnel sur le document donné s'il apparaît également dans plusieurs autres documents dans le corpus. Une fréquence plus élevée dans d'autres documents se traduit par une faible IDF [64].

— Étiquetage *Part-of-Speech Tagging*
POS Tagging est une procédure d'extraction de données de texte utile essentielle à l'obtention de métadonnées grammaticale structurelle du corpus de textes. C'est l'acte d'attribution des étiquettes grammaticales descriptives aux mots dans un texte, par exemple décider si un mot est un adjectif, adverbe, nom, etc [65]. Ces marqueurs sont souvent créés en utilisant des algorithmes d'apprentissage machine statistiques telles que Entropie maximale [66].

Pour les tâches de classification d'opinion, le POS Tagging peut être un outil utile. La présence de certaines classes de mots peut par exemple aide à indiquer si un texte est subjectif ou non. Par exemple, une présence des pronoms personnels dans un texte peut être un indicateur que le texte contient un sentiment subjectif [67].

— Autres représentations
Adjectifs, adverbes : Ces deux traits comptent le nombre d'adjectifs et d'adverbes dans la phrase respectivement. Les adjectifs et les adverbes sont révélés d'être très bons indicateurs de la subjectivité [68, 12, 69].
Indices de subjectivité, indices de subjectivités forts et faibles : Ce sont trois traits connexes dépend d'une liste de 8000 mots et expressions subjectives qui ont été annotés manuellement par Rilof et Wiebe (2003) [70] à partir des articles de journaux. Ils utilisent un système de classification binaire et étiquettent un indice *strongsubj* si elle est subjective dans la plupart des contextes, et *weaksubj* autrement. La premier trait est binaire, qui est défini sur *true* si un indice de subjectivité est présent dans une phrase et *false* sinon. Les autres traits sont numériques et leur valeur est définie sur le nombre d'indices subjectifs forts ou faibles présents dans la phrase.
Nombres cardinaux : Ceci est un trait binaire, qui est défini sur *true* si un nombre cardinal est présent dans une phrase et *false* sinon. Dans des travaux antérieurs [71], les nombres cardinaux ont été trouvés pour être utile à la tâche de classification de la subjectivité que leur présence pourrait indiquer un fait ou une déclaration objective, évitant ainsi le sentiment.
Verbes auxiliaires modaux : Ceci est un trait binaire qui est défini sur *true* si un verbe auxiliaire modal est présent dans une phrase et *false* sinon. Ces verbes sont normalement utilisés pour la couvrir ou assumer une position, ce qui peut être un indicateur de la subjectivité.
Négation : La négation a pour effet d'inverser la polarité de toute expression subjective qui lui succède. Ce trait est défini sur *true* si un mot de négation se trouve dans la phrase, et *false* sinon.

- Classificateurs
  Un classificateur est une procédure qui, à l'aide d'un ensemble d'exemples, produit une prédiction de la classe de toute donnée. Beaucoup de méthodes de classification supervisée existent et beaucoup d'entre elles ont été testées pour la classification d'opinions. On peut citer les arbres de décision, les réseaux de neurones, la régression logistique, les règles de décision ainsi que des méthodes combinant différents classificateurs comme les systèmes de votes ou les algorithmes de Boosting. Toutefois, les méthodes les plus présentes dans la littérature, et qui semblent également être les plus performantes sur les textes, sont les Machine à Vecteurs de Support (SVM) [61, 62, 63, 72, 32] et les classificateurs Naïfs Bayésiens (NB) [50, 62, 63].

  — Plus proche voisin *(ang.Nearest Neighbor)*
  En commençant par le plus simple des algorithmes, l'algorithme du plus proche voisin*(ang. Nearest Neighbor)*. Cet algorithme n'a pas été appliqué dans l'analyse des sentiments. Ceci pourrait être expliqué par la malédiction de la dimensionnalité en combinaison avec des variables binaires, comme la présence du terme.Les plus proches voisins algorithmes ont été appliqués dans des domaines où le nombre de dimensions est faible.

  Il est connu de la littérature [73, 74] que le plus proche voisin a plus de difficultés dans l'espace de dimension supérieure. La variance des distances diminue et donc la classification devient moins significative. Avec les variables binaires, soit la distance entre les points pour une variable spécifique est minimale ou maximale, parce qu'il n'y a pas d'échelle continue.

  Contrairement à l'analyse des sentiments, dans la classification de texte le plus proche voisin obtient des résultats raisonnables en combinaison avec la mesure de TF-IDF. Les inconvénients de l'application de la mesure TF-IDF pour la sélection des traits ont été déjà discutées. Par conséquent, l'algorithme du plus proche voisin n'est pas appliquée dans cette recherche.

  — Naïf bayes
  Naïf Bayes (NB) est un algorithme de classification rapide et polyvalent, et il est probablement l'algorithme de classification de texte supervisée qui est plus facile à mettre en œuvre. La performance du classificateur NB reflète souvent sa nature simple, en ce qu'il est souvent dépassé par d'autres classificateurs plus sophistiqués tels que Support Vector Machines (SVM)[75].

  Le classificateur NB est basée sur le théorème de Bayes, qui nous donne la relation entre $P(A)$ et $P(B)$, à savoir la probabilité de l'événement A et la probabilité de l'événement B. Cette relation prend la forme du théorème ci-dessous :

$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)} \tag{2.6}$$

En bref, ce théorème permet un classificateur de calculer la probabilité a posteriori de B sachant A , en utilisant les probabilités a priori. Depuis ce théorème doit être utilisé dans un classificateur NB pour les documents, nous formulons comme ceci :

$$P\left(c_p|\vec{dj}\right) = \frac{P(c_p) \times P\left(\vec{dj}|c_p\right)}{P\left(\vec{dj}\right)} \tag{2.7}$$

où $P\left(\vec{dj}\right)$ est la probabilité qu'un texte choisi au hasard sera représenté par $\vec{dj}$ et $P(c_p)$ est la probabilité qu'un texte sélectionné au hasard appartient à la classe c_p [64].

La fonctionnalité de classification devient alors de trouver la classe avec la plus grande fonction de probabilité donnée par le produit de toutes les probabilités des traits, compte tenu de leurs étiquettes de classe. Cette fonctionnalité est décrite par l'équation ci-dessous :

$$classify(f_1, ..., f_n) = argmaxp(C = c)\prod_{i=1}^{n} p(F_i = f_i|C = c) \tag{2.8}$$

Cette équation montre la simplicité du classificateur naïf bayésien. En substance, tout ce que nous devons faire pour former notre classificateur est de compter toutes les traits et dans quelles classes ils apparaissent , et d'utiliser ces fréquences pour calculer leurs probabilités. Lors du classement, nous sélectionnons la classe qui obtient le meilleur produit des traits données par le vecteur de trait cible.

— Machines à vecteurs de support *(ang. Support Vector Machines)*
Support Vector Machines (SVM) est une technique relativement nouvelle pour la classification de texte, la méthode a été d'abord utilisé à cet effet par Joachim de 1999 [76]. Par rapport au classificateur NB, la méthode SVM est beaucoup plus complexe, et par conséquent plus difficile à mettre en œuvre.

Comme toutes les méthodes d'apprentissage machine supervisé, SVMs produisent un modèle basé sur les données d'apprentissage, qui peuvent ensuite être utilisés pour prédire les étiquettes de classe des données de test. Chaque instance des données d'apprentissage et de test est d'abord converti en un ensemble de points dans un espace de traits. Par exemple, pour la classification

de texte basé sur uni-gramme, chaque phrase peut être considérée comme un point dans un espace de traits de grande dimension avec chaque dimension représentant un uni-gramme particulier dans cette phrase.

Considérons un système de classification binaire avec *l* instances d'apprentissage dans un espace de traits *n*-dimensionnel de telle sorte que chaque instance $x_i \in R^n, i = 1, 2, ..., l$ a une étiquette de classe $y_i \in \{1, -1\}$.

La figure 2.2 montre un tel problème dans lequel les points bleus $(y_i = -1)$ doivent être séparés de ceux qui sont rouges $(y_i = 1)$.

Un tel ensemble d'instances est dit linéairement séparables si nous pouvons tracer une ligne telle qu'elle toutes les instances des deux classes sont sur les côtés opposés de cette ligne. Plus formellement, cela signifie qu'il existe un vecteur w et un scalaire b, pour laquelle l'inégalité suivante est valable pour toutes les paires d'étiquettes des instances (x_i, y_i) :

$$y_i \left(w.x_i + b\right) \geq 1 \tag{2.9}$$

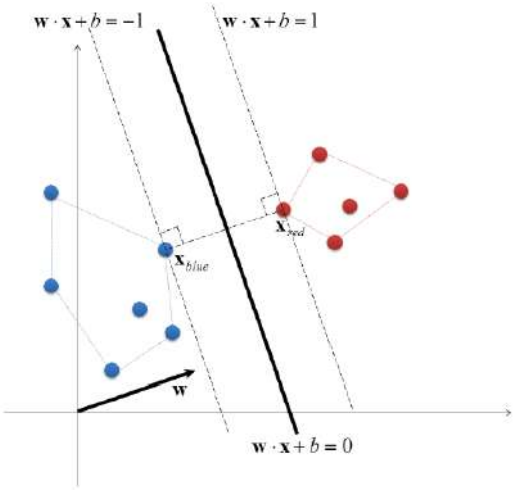


FIGURE 2.2 – Exemple d'un clasificateur SVM

La fonction de décision qui attribue une étiquette à un vecteur invisible x est donné par :

$$f\left(x\right) = sign\left(w.x + b\right) \tag{2.10}$$

Alors que beaucoup de ces vecteurs existent, on peut montrer que le vecteur optimal entraînerait une maximisation de la marge de séparation entre les deux classes. Ce vecteur serait perpendiculaire à la ligne la plus courte reliant les enveloppes convexes des deux classes, et il serait également bissectrice. Dans

la figure ci-dessus, la ligne la plus courte est la ligne entre les points x_{red} et x_{blue}. Pour le vecteur de poids w, ces points sont situés sur les lignes parallèles suivantes :

$$w.x_{red} + b = 1 \tag{2.11}$$

$$w.x_{blue} + b = -1 \tag{2.12}$$

Ces lignes sont appelés vecteurs de support et sont représentés en traits pointillés sur la figure. Soustraire ces équations, nous obtenons :

$$w.\left(x_{red} - x_{blue}\right) = 2 \tag{2.13}$$

$$\left(\frac{w}{||w||}.\left(x_{red} - x_{blue}\right)\right) = \frac{2}{||w||} \tag{2.14}$$

Par conséquent, afin de maximiser la marge, il faut réduire au minimum ||w|| soumis à la contrainte donnée dans l'équation 2.9. Cortes et Vapnik (1995) montrent que la solution à ce problème d'optimisation sous contrainte a une expansion $w = \sum_i v_i x_i$où v_i sont les instances d'apprentissage qui se trouvent sur les lignes données par l'équation 2.11 et l'équation 2.12. La fonction de décision peut donc être révisée pour :

$$f(x) = sign(w.x + b) = sign(\sum_i .v_i(x.x_i) + b) \tag{2.15}$$

Nous pouvons voir que la solution au problème d'optimisation, ainsi que la limite de la décision finale ne dépendent que des produits scalaires entre ces instances. Ce fait est important car il permet de résoudre des problèmes de classification non linéaires en utilisant SVMs en les tapotant à des problèmes linéaires dans un espace de dimension supérieure. À titre d'exemple, consi-

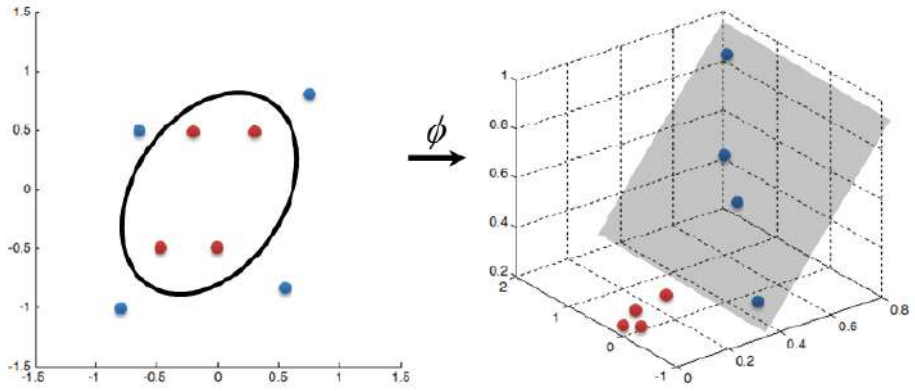


FIGURE 2.3 – Utilisation d'une fonction noyau

dérons le problème à la figure 2.3. Les données sont linéairement inséparable dans un espace vectoriel à deux dimensions. Cependant, en utilisant une fonc-

tion $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ convertit le problème à des données linéairement séparables dans un espace de dimension supérieure.Ici, chaque point de données est mappé à un autre point dans un espace vectoriel en trois dimensions. La fonction de décision pour un vecteur invisible x dans l'équation 2.15 peut donc être mis à jour :

$$f(x) = sign\left(\sum_i v_i \left(\phi(x).\phi(x_i)\right) + b\right) \tag{2.16}$$

Comme indiqué précédemment, la fonction de décision nécessite le calcul des produits scalaires uniquement. [77] ont suggéré d'utiliser la forme générale des produits de points dans un espace de Hilbert [78] à cet effet. Cette forme générale est donnée par :

$$k(u, v) =< \phi(u).\phi(v) > \tag{2.17}$$

À l'aide d'un noyau simple k, le produit scalaire peut être calculé efficacement dans l'espace de dimension supérieure aussi bien. Par exemple, pour le problème à la figure précédente avec deux points u = (u1,u2) et v = (v1,v2), considérons le noyau polynôme suivant :

$$\begin{aligned} k(u, v) &=< u, v >^2 \\ &= (u_1 v_1 + u_2 v_2) \\ &= u_1^2 v_1^2 + 2u_1 v_1 u_2 v_2 + u_2^2 v_2^2 \\ &=< \left(u_1^2, \sqrt{2}u_1 u_2, u_2^2\right) . \left(v_1^2, \sqrt{2}v_1 v_2, v_2^2\right) > \\ &=< \phi(u).\phi(v) > \end{aligned}$$

Ceci implique que la fonction de mappage est donnée par :

$$\phi(x) \rightarrow \left(x_1^2, \sqrt{2}x_1 x_2, x_2^2\right)$$

Par conséquent, la fonction de décision finale pour un point de données invisible (x) en utilisant un noyau k est donné par :

$$f(x) = sign\left(\sum_i v_i.k(x, x_i) + b\right) \tag{2.18}$$

Les valeurs v_i sont calculées en résolvant la fonction comme un problème de programmation contraint. De nombreuses implémentations pour résoudre ce problème sont accessibles au public. *Exemple :* l'mplémentation LIBSVM Library disponible via la boîte à outils WEKA.

— Entropie maximale *(ang. Maximum Entropy)*
Un classificateur Maximum Entropy (MaxEnt) est un classificateur probabiliste conditionnelle. Les implémentations de celui-ci utilisent la régression logistique afin de trouver la distribution de probabilité avec la plus grande entropie, qui est donnée par la théorie de l'entropie maximale [79] devrait être la mieux pour représenter l'état actuel des connaissances, compte tenu des données antérieures précisément indiquées [80] .

Contrairement au classificateur NB, MaxEnt décline toute indépendance conditionnelle pour les fonctionnalités. Cela signifie que MaxEnt gère le chevauchement des traits mieux que le classificateur NB [81]. Cela signifie également que pour les fonctions de texte seulement, le classificateur MaxEnt fonctionnera mieux étant donné que la plupart du temps, nous travaillons avec des mots qui sont conditionnellement dépendants les uns des autres.

Cette formule présente le modèle de la façon suivante[81] :

$$P\left(c|d,\lambda\right)=\frac{\exp\left[\sum_i \lambda_i f_i(c,d)\right]}{\sum_{c'}\exp\left[\sum_i \lambda_i f_i(c,d)\right]} \tag{2.19}$$

où, dans ce cas, c est la classe et d est le document.Ici λ représente un facteur de poids, ce qui correspond à la pertinence d'une caractéristique(feature) pour une classe donnée.Les opérations numériques de la tâche de l'optimisation de ces lambdas est ce qui rend la mise en œuvre MaxEnt non-trivial et long.

Pour les tâches de classification de texte, les classificateurs MaxEnt ont été montré pour avoir une performance de précision qui est à égalité avec SVM [80] .

— Arbre de décision
Un classificateur de texte basé sur la méthode d'arbre de décision est un arbre de noeuds internes qui sont marqués par des termes, les branches qui sortent des noeuds sont des tests sur les termes, et les feuilles sont marquées par catégories [82]. Ce classificateur classe un document du test d_j en testant récursivement les poids des noeuds internes de vecteur $\overrightarrow{d_j}$, jusqu'à ce qu'une feuille soit atteinte. L'étiquette de ce noeud est alors attribuée à d_j. La plupart de ces classificateurs utilise une représentation du document binaire, et sont donc crée par des arbres binaires.

Il existe un certain nombre d'approches pour l'apprentissage de l'arbre de décision. Les plus populaires sont ID3 [83], C4.5 [84] et C5 [85].

Une méthode pour effectuer l'apprentissage d'un arbre de décision pour la catégorie c_i consiste à vérifier si tous les exemples d'apprentissage ont la même étiquette $(c_i \ ou \ \overline{c_i})$, dans le cas contraire nous sélectionnons un terme t_k, et nous partitionnons l'ensemble d'apprentissage en classes de documents qui ont la même valeur pour t_k, et à la fin l'on crée les sous-arbres pour chacune de ces classes. Ce processus est répété récursivement sur les sous-arbres jusqu'à ce que chaque feuille de l'arbre généré de cette façon contienne des exemples d'apprentissage attribués à la même catégorie c_i, qui est alors choisie comme l'étiquette de la feuille. L'étape la plus importante est le choix du terme de t_k pour effectuer la partition.

Toutefois, une telle méthode de construction d'arbre peut faire l'objet de surapprentissage, comme certaines branches peuvent être trop spécifiques pour les données d'apprentissage. La plupart des méthodes d'apprentissage des arbres incluent une méthode pour la construction d'arbre et pour élaguer les branches trop spécifiques [82].

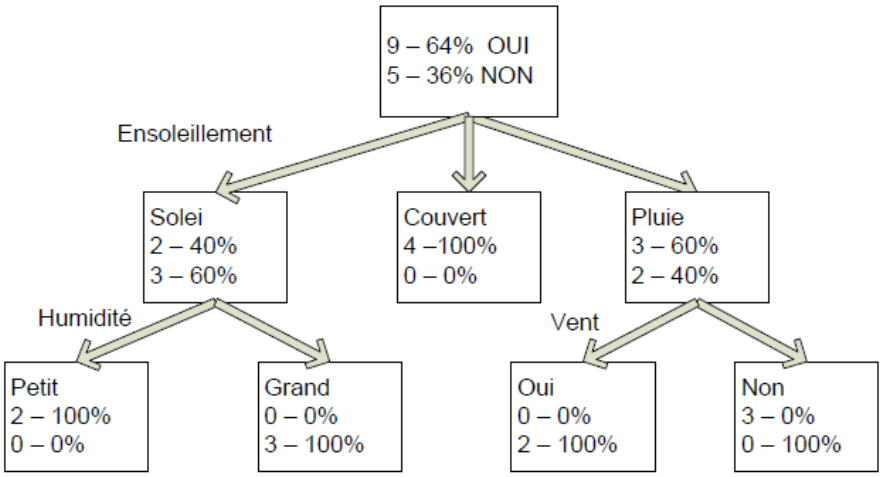


FIGURE 2.4 – Exemple d'un arbre de décision

L'algorithme d'apprentissage cherche à produire des groupes d'individus les plus homogènes possible du point de vue de la variable à prédire à partir des variables de météo.

L'exemple d'un arbre de décision décrit dans l'ouvrage [86] est montrée sur la figure 2.4. Il s'agit de prédire le comportement de sportifs (Jouer ; variable à prédire) en fonction de données météo (Ensoleillement, Température, Humidité, Vent ; variables prédictives). Sur chaque sommet de l'arbre est décrite la distribution de la variable à prédire. Dans le cas du premier sommet, la ra-

cine de l'arbre, nous constatons qu'il y a 14 observations dans notre fichier, 9 d'entre eux ont décidé de jouer (Jouer = oui), 5 ont décidé le contraire (Jouer = non). Ce premier sommet est segmenté à l'aide de la variable Ensoleillement, 3 sous-groupes ont été produits. Le premier groupe à gauche (Ensoleillement = Soleil) comporte 5 observations, 2 d'entre eux correspondent à Jouer = oui, 3 à Jouer = non. Chaque sommet est ainsi itérativement traité jusqu'à ce que l'on obtienne des groupes suffisamment homogènes. Elles correspondent aux feuilles de l'arbre, des sommets qui ne sont plus segmentés.

— Réseaux de neurones
Un classificateur de texte basé sur les réseaux de neurones *(ang. Neural Networks NN)* est un réseau d'unités, où les unités d'entrée représentent les termes, l'unité(s) de sortie représentent la catégorie ou les catégories d'intérêts, et le poids sur les bords reliant les unités représentent les relations de dépendance. Pour classer un document de test d_j , ses poids w_{kj} sont chargés dans les unités d'entrée ; l'activation de ces unités se propage à travers le réseau, et la valeur de l'unité de sortie (s) détermine la décision du classement.

Une manière typique d'apprentissage de réseau de neurones est la rétro propagation, qui consiste à rétro propager l'erreur commise par un neurone à ses synapses et aux neurones qui y sont reliés. Pour les réseaux de neurones, on utilise habituellement la rétro propagation du gradient de l'erreur, qui consiste à corriger les erreurs selon l'importance des éléments qui ont justement participé à la réalisation de ces erreurs.

Le type le plus simple de classificateur de NN est le perceptron [87], qui est un classificateur linéaire. Dans cet algorithme, le classificateur de c_i est d'abord initialisé par la mise de tous les poids w_{kj} à la même valeur positive. Quand un exemple d'apprentissage d_j (représenté par un vecteur $\overline{d_j}$ de poids binaire) est examiné, et si le résultat de la classification est correct, rien n'est fait, alors que si le résultat est faux, les poids du classificateur sont modifiés :

- si d_j est un exemple positif de c_i, le poids de w_{ki} des termes actifs (c'est-à-dire, les termes t_k tels que $w_{kj} = 1$) sont "promus" par augmentation d'une valeur $\alpha > 0$ (appelé taux d'apprentissage),
- si d_j est un exemple négatif de ci les mêmes poids sont "rétrogradés" en diminuant leur valeur parα.

Lorsque le classificateur a atteint un niveau raisonnable d'efficacité, le fait qu'un poids w_{ki} est très faible signifie que t_k a contribué négativement à la procédure de classement, il peut donc être éliminé de la représentation. Le classificateur perceptron a montré une bonne efficacité [87].

Un réseau de neurones est en général composé d'une succession de couches dont chacune prend ses entrées sur les sorties de la précédente. Chaque couche est composée de n_i neurones, prenant leurs entrées sur les n_{i-1} neurones de la couche précédente. A chaque synapse est associé un poids synaptique, de sorte que les n_{i-1} sont multipliés par ce poids, puis additionnés par les neurones de niveau i, ce qui est équivalent à multiplier le vecteur d'entrée par une matrice de transformation.

Le réseau de neurones peut également contenir des boucles qui en changent radicalement les possibilités mais aussi la complexité. De la même façon que des boucles peuvent transformer une logique combinatoire en logique séquentielle, les boucles dans un réseau de neurones transforment un simple dispositif de reconnaissance d'entrées en une machine complexe capable de toutes sortes de comportements. La représentation schématique d'un neurone artificiel avec un index j est montré sur la fiqure 2.5.

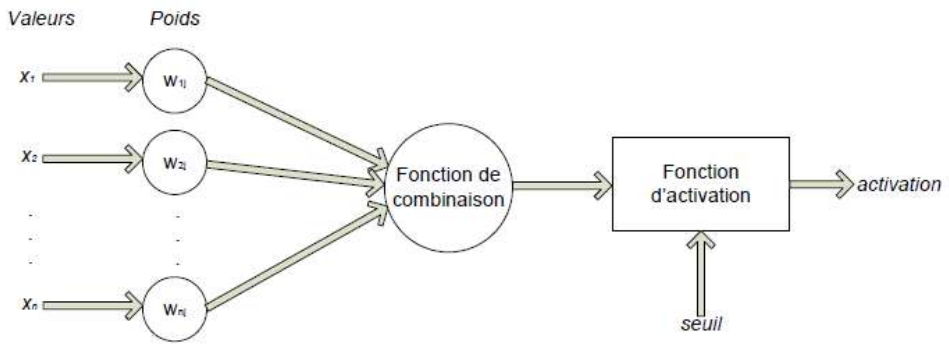


FIGURE 2.5 – Structure d'un neurone artificiel - Le neurone calcule la somme de ses entrées puis cette valeur passe à travers la fonction d'activation pour produire sa sortie.

3.2.2 Classification d'opinion utilisant la fonction de score personnalisé

Au lieu d'utiliser la méthode d'apprentissage machine standard, les chercheurs ont proposé des techniques personnalisées spécifiquement pour la classification de sentiment des critiques. La fonction de score dans Dave et al. (2003), est une telle technique. Elle est basée sur les mots des critiques positives et négatives. L'algorithme est constitué de deux étapes :

Etape 1 : marquer chaque terme (uni-gramme ou n-gramme) dans l'ensemble d'apprentissage en utilisant l'équation suivante :

$$score(t_i) = \frac{P(t_i|C) - P(t_i|C\prime)}{P(t_i|C) + P(t_i|C\prime)} \tag{2.20}$$

où t_i est un terme, C est une classe,C' est son complément, qui n'est pas C et $P(t_i|C)$ est la probabilité conditionnelle du terme t_i en classe C, qui est calculé en prenant le nombre de fois qu'un terme t_i se produit dans la classe de critiques C et en le divisant par le nombre total de termes dans les critiques de la classe C. un score à long terme est donc une mesure de termes biais vers soit la classe allant de -1 à 1.

Etape 2 :classer un nouveau document $d_i = t_1...t_n$ en additionnant les scores de tous les termes et en utilisant le signe du total pour déterminer la classe :

$$class(d_i) = \begin{cases} C & eval(d_i) > 0 \\ C' & sinon. \end{cases} \tag{2.21}$$

Où,

$$eval(d_i) = \sum_j score(t_j). \tag{2.22}$$

Des expériences ont été menées sur la base de plus de treize mille commentaires de sept types de produits. Les résultats ont montré que les bi-grammes (deux mots consécutifs) et tri-grammes (trois mots consécutifs) comme des termes ont donné (similaire) des meilleures pertinences (84,6 à 88,3%). Ni normalisation ni suppression de mots vides a été appliquée.

Les auteurs ont également expérimenté avec beaucoup de nombreuses techniques alternatives de classification, par exemple, naïf Bayes, SVM, et plusieurs fonctions de score variante, et des stratégies de substitution de mots pour améliorer la généralisation, par exemple,

- remplacer les noms de produits avec un token("$_productname$")
- remplacer les mots avec un token ("$_unique$")
- remplacer les mots des catégorie spécifique avec un token ("$_producttypeword$")
- remplacer les tokens numériques avec NUMBER

Quelques modifications linguistiques en utilisant WordNet, la normalisation, la négation, et de la colocalisation ont été testés aussi.Cependant, ne sont pas utiles effectivement elle dégradent la précision de la classification.

3.3 Classification d'opinion semi supervisée (approches hybrides)

Plusieurs types d'hybridation sont présents dans la littérature. Dans tous les cas, elles utilisent des éléments décrits dans les deux approches précédentes. On peut distinguer parmi ces approches trois méthodes distinctes :

3.3.1 Linguistique au service de l'apprentissage automatique

Une approche hybride consiste à utiliser les outils linguistiques afin de préparer le corpus avant de classer les textes à l'aide de l'apprentissage supervisé. Wilson et al. [32] préparent les données à l'aide d'outils de TALN afin de sélectionner un vocabulaire d'opinions. Ces mots pré-sélectionnés sont ensuite utilisés comme vecteurs de représentation des textes pour les outils d'apprentissage supervisé.

Trois algorithmes d'apprentissage sont comparés : BoostTexter [88], Ripper [89] et SVM-light [76]. Nigam et Hurst [61] utilisent des techniques provenant du Traitement Automatique des Langues afin de détecter dans les textes les mots et expressions porteurs d'opinions et ajoutent des marques dans le texte (traits grammaticaux et + ou - pour opinion positive et opinion négative). Ils utilisent ensuite l'apprentissage automatique pour classer les textes selon leur opinion générale.

3.3.2 Apprentissage automatique au service de la linguistique

Une autre façon de combiner les méthodes est d'utiliser les techniques d'apprentissage automatique dans le but de construire les dictionnaires d'opinions nécessaires à l'approche linguistique.

Hatzivassiloglou et McKeown [68] présentent une méthode ayant pour objectif de définir l'orientation sémantique des adjectifs pour la construction du dictionnaire d'opinions. Ils extraient tout d'abord tous les adjectifs du corpus à l'aide d'un analyseur syntaxique, puis utilisent un algorithme de clustering afin de classer les adjectifs selon leur polarité.

Riloff et Wiebe [70] combinent les deux approches afin de répertorier les expressions porteuses d'opinion qui, selon eux, sont plus riches que des mots pris individuellement.

Turney et Littman [43] utilisent une approche statistique pour classer un plus grand nombre de types de mots selon leur polarité : adjectifs, verbes, noms...

3.3.3 Fusion a posteriori des résultats des deux approches

Une dernière façon d'utiliser conjointement les approches basées sur la linguistique et celles basées sur l'apprentissage automatique est de construire plusieurs types de classificateurs afin de combiner leurs résultats, soit par des systèmes de vote, soit par un algorithme d'apprentissage [90].

Dans le cas d'un système de vote, on attribue généralement des poids suivant les performances des classificateurs utilisés. Si les résultats diffèrent, on conserve alors le résultat du classificateur ayant le poids le plus fort. Dans le cas de l'utilisation d'un algorithme

d'apprentissage, les poids sont calculés automatiquement selon l'indice de confiance attribué par chaque classificateur à ses propres résultats.

La classification d'opinion au niveau du document donne un avis global sur une entité, un sujet ou événement. Elle a été étudiée par un grand nombre de chercheurs. Toutefois, ce niveau de classification comporte des lacunes pour les applications :

Dans de nombreuses applications, l'utilisateur a besoin de connaître les détails supplémentaires, par exemple, quels sont les traits des entités qui sont aimé et détesté par les consommateurs. Dans les documents d'opinion typiques, ces détails sont fournis, mais la classification de documents d'opinions ne les extrait pas pour l'utilisateur.

La classification d'opinions au niveau du document est difficilement applicable aux non-critiques tels que des forums de discussion, des blogs et des articles de presse, parce que beaucoup de ces affichages peuvent évaluer plusieurs entités et de les comparer. Dans de nombreux cas, il est difficile de déterminer si une affectation évalue effectivement les entités que l'utilisateur est intéressé, et si l'affichage exprime une opinion à tous, et encore moins de déterminer le sentiment à leur sujet.

La classification d'opinions au niveau du document ne fonctionne pas de telles tâches à grains fins, qui nécessitent un traitement en profondeur le langage naturel. En fait, les commentaires en ligne n'ont pas besoin de classification d'opinion parce que presque tous les avis ont déjà le nombre de classes attribuées par l'utilisateur. Dans la pratique, ce sont les discussions et les blogs qui ont besoin de classification d'opinions pour déterminer l'opinion des gens au sujet de différentes entités (par exemple, les produits et services) et les sujets du forum [42].

4 Classification d'opinions au niveau des phrases

Si une phrase est classée comme étant subjective, nous déterminons si elle exprime une opinion positive ou négative. L'apprentissage supervisé et les méthodes basées sur le lexique peuvent être appliqué aussi juste comme pour la classification d'opinion au niveau du document (certains algorithmes n'utilisent pas l'étape de la classification de subjectivité).

Dans [91], une classification d'opinions de messages Twitter (ou tweets) a été étudiée. Chaque tweet est essentiellement une seule phrase. Les auteurs ont adopté une approche d'apprentissage supervisé. Outre les caractéristiques traditionnelles, la méthode a également utilisée hashtags, smileys, ponctuations, et leurs motifs fréquents. Ces caractéristiques se sont avérées très efficaces.

Dans [92], un algorithme d'apprentissage semi-supervisé a été utilisé pour apprendre d'un petit ensemble de phrases marquées et un grand ensemble de phrases non marquées. L'algorithme d'apprentissage a été basé sur Expectation Maximization (EM) en utilisant le Bayes naïf comme un classificateur de base. Ce travail a effectué une classification à trois classes, positif, négatif, et «autres» (pas d'opinion ou opinion mixte).

5 Classification d'opinions au niveau des traits linguistiques

Les gens n'expriment pas leurs opinions sur des documents et des phrases seulement, ils les expriment aussi sur des traits et des entités. Le niveau des informations fournies dans le niveau du document ou de phrase ne suffit pas pour prendre une bonne décision et donc la recherche approfondie sur les traits et les entités a donné une nouvelle direction pour la recherche appelé fouille d'opinion basée sur les entités ou les traits *(ang. features ou aspects)*.

Un inconvénient au niveau document ou phrase est qu'ils ne peuvent pas fournir des informations complètes d'un produit, par exemple, une critique positive ou négative d'un produit particulier ne signifie pas que l'examinateur aime ou déteste tous les caractéristiques de ce produit. Une personne qui a besoin d'acheter un mobile avec une excellente qualité de la caméra va chercher seulement pour les commentaires sur cette caractéristique particulière à savoir "la qualité de l'image" plutôt que la critique globale de ce mobile.

Les traits d'opinions uniques sont des commentaires où les gens se concentrent sur un seul trait du produit, alors que dans plusieurs traits, les gens expriment des opinions divergentes sur plus d'un trait à la fois dans la même critique, et même dans la même phrase de ce produit. Par exemple, *«This book had a good storyline, but the paper quality is bad»*. Ici, l'examinateur a donné une opinion positive sur le trait *«story»* et une opinion négative sur le trait du livre*«paper quality»*.

Deux tâches principales dans la fouille d'opinions au niveau du caractéristique sont l'extraction des traits et la classification d'opinions au niveau du trait. Le processus d'identification des mots d'opinion de la phrase donnée est appelée extraction de traits et de classer les mots d'opinion extraits dans l'une des échelles de polarité est appelée classification d'opinion au niveau du trait. Les gens expriment les opinions soit implicitement (indirects) ou explicitement (directs). L'extraction des expressions des traits implicites est une tâche difficile, car les mots d'opinions diffère entre les gens.

Pour identifier les traits implicites, certaines études antérieures [93, 94, 95, 96] ont tenté d'utiliser différentes approches. Une étude montre que les adjectifs, les adverbes et les noms subjectifs sont considérés comme traits des mots liés dans la plupart des cas et

donne une haute performance [12, 70, 53].

Considérons cette critique *«This bike is very smooth to ride, breaks are good, engine is not noisy, suspension of wheels are ok, ignition is digital and also has self-start option, mileage is not up to the mark.»*. Il est clairement connu que l'auteur a donné son point de vue sur plusieurs traits de la moto. La polarité de chaque trait est évalué et donné dans le tableau ci-dessous 5.3.

Caractéristique	**Positive(+)**	**Neutre(=)**	**Négative(-)**
Engine	*		
Ignition	*		
Mileage			*
Breaks	*		
Suspension		*	

TABLE 5.3 – Polarité des traits

En général, pas tous les gens aiment ou détestent tous les traits d'un produit, il diffère de chaque individu, d'une critique de 1000 phrases, il est difficile pour les clients potentiels à lire les opinions sur toutes les traits à la fois et faire une décision d'acheter le produit ou non. De nombreux travaux de recherche ont été menées afin de surmonter cette difficulté. La plupart des premiers travaux de fouille d'opinions au niveau des traits sont classés comme une de ces approches : basée sur la fréquence, sur les relations et sur les modèles [97].

6 Classification d'opinions avec croisement de langues

La classification d'opinions en langues Croisées signifie d'effectuer la classification d'opinions de documents d'opinions dans plusieurs langues. Il existe deux principales motivations pour la classification inter-langues. Les chercheurs de différents pays veulent construire des systèmes d'analyse des sentiments dans leurs propres langues. Cependant, une grande partie de la recherche a été fait en anglais. Il n'y a pas beaucoup de ressources ou d'outils dans d'autres langues qui peuvent être utilisées pour construire de bonnes classificateurs de sentiment rapidement dans ces langues.

La question naturelle est de savoir s'il est possible de tirer parti de la capacité automatisée de la machine de traduction et les ressources existantes d'analyse des sentiments et des outils disponibles en anglais pour aider à construire des systèmes d'analyse des sentiments dans d'autres langues. La deuxième motivation est que dans de nombreuses applications, les entreprises veulent connaître et comparer les opinions des consommateurs sur leurs produits et services dans les différents pays. Si elles ont un système d'analyse des sentiments en anglais, ils veulent construire rapidement des systèmes d'analyse des sentiments

dans d'autres langues grâce à la traduction.[42]

Dans [98], les auteurs ont présenté leurs opinions sur la recherche de la classification d'opinions inter-langues. Sur la base de leur analyse, ils ont affirmé que le domaine non con-cordonné n'a pas été causé par la traduction automatique *(ang. Machine Translation)* des erreurs, et la dégradation de la précision se produirait même avec MT parfaite. Il a également fait valoir que le problème d'adaptation inter-langues était qualitativement différent des autres (monolingues) problèmes d'adaptation dans le traitement automatique des langues, ainsi de nouveaux algorithmes d'adaptation doivent être considérés.

Dans [99], les auteurs ont proposé une méthode d'apprentissage de la représentation en deux étapes pour combler les espaces de traits linguistiques de différentes langues en exploitant un ensemble de documents bilingues parallèles. Plus précisément, ils ont formulé d'abord un problème de matrice d'achèvement pour produire une matrice parallèle complète (document-terme) pour tous les documents dans les deux langues. Puis induire une faible dimension à la représentation du document en langues croisées en appliquant l'indexation sémantique latente *ang. Latent semantic Indexing* sur la matrice obtenue. Ils ont utilisé un algorithme de descente de gradient projeté pour résoudre le problème de l'achèvement de la matrice formulée avec des garanties de convergence. La méthode proposée est évaluée par la réalisation d'une série d'expériences avec les tâches de classification d'opinions en langues croisées sur les critiques de produits Amazon.

Afin de faciliter l'analyse des sentiments inter-langues,[100] ont proposé de comparer le sentiment véhiculé par le texte non structuré à travers les langues à travers le nombre de classes pour le sentiment prévu. Ils ont démontré que la façon dont le langage naturel révèle le sentiment destiné aux gens diffère selon les langues. Les résultats des expériences en ce qui concerne la modélisation de cette relation pour néerlandais et anglais au moyen d'une fonction de l'étape monotone croissante suggèrent essentiellement que les scores de sentiment spécifiques à la langue peut séparer les classes de sentiment destinés les uns des autres dans une mesure limitée.

La performance de classification d'opinions repose sur les ressources de sentiment de haute qualité. Cependant, ces ressources sont déséquilibrés dans différentes langues. La classification d'opinions en langues Croisées *(ang. Cross Language Sentiment Classification)* peut tirer parti de la richesse des ressources dans une langue (langue source) pour la classification d'opinions dans une langue (langue cible) de ressources rares. Les plongements bilingues pourraient éliminer le fossé sémantique entre deux langues pour CLSC, mais ignorer les informations de sentiment de texte.

[101] propose une approche d'apprentissage de plongement de mot de sentiment bilingue*(ang. Bilingual Sentiment Word Embeddings)* pour le CLSC anglais-chinois. La proposition BSWE incorpore des informations de sentiment de texte dans les plonge-

ments bilingues. De plus,ils peuvent apprendre BSWE de haute qualité en utilisant un simple corpus étiqueté et leurs traductions, sans compter les corpus parallèles sur grande échelle.

7 Classification d'opinions avec croisement de domaines

Il a été démontré que la classification d'opinions est très sensible au domaines à partir duquel les données d'apprentissage sont extraient. Un classificateur formé en utilisant des documents d'opinions d'un domaine effectue souvent mal sur les données de test à partir d'un autre domaine. La raison est que les mots et même les constructions de langage utilisé dans différents domaines pour exprimer des opinions peuvent être très différents. Pour aggraver les choses, le même mot dans un domaine peut signifier positif, mais dans un autre domaine peut signifier négatif. Ainsi, l'adaptation du domaine ou le transfert d'apprentissage est nécessaire.[42]

Les recherches actuelles sont basées principalement sur deux paramètres. Le premier paramètre a besoin d'une petite quantité de données d'apprentissage marqués pour le nouveau domaine [102]. La seconde nécessite aucune étiquette de données pour le nouveau domaine [103]. Le domaine d'origine avec des données d'apprentissage marqué est souvent appelé le domaine source et le nouveau domaine qui est utilisé pour le test est appelé le domaine cible.

Xia et Zong (2011)[104] ont constaté que dans les différents domaines, les traits de certains types de balises de partie du discours *(ang. Part Of Speech tags)* sont généralement dépendant du domaine, tandis que certains autres sont indépendant du domaine. Partant de ce constat, ils ont proposé un modèle d'ensemble sur la base POS à intégrer des traits avec différents types de POS tags pour améliorer les performances de classification.

He, Lin et Alina (2011)[105] ont utilisé la modélisation des sujets communs pour identifier les sujets d'opinions de deux domaines pour les combler. Les sujets résultant qui couvrent les deux domaines sont utilisés comme des traits linguistiques supplémentaires pour augmenter l'ensemble original des traits linguistiques pour la classification.

Dans [106], le sujet de modélisation a été utilisé aussi pour trouver un espace sémantique commun basé sur les correspondances des termes par rapport au domaine et les co-occurrences des termes dans les deux domaines. Cet espace sémantique commun a ensuite été utilisé pour l'apprentissage d'un classificateur qui a été appliquée au domaine cible.

8 Mesures d'évaluations

L'évaluation du classificateur des documents est effectuée de façon expérimentale, plutôt que de façon analytique. Le traitement est expérimental car pour évaluer un système de façon analytique, donc pour prouver que le système est correct et complet, nous aurions besoin d'une spécification formelle du problème que le système tente de résoudre. La notion de systèmes de catégorisation de texte est, en raison de son caractère subjectif, par nature non-formalisable comme il s'agissait de déterminer l'appartenance d'un document à une catégorie.

L'efficacité de classification se mesure généralement par les paramètres classiques du domaine de la recherche d'information :

- la précision peut être considérée comme le "degré de solidité" du classificateur, elle est définie comme la probabilité conditionnelle :

  $P\left(\phi(d_x, c_i) = T | \phi(d_x, c_i) = T\right)$, qui signifie la probabilité que, si d_x est classé dans c_i, cette décision est correcte.

  $Précision_i = \frac{Objets\ correctement\ attribués\ à\ la\ classe_i}{Nombres\ d'objets\ attribués\ à\ la\ classe_i}$

  Elle mesure le bruit, et plus elle est proche de 100%, moins il y a de bruit, et donc meilleure est la réponse.

  $Précision\ totale = \frac{\sum_{i=0}^{n} Précision_i}{n}$, qui effectue la moyenne des précisions de chaque classe.

- le rappel peut être considéré comme le "degré d'exhaustivité" du classificateur, il est défini comme :

  $P\left(\phi(d_x, c_i) = T | \phi(d_x, c_i) = T\right)$, qui signifie la probabilité que, si un document quelconque d_x devait être classé sous c_i, cette décision était prise. Ce dernier est le nombre de documents bien classés en rapport au nombre de documents pertinents contenus dans le corpus.

  $Rappel_i = \frac{Documents\ correctement\ attribués\ à\ la\ classe_i}{Nombre\ de\ documents\ attribués\ à\ la\ classe_i}$

  Plus il est proche de 100%, moins il y a de silence, et meilleure est la réponse.
  L'évaluation de la pertinence dans les recherches documentaires essaie de quantifier le rappel et la précision, avec les indices de bruit et de silence.

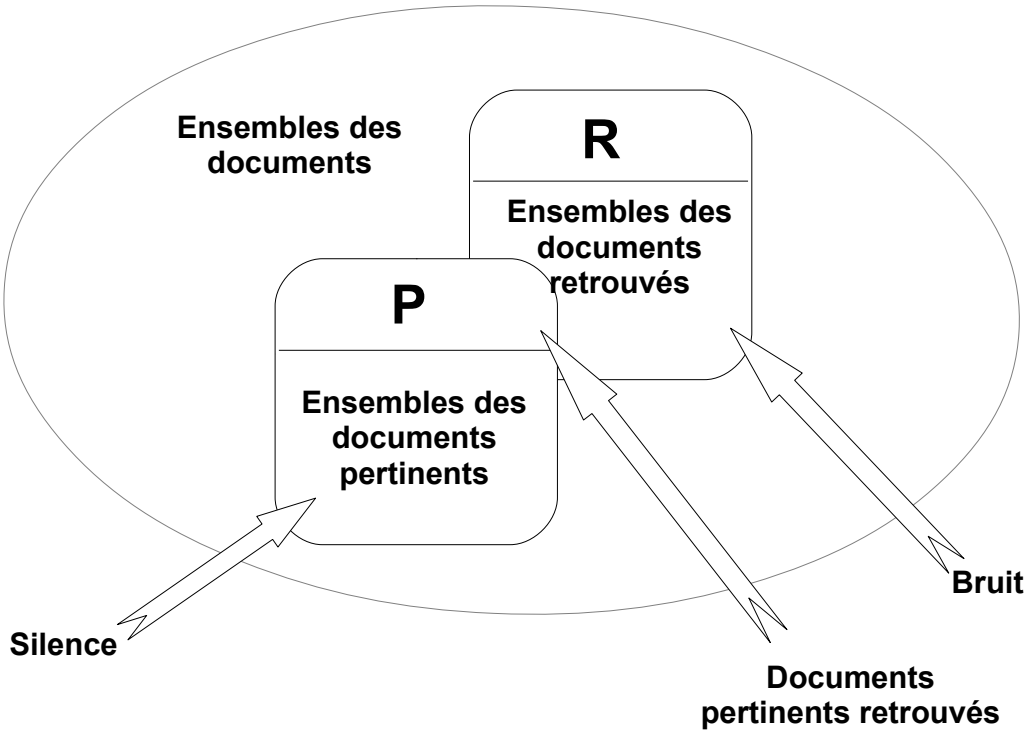


FIGURE 2.6 – Représentation du bruit et du silence en recherche d'information.

Le bruit est un ratio entre le nombre de documents ramenés en réponse, mais qui ne sont pas pertinents par rapport à la question posée, et le nombre total de documents. Donc :
$Bruit = 1 - Précision$

Le silence est un ratio entre le nombre de documents pertinents qui n'apparaissent pas dans le résultat de la recherche et le nombre total de documents. Donc :
$Silence = 1 - Rappel$

La précision et le rappel peuvent être définis comme les probabilités subjectives, qui mesurent l'attente de l'utilisateur d'un système se comportant correctement lors de la classification d'un document inconnu pour une catégorie c_i. Ces probabilités peuvent être estimées par le tableau 8.4 qui montre toutes les possibilités d'un classificateur.

Expert / Classificateur	Vrai	Faux
Positif	VP_i	FP_i
Négatif	FN_i	VN_i

TABLE 8.4 – Exemple d'une matrice de confusion pour une classification binaire

FPi (erreurs de commission) : nombre de documents de test mal classés dans la catégorie c_i ;
VNi : ensemble de documents bien classés qui n'appartiennent pas à la catégorie c_i ;
VPi : ensemble de documents bien classés qui appartiennent à la catégorie c_i ;
FNi (erreurs d'omission) : ensemble des documents de catégorie c_i non classés par le classificateur.

Donc, la précision et le rappel peuvent être exprimés de la façon suivante :

$$Précision_i = \frac{VP_i}{VP_i + FP_i} \tag{2.23}$$

$$Rappel_i = \frac{VP_i}{VP_i + FN_i} \tag{2.24}$$

Des mesures alternatives de précision et de rappel sont couramment utilisées dans la littérature, telles que :

La pertinence *(ang.accuracy)* :

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.25}$$

L'erreur :

$$E = \frac{FP + FN}{VP + VN + FP + FN} = 1 - A \tag{2.26}$$

Cependant,en général la précision et le rappel ne sont pas les seuls pris en compte à cause de la variabilité. La variabilité peut induire en erreur certaines mesures de performance. Pour cette raison, deux nouvelles alternatives sont utilisées :

- Seuil de rentabilité : c'est une valeur lorsque la précision et le rappel sont les mêmes et fournit un score unique qui relie les deux valeurs.
- F-mesure :utilisé comme mesure de l'efficacité de la classification. F-mesure est définie comme suit :

$$F - mesure = \frac{2 \times Précision \times Rappel}{Précision + Rappel} \tag{2.27}$$

  F-mesure a été conçue pour équilibrer le poids de précision et de rappel. Les valeurs F-mesure sont dans l'intervalle [0,1] et les grandes valeurs F-mesure correspondent à une meilleure qualité de classification. La mesure populaire pour l'évaluation des systèmes de classification et est le plus souvent utilisé pour comparer les performances des classificateurs.

9 Fouille d'opinions en langue arabe

Contrairement à la langue anglaise, la langue arabe n'a pas attiré beaucoup d'attention des chercheurs dans le domaine d'opinion-mining. Plusieurs raisons peuvent être à l'origine de l'absence de travaux de recherche : l'arabe est syntaxiquement et morphologiquement complexe qui provoque une ambiguïté lexicale et sémantique élevée, surtout quand les diacritiques ne sont pas employés, un manque remarquable de ressources linguistiques pour l'arabe comme les lexiques de sentiments et les ressources subjectives sur le Web pour la création des corpus.

Dans cette section, nous présentons quelques travaux de recherche en fouille d'opinions qui se focalise sur la langue arabe et qui peuvent donner plus de motivation à notre travail de recherche :

Dans [107], les auteurs ont proposé une approche pour la classifications des commentaires Facebook, écrits dans la langue de l'argot arabe, basée sur l'utilisation d'un classificateur SVM, leur approche utilise un lexique construit manuellement regroupant les mots de sentiment de l'argot arabe, le lexique a été utilisé pour extraire les mots d'opinion dans les commentaires, ce qui a amélioré le résultat de classification par rapport à l'utilisation d'un lexique de sentiments classique.

L'approche proposée dans [108], centrée autour de la création d'un lexique de sentiments pour la langue arabe en traduisant les lexiques disponibles pour l'anglais vers l'arabe.

Le travail de recherche présenté dans [109], a porté sur l'extraction du détenteur de l'opinion (ang. Opinion holder). Les chercheurs ont combiné des techniques d'apprentissage automatique et des règles linguistiques pour identifier les détenteurs d'opinion. Le corpus utilisé a été créé manuellement à partir des sites Web de la presse arabe.

L'auteur a présenté dans [110] une approche pour extraire les opinions dans les documents arabe, en combinant plusieurs techniques. Les résultats ont montré que les performances de classification s'améliorent lorsque de nombreux classificateurs collaborent.

[111] présente une approche de niveau phrase à base de traits pour l'analyse des sentiment arabes en utilisant des idiomes arabes comme une importance clé pour améliorer la détection de la polarité du sentiment dans les phrases arabes ainsi que d'un certain nombre de romans et un ensemble riche de traits linguistiquement motivés (intensificateurs contextuels et manipulation de négation), les traits syntaxiques pour les phrases contradictoires qui améliorent la pertinence de la classification de sentiment.

10 Travaux connexes à l'identification du genre

Classer les attributs utilisateur latents sur Twitter a été présenté par [112], en annotant manuellement 500 utilisateurs anglophones dont le genre a été marqué. Leur travail est le premier de son application de type attribut-auteur-latent découverte sur les données Twitter. Leur travail consiste à classer les attributs d'utilisateur latents pour déterminer leur genre, leur orientation politique, l'âge et l'origine régionale.

Burger et al [113] affirment que des éléments de profil ne sont pas très utiles pour prédire le genre lors de l'utilisation des approches d'apprentissage supervisé pour classer les tweets dans le but d'identifier le genre. Au lieu de cela, ils ont passé au peigne fin les utilisateurs qui ont inclus un URL dans leur champ de profil, ce qui leur a permis de créer un lien vers les blogs et autres médias en ligne qui identifient clairement le genre.

Employant des réseaux de neuronnes et des techniques de data mining, William et al [114] ont été en mesure d'identifier le genre des utilisateurs de Twitter. Ils ont démontré que les réseaux de neurone basées sur les flux pourraient automatiquement discriminer le genre d'un utilisateur sur les tweets marqués manuellement.

Il y a beaucoup dans la littérature sur l'utilisation du contenu du message d'état de l'utilisateur (à savoir, tweets) pour déterminer le genre, Cependant, Wendy et Derek [115] ont fait valoir que l'utilisation de seulement le premier nom de l'utilisateur peut prédire le genre des utilisateurs de Twitter. Ils ont en particulier fait valoir que les travaux décrits dans [113] n'étaient pas représentatifs de la base d'utilisateurs du réseau social Twitter.

En approchant de la tâche de classification de genre à partir d'un point de vue plus réactif et efficace, Wendy et al [116] ont appliqué [117] approche à un domaine spécifique. Leur travail a porté sur l'identification de la composition par genre de la population à Toronto, au Canada, en utilisant le transport, tels que les voitures, les transports en commun et les vélos, faire la navette.

Clay et al [118] a présenté un exemple spécifique à la région pour identifier le genre des utilisateurs de Twitter à partir du contenu de leurs tweets. Ils ont utilisé des traits unigramme à partir des tweets seulement, en omettant d'autres traits prédictives qui ont prouvé leur utilité dans la détermination de la composition par genre des utilisateurs de Twitter.

Bauman et al [119] a fait valoir contre l'utilisation de l'information pour identifier le genre des utilisateurs de Twitter comme une variable binaire.

11 Conclusion

Dans ce chapitre, nous avons présenté un état de l'art sur la classification d'opinions, qui nous intéresse particulièrement dans notre travail de recherche.

Nous avons dressé l'état de l'art des méthodes existantes supervisées et non supervisées telles que les méthodes basées sur les patrons syntaxiques et les méthodes linguistiques, les méthodes basées sur l'apprentissage automatique et les méthodes basées sur la fonction de score, et les méthodes hybrides qui mêlent les méthodes linguistiques et les méthodes basées sur l'apprentissage machine dans différents niveaux de granularité.

Les méthodes basées sur les lexiques offrent une analyse fine mais ont besoin toujours d'une intervention humaine que les méthodes statistiques et nécessitent des connaissances sur le langage. Les méthodes hybrides permettent de trouver un compromis : minimisation de l'intervention humaine tout en permettant d'analyser les opinions de manière précise.

Deuxième partie

Contribution

Chapitre 3

GenderSentiment : Approche proposée

Sommaire

1 Introduction 68
2 Motivation 69
3 Description 70
4 Entrées de notre système 70
4.1 Profils utilisateurs 70
4.2 Lexique des sentiments 71
4.3 Textes évaluatifs 74
5 Architecture 76
5.1 Détection du genre 76
5.2 analyse des sentiments 79
5.3 Évaluation des méthodes de classification 83
5.4 Architecture générale 83
6 Conclusion 84

1 Introduction

Dans les dernières années, et notamment avec l'apparence du Web 2.0 qui a permet aux développeurs de mettre à la disposition des internautes des outils simples et efficaces comme les réseaux sociaux et les blogs, l'information est devenue accessible et à la portée de tout le monde. À travers ces outils, les utilisateurs peuvent poster librement des commentaires, des jugements, ou des critiques sur une personne, un service ou un produit.

Pour bénéficier des textes évaluatifs disponibles, en vue d'améliorer la qualité d'un produit ou d'un service offert aux clients ou de connaître l'avis du grand public sur une personne politique, etc., les entreprises, les politiciens, ainsi que les clients ont besoin d'outils puissants pour suivre les opinions, les sentiments, les jugements et les croyances qui se cachent derrière ces textes qui se diffèrent dans les styles d'écriture qui se produisent entre le genre, l'âge et les groupes sociaux.

Le genre est une question de vie importante. Surtout quand nous ferons face aux défis de la vie, nous avons tendance à prendre en compte des solutions spécifiques qui correspondent le mieux que le genre. Quand la discussion va de la mode, à la santé, à la richesse ou à l'emploi, la distinction de genre fait partie intégrante. Dans chacun de ces sujets, la discussion diffère en ce qui concerne le genre, afin de comprendre et d'offrir la meilleure solution unique à chaque genre.

La nouvelle discipline, fouille d'opinion et analyse des sentiments (ang. Opinion Mining and Sentiment Analysis), est née pour répondre à la demande croissante en qualité d'analyse d'opinions, en mettant en place des systèmes capables de traiter automatiquement des textes évaluatifs, et rechercher les valeurs prédictives des jugements. Dans ce contexte, plusieurs approches ont été proposées par les chercheurs, et plusieurs ressources lexicales ont été construites en prenant pas compte au genre.

La fouille d'opinions est un domaine qui s'intéresse essentiellement à la détection et la catégorisation d'opinions. La prise en compte du contexte des mots dans les évaluations, le domaine étudié, ainsi la complexité grammaticale et morphologique de certaines langues comme l'arabe dans notre étude, peuvent apporter une forte ambiguïté, pas seulement au niveau sémantique mais aussi au niveau subjectif, qui rend difficile la tâche d'identification des opinions et leurs cibles.

Les approches, qui se basent sur les lexiques des sentiments (Sous section 3.1.2 du chapitre 2) pour détecter les indices de subjectivité dans les textes, attribuent une polarité positive ou négative à chaque entrée sans la prise en compte du domaine étudié et du genre du détenteur d'opinion.

Contrairement à ces approches, nous proposons dans notre étude une approche de classifi-

cation d'opinions par genre qui vise essentiellement à connaitre le genre d'utilisateur arabe puis une classification d'opinion selon le genre en se basant sur un domaine spécifique, où l'attribution de polarité aux mots de lexique dépend fortement du domaine étudié.

Nous pouvions résumer les grandes lignes de notre approche dans les points suivants :

- Classification manuelle des profils utilisateurs selon leurs genre ;
- Extraction des traits des profils utilisateurs ;
- Extraction des traits de comportement de messagerie ;
- Extraction des traits linguistique ;
- Association de paires utilisateur-traits ;
- Classification des utilisateurs selon leurs genre ;
- Annotation manuelle des textes évaluatifs des sentiments selon une polarité (négatif, positif) ;
- Identification des textes subjectifs en employant le lexique de sentiments ;
- Extraction des traits explicites du domaine étudié, à partir des textes évaluatifs,
- Association de paires opinion-traits ;
- Classification des opinions identifiées au niveau des documents *(ang. Document-level classification)* ;
- Statistiques opinion-genre.

Le chapitre est organisé comme suit : nous présentons dans la section 2, les motivations de la démarche de notre travail en citant les principales approches proposées dans le domaine de détection du genre et de fouille d'opinions. Dans la section 3, nous présentons une description générale de notre approche. Nous décrivons dans la section 4, les entrées de notre approche qui sont : les profils des utilisateurs, le lexique des sentiments, et l'ensemble des textes évaluatifs. Enfin, dans la section 5, nous expliquons l'architecture de notre approche et ses différents modules illustrés par des exemples.

2 Motivation

Notre étude sur l'identification du genre des utilisateurs montre qu'aucun travail sur des textes arabes n'était fait.

Le domaine de fouille d'opinions réunis différentes problématiques de recherche : construction de ressources lexicales [54, 120, 121], classification d'opinions [13, 9, 12, 72, 71, 122], détection de passage d'opinions et leurs catégorisations sémantiques [123, 124], résumé automatique de textes d'opinion [72, 125], etc.

Dans le domaine de fouille d'opinions, deux grandes familles d'approches se distinguent : celles qui font une simple extraction des caractéristiques d'objets [46, 126, 124, 122] et celles

qui les organisent en une hiérarchie à l'aide des taxonomies [127, 92, 128] ou ontologies [129, 125, 130, 72].

Le processus d'extraction des caractéristiques concerne principalement les caractéristiques explicites.

3 Description

Dans la section précédente 2 de ce chapitre, nous constatons que les chercheurs qui ont travaillé sur l'identification du genre n'abordent pas la langue arabe et les utilisateurs arabophones. Cependant, 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 détection du genre et le processus de classification d'opinion.

Dans le contexte de notre approche, nous proposons une approche de classification d'opinions par genre dont l'objectif de la combinaison des deux domaines est relatif à deux aspects :

- Faire des statistiques sur les opinions mâle et femelle ;
- Utiliser la classification d'opinion comme un trait qui améliore le système du détection du genre dans le futur et selon le domaine d'étude.

Nous avons opté pour une architecture modulaire, où chaque module est décrit séparément des autres, qui enfin vont coopérer pour décrire l'architecture générale du système.

4 Entrées de notre système

Nous découpons les entrées de notre approche autour des trois axes suivants : la collecte des profils utilisateurs, la construction de lexique des sentiments et la collecte des textes évaluatifs.

4.1 Profils utilisateurs

Notre système s'appuie sur l'utilisation des informations des utilisateurs. Nous décrivons en détail trois types d'informations qui peuvent aider à caractériser un utilisateur d'un microblog[131] : profil, comportement de messagerie (tweeting), contenu linguistique des messages.

Nous utilisons ces trois types d'informations pour dériver un riche ensemble de traits pour une utilisation dans un modèle de classification de l'utilisateur.

Notre objectif est double : premièrement, de fournir une évaluation générale de la valeur relative, la robustesse et la généralisation potentielle des traits dans le but d'une classification

de l'utilisateur et d'autre part, d'explorer la valeur de l'information linguistique pour classer les utilisateurs.

4.1.1 Traits du profils

La plupart des services (tels que Twitter) montrent publiquement des informations de profil tels que le nom d'utilisateur, et une courte biographie par défaut. L'API Twitter permet également d'accéder à d'autres informations de base de l'utilisateur, tels que l'Url, les tweets et les retweets d'un utilisateur.

4.1.2 Traits de comportement de messagegerie

Le comportement des tweets se caractérise par un ensemble de statistiques capturant la façon dont l'utilisateur interagit avec le service de micro-blogging : le nombre moyen de messages par jour, le nombre de réponses, etc. Intuitivement, une telle information est utile pour construire un modèle de l'utilisateur.

4.1.3 Traits linguistiques

Le contenu linguistique résume les principaux sujets d'intérêt pour l'utilisateur, ainsi que l'utilisation lexicale de l'utilisateur. Une information linguistique simple est utile pour classer les utilisateurs en mâle ou en femelle.

4.2 Lexique des sentiments

Pour exprimer nos idées, nos préférences ou nos opinions sur un sujet, ou sur un objet ou l'un de ses caractéristiques, nous utilisons un langage évaluatif. Ce langage incorpore différents éléments linguistiques qui peuvent permettre la détection et l'analyse de la subjectivité.

Ces éléments découlent de la théorie de l'évaluation cognitive *(ang. Appraisal Theory)* qui suggère que les opinions et les sentiments en générale proviennent des évaluations cognitives personnelles que l'individu exprime sur un objet ou un évènement.

L'ensemble des mots qui expriment l'opinion peuvent être subdivisés en catégories représentant la modélisation. La modélisation se définit comme l'inscription dans le discours de la prise d'attitude du locuteur à l'égard du contenu d'un énoncé. Les cinq catégories de modalité de l'évaluation sont :

- **Opinion** : un fait présupposé est évalué par le locuteur qui révèle du même coup son point de vue ;
- **Jugement** (favorable ou défavorable) : le locuteur pose dans son énoncé, une action réalisée et juge si cet acte est bon ou mauvais ;
- **Accord** ou **désaccord** : nous présupposons qu'une demande a été adressée au locuteur qui va dire s'il adhère ou non à la vérité d'un propos tenu par un autre ;

- **Appréciation** (jugement intellectuel) : le locuteur évalue la valeur d'un fait en révélant ses propres sentiments selon son champ d'appréciation ;
- **Acceptation** ou **refus** : nous présupposons qu'une demande d'accomplissement d'un acte a été adressée au locuteur qui peut répondre favorablement ou non à cette demande.

Dans chacune de ses modalités, la valeur peut être graduée selon sa force : soit du négatif au positif dans les cas de l'opinion, du jugement ou de l'appréciatif soit du plus ou moins pour l'accord/désaccord et acceptation/refus.
À ces modalités, viennent se combiner trois types de vocabulaire de l'opinion :

- Vocabulaire **affectif** (l'ensemble des mots impliquant une réaction émotionnelle : effrayant, pleurer, séduire, étonnant) ;
- Vocabulaire **appréciatif** (impliquant un jugement de valeur positif ou négatif : intéressant, de qualité, trop cher, banalité) ;
- Vocabulaire **connoté** (mots possédant une signification affective en plus de son sens premier : force, liberté, original).

Trancher entre le caractère objectif ou le caractère subjectif d'un texte revient à détecter dans le texte les indices de la présence du subjectif. Plusieurs indices peuvent en fait témoigner de la présence de la subjectivité dans un texte ou une phrase. La présence d'adjectifs constituerait un bon indicateur du caractère subjectif des phrases [132].

Enfin, certains ont tenté de séparer les opinions des faits, au niveau du document et au niveau de la phrase, en utilisant un classificateur naïf Bayes [50].

Bing Liu [6] admet aussi qu'en matière d'analyse de l'opinion, la distinction entre l'objectif et le subjectif n'est pas forcément pertinente. Ainsi, une phrase contenant de la subjectivité peut ne pas exprimer une opinion ni positive ni négative. Ex. : Je pense qu'elle me l'a dit. Aussi, une phrase objective peut être porteuse d'opinion, ce qui n'est pas vraiment traité par l'analyse des sentiments. L'expression peut comporter une opinion implicite ou transmettre une information qui peut produire chez le lecteur des effets négatifs d'appréciation de la cible. Ex. : Mon téléviseur X est tombé en panne hier.

Traditionnellement, l'extraction d'opinion dans les textes est basée sur la recherche des adjectifs, pour cela, les méthodes existantes sont souvent basées sur des dictionnaires généraux. Malheureusement, ce type d'approche trouve ses limites : pour certains domaines, des adjectifs peuvent être inexistants, voire contradictoires [123].

Les approches lexicales utilisent des dictionnaires de mots subjectifs, considérés comme des références universelles à partir de l'anglais. Ces dictionnaires peuvent être généraux comme le *General Inquirer* [1], *Sentiwordnet* [2], *Opinion Finder* [3], *NTU Sentiment Dictionary (NTUSD)* [4].

1. http ://www.wjh.harvard.edu/ inquirer/
2. http ://sentiwordnet.isti.cnr.it/
3. http ://mpqa.cs.pitt.edu/opinionfinder.html
4. http ://onlinelibrary.wiley.com/doi/10.1002/asi.20630/full

Dans ces dictionnaires, une polarité est associée a priori à chacun des mots. Quel que soit le contexte dans lequel il sera inséré, le mot devrait ainsi avoir toujours la même polarité. Nous donnons ensuite au document un score d'opinion en fonction de la présence de mots issus de ces dictionnaires dans le texte.

4.2.1 Initialisation du dictionnaire

Dans notre approche, nous proposons de construire un lexique de sentiments pour notre domaine étudié . Nous croyons que la séparation des lexiques par domaine peut diminuer considérablement l'ambiguïté sémantique des mots d'opinion.

Par exemple le mot "arabe" (petit) peut être positif dans un domaine et négatif dans un autre.
Exemple :

"texte arabe "(le mobile est petit) "arabe"(petit) est positif dans le domaine d'électronique.

"texte arabe "(la chambre est petite)"arabe"(petit) est négatif dans le domaine d'urbanisme.

Une idée inspirée des travaux de Harb et al. (2008) [123] consiste à construire un lexique avec les mots de sentiments les plus connus dans le domaine étudié. Pour cela, nous considérons deux ensembles P et N de mots germes classiquement utilisés dans la littérature dont les orientations sémantiques sont respectivement positives et négatives.

Exemple :
"texte arabe ""texte arabe ""texte arabe " aimable, correct, positive, excellent, beau, bien ;
"texte arabe ""texte arabe ""texte arabe " désagréable, faux, mauvais, laid, mauvais ;

Le dictionnaire 4.1 aura alors pour chaque entrée : le mot de sentiment et sa polarité positive ou négative.

Mots de sentiment	Polarité
	+
	+ -
	-
...etc.	

TABLE 4.1 – Structure simplifiée d'un dictionnaire des sentiments

4.2.2 Enrichissement du dictionnaire

Certaines approches de détection des sentiments utilisent des dictionnaires généraux existants [122, 121] ou construits manuellement [120], d'autres utilisent des dictionnaires construits automatiquement, comme dans [123], où les auteurs ont proposé une nouvelle approche de création automatique de dictionnaire d'adjectifs qui intègre les connaissance du domaine, puis l'enrichissement de ce dictionnaire à partir d'un corpus d'apprentissage.

Afin d'établir des associations entre les différentes expressions d'opinion pour enrichir le dictionnaire d'opinions, il est tout d'abord nécessaire de connaître la fonction grammaticale de chacun des mots du corpus d'apprentissage. Pour ce faire, un outil d'étiquetage automatique de textes qui attribue à chaque mot une catégorie grammaticale et fournit les mots sous une forme lemmatisée (forme canonique) est indispensable dans ce cas.

En fait, la création des ressources lexicales constitue un sous-domaine de recherche à part [123, 120, 121, 25], et pour ne pas s'écarter de notre sujet de recherche qui focalise sur l'identification d'opinions , nous ne donnerons pas une grande importance à la création du dictionnaire des sentiments, et nous nous limiterons à l'adaptation des lexiques existants, qui regroupe des listes prédéfinis d'expression d'opinion.

4.3 Textes évaluatifs

Afin d'évaluer l'efficacité de notre approche, nous aurons besoin d'un ensemble de textes évaluatifs propres au domaine étudié. Le réseau Internet est considéré comme la source la plus riche pour collecter les textes à caractère subjectif : les réseaux sociaux comme Facebook et Twitter, les blogs, les sites de commerces électroniques,...etc. Les méthodes de collecte des textes peuvent être manuelles ou automatiques.

Nous supposons qu'une opinion est exprimée dans le même segment dans lequel les traits sont présents, i.e. les traits et les expressions d'opinion sont près les uns des autres. Dans ce cas, la segmentation du texte joue le rôle d'éloigner les traits inutiles des expressions d'opinion.

Pour extraire les unités lexicales et déterminer si une unité est un trait ou un mot d'opinion, plusieurs tâches linguistiques peuvent être appliquées :

4.3.1 Segmentation

Pour analyser un texte, nous devons mener à bien sa segmentation en paragraphes, phrases et propositions. Dans notre travail, nous avons utilisé MADAMIRA pour segmenter le texte en segments bruts. La figure suivante illustre un texte arabe segmenté en segments bruts.

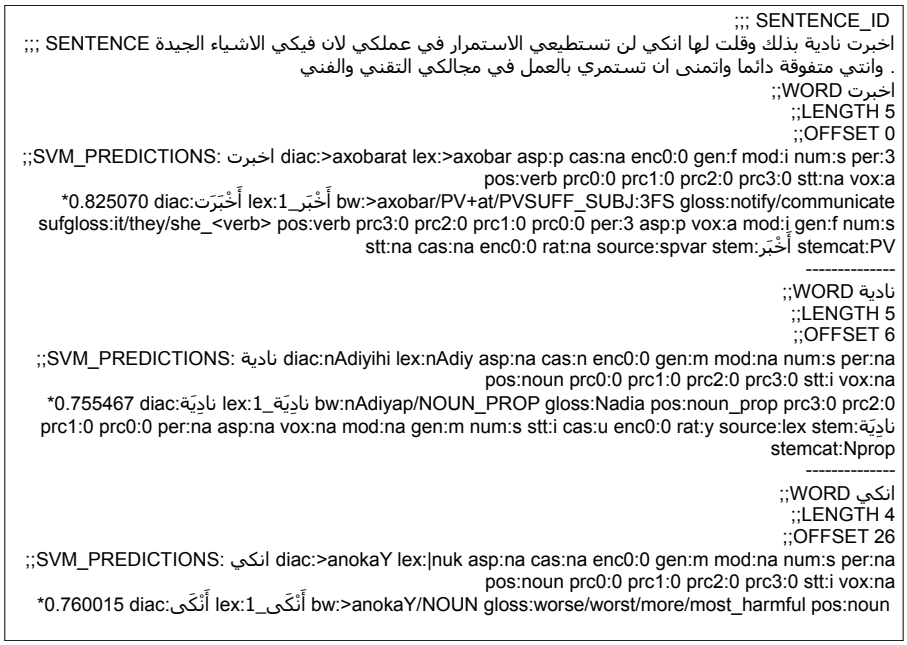


FIGURE 3.1 – Exemple d'un texte arabe segmenté

4.3.2 Étiquetage

Après la segmentation du texte (étape précédente), nous devons associer à chaque terme son étiquette lexicale. L'étiquetage *(ang. POS Tagging : Part-Of-Speech Tagging)* est le processus de marquage des unités lexicales, i.e. l'attribution des classes grammaticales aux unités lexicales. Dans notre travail de recherche, cette étape est primordiale, car sans étiquetage des unités lexicales nous ne pouvons pas extraire ni les expressions d'opinion ni les traits linguistiques. Nous avons utilisé MADAMIRA POS Tagger pour localiser toutes les unités lexicales dans le texte. La figure suivante montre un simple texte arabe étiqueté.



FIGURE 3.2 – Exemple d'un texte arabe étiqueté

5 Architecture

Nous détaillons dans ce qui suit les différents modules de notre approche.

5.1 Détection du genre

Un profil utilisateur est composé de plusieurs segments qui sont organisé dans notre étude selon trois catégories :

5.1.1 Extraction des traits du profils

Les informations d'un profil se diffèrent entre les gens ce qui peuvent révéler beaucoup sur eux tels que :

- **Numéro de téléphone :** Nous trouvons généralement dans la société arabe que les femmes évitent de mettre un numéro de téléphone en public par contre les hommes qui peuvent le mettre sans aucun problème.
- **Instagram** [5] **:** Selon notre recherche, nous avons constaté que les femmes utilisent beaucoup Instagram par rapport aux hommes.
- **Émotions :** La tendresse de la femme rend elle attiré aux émotions plus qu'aux hommes et cela peut faciliter la classification du genre.
- **Couleurs :** Ce sont des traits utile pour la classification du genre. [133] ont fait une classification du genre sur Twitter indépendante aux langues qui s'intéresse aux couleurs sachant que : la couleur d'arrière plan, la couleur du texte et des liens, la couleur du remplissage d'un SideBar [6] et la couleur de la bordure d'un SideBar.
- **Luminosté des couleur :** se diffère d'un compte à l'autre.

5.1.2 Extraction des traits de comportement de messagerie

Des calculs sont faites pour avoir un comportement générale sur un compte utilisateur tels que :

- **Nombre des messages :** Nous avons besoin de compter le nombre total des messages.
- **Nombre des réponses :** Nous avons besoin de compter le nombre total des réponses.
- **Nombre des messages et réponses contenant des émotions :** Cela aide à faire d'autres calculs qui sont indispensable pour la classification.
- **Moyenne des messages et réponses contenant des émotions :** à l'aide des calculs précédents nous pouvons effectuer des calculs de moyennes qui facilite l'interaction avec ces derniers.

5. https ://www.instagram.com/
6. **SideBar :** un court article dans un journal ou un magazine, généralement en boîte, placée à côté d'un article principal, et contenant du matériel supplémentaire ou explicatif.

5.1.3 Extraction des traits linguistiques

L'utilisation lexicale des utilisateurs et le style linguistique peuvent fournir des conseils utiles pour classer les utilisateurs tels que :

- **Formule féminine :** montre la féminine de la personne.

  Exemple :

  (ingénieur)
  (étudiante)
- **Destinataire de la formule de féminisation :** nous avons utilisé la formule de féminisation afin de savoir le genre d'un destinataire en utilisant les réponses qu'ils reçoit.

  Exemple :

- **Mots prototypes :** ces mots nous aide à différencier entre les intérêts d'un homme et d'une femme sachant que les femmes s'intéressent beaucoup à la mode, au shopping ...etc, et les hommes sont généralement intéressé au sport et à la politique... etc, nous avons aussi trouvé que les arabophones ont des appellations qui peuvent faire une distinction entre les genres.

  Exemple :

  indique que la personne est mâle.

  indique que la personne est femelle.

5.1.4 Association des traits avec les utilisateurs

Cette étape cherche à associer tous les utilisateurs avec les traits extraits. Autrement dit, nous devons associer à chaque U l'ensemble des traits, sous forme de couples(Ui, Fj).

5.1.5 Classification

Après l'extraction de tous les couples (U_i, F_j) ie. U_i est un utilisateur et F_j est l'ensemble des traits ou features, nous arrivons à la dernière étape de ce module qui est la catégorisation des utilisateurs selon leurs genre. Notons qu'à un utilisateur mâle est assigné à une classe qui est égale à 1 et un utilisateur femelle est assigné à une classe 0. Si le résultat donne 1, alors l'utilisateur sur F est mâle. Si le le résultat donne 0, alors l'utilisateur sur F est femelle.

5.1.6 Classificateur

Il existe beaucoup de méthodes de classification supervisée dans la littérature. Nous avons utilisé le classificateur naïf bayes qui supporte la classification binaire. En regroupant les segments présentés précédemment, nous obtenons l'architecture de la détection du genre suivante :

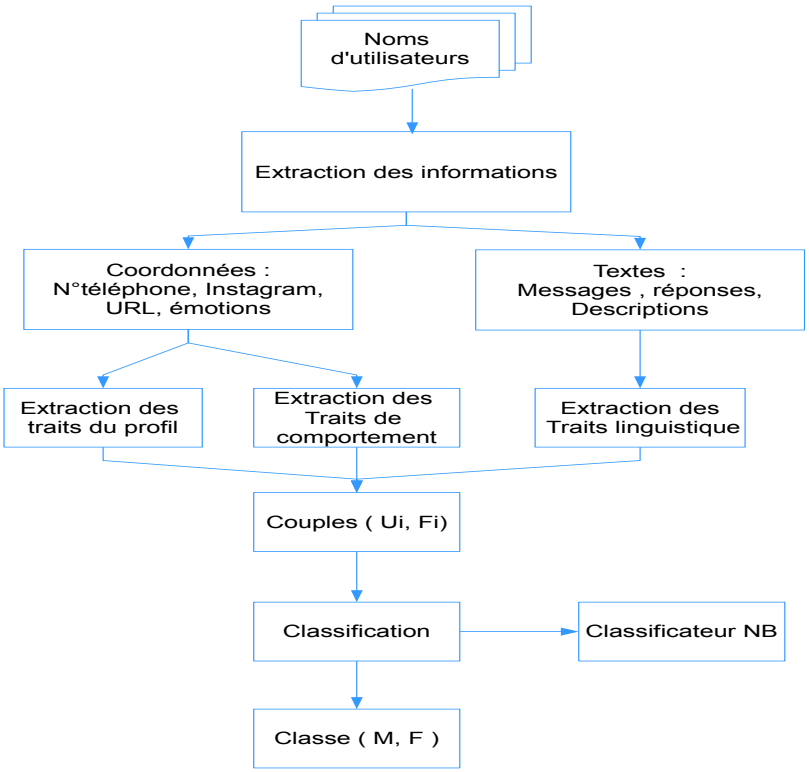


FIGURE 3.3 – Architecture de la détection du genre

5.2 analyse des sentiments

Un texte évaluatif est composé de un ou plusieurs segments, chaque segment peut être subjectif ou objectif, chaque segment subjectif contient au moins une expression d'opinion (EO), notons qu'une EO est une expression explicite exprimée sur un objet, elle est composée d'un nom, un adjectif, ou un verbe avec certains modificateurs comme les négations et les adverbes.

Exemple :

Ce simple exemple contient une expression d'opinion composée d'un adjectif(belle) et un modificateur

Notre approche qui s'intéresse à la classification d'opinion, a besoin de deux composants de base :

— Un lexique de sentiments : comme nous l'avons vu précédemment, il s'agit d'une ressource lexicale L des expressions d'opinion propres à un domaine étudié ;

— Un ensemble de textes évaluatifs T collectés d'Internet.

Nous détaillons dans ce qui suit les différents segments de ce module.

5.2.1 Identification des textes subjectifs

Nous rappelons que l'EO est la plus petite unité au sein d'un segment subjectif (SS). Elle est composée d'un seul mot d'opinion (un nom, un adjectif ou un verbe), éventuellement associé à certains modificateurs comme les mots de négation et les adverbes. Par exemple :

Nous croyons qu'il est difficile, voire impossible de créer un lexique regroupant tous les mots de sentiments propres a un domaine étudié, et le problème qui se pose dans ce cas est comment identifier les autres mots de sentiments dans les textes évaluatifs sans leur présence dans le dictionnaire ?

Dans notre approche, et pour extraire les textes porteurs d'opinions explicites, nous avons proposé de faire une projection simple du dictionnaire de sentiments sur les textes évaluatifs, et nous avons fait un enrichissement du dictionnaire manuellement qui repose sur l'expertise humaine afin d'avoir le maximum de textes subjectifs. La figure suivante illustre le processus d'identification des textes subjectifs avec un enrichissement du lexique.

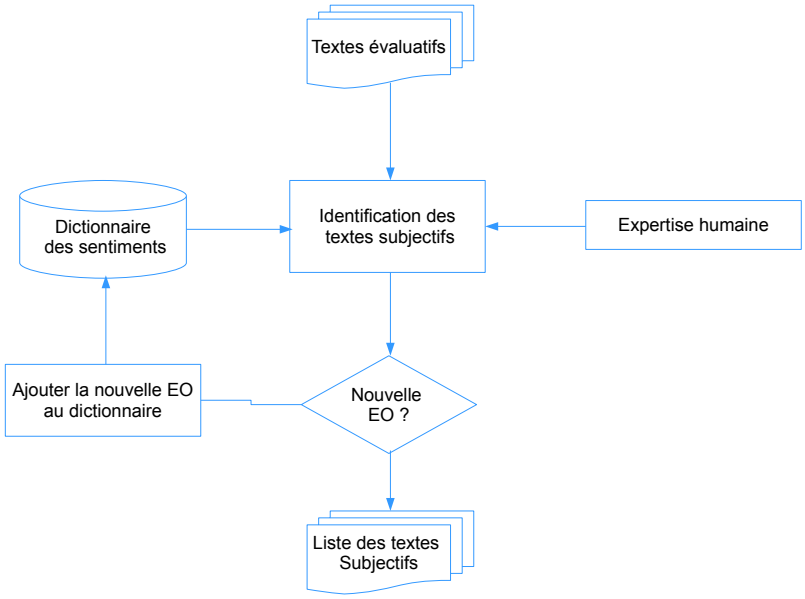


FIGURE 3.4 – Processus d'identification des textes subjectifs avec enrichissement de lexique des sentiments

5.2.2 Extraction des traits explicites

Dans cette section, nous nous intéressons spécifiquement à l'extraction des traits d'une opinion. Cette étape vise à extraire à partir des textes évaluatifs toutes les expressions d'opinions. Étant donné que chaque EO a une polarité positive ou négative selon leur emplacement dans le texte et selon plusieurs autres traits.

- **Traits standard**
  Pour la classification d'opinions, nous appliquons deux traits principales binaires pour les mots d'opinion, qui déterminent si le mot a une polarité de sentiment positif ou négatif en utilisant notre lexique. Et d'autres traits qui sont tout simplement le calcul de nombre de mots positifs et négatifs de sentiment dans le sujet.
- **Traits linguistiques**
  Nous avons exploré une grande variété de traits linguistiques et lexicales qui peuvent gérer avec précision les structures linguistiques complexes dans la langue arabe comme la négation, les modificateurs, etc. Dans ce travail, nous avons utilisé l'ensemble des traits suivants :
  - **Négation :** La manipulation de la négation peut être une préoccupation importante dans l'opinion et l'analyse des sentiments. Par exemple : (j'aime ce livre) ou (je n'aime pas ce livre). les deux phrases considérées comme très similaires par des mesures de similarité les plus couramment utilisés, la seule différence est le terme de négation, qui rend la deuxième phrase en polarité opposée. Donc, il y a une liste de termes de négation en langue arabe standard et la dialecte saoudienne qui peut changer la polarité du sentiment des termes du positif au négatif et vice versa lorsque l'un d'entre eux se sont produits avant le mot d'opinion. Nous avons utilisé

deux traits, à savoir : une trait binaire indique qu'il a une négation dans la phrase et un trait pour exprimer combien de termes de négation apparaissent dans la phrase.

— **Questions :** Les questions sont généralement destinés à la recherche de quelque chose que vous avez besoin ou ne pas exister. Dans les dialogues de sentiment, les questions ont la même signification, mais expriment un sentiment négatif parce que le détenteur d'opinion se sent triste ou en colère ou étonnant à cause du manque de ce qu'il a besoin de son point de vue. Ceci ce que les gens considèrent comme une façon ou une méthode efficace pour la dénonciation, le rejet et la condamnation. Par conséquent, nous considérons l'existence de la question dans une phrase donne une polarité négative dans la plupart des phrases que nous avons observées.

— **Exclamation :** Ce sont comme les questions, expriment un sentiment négatif parce que le détenteur d'opinion se sent surpris et nous avons constaté que la plupart des phrases exclamatives donne une polarité négative.

— **Fautes d'orthographe :** ne sont pas des uni-grammes standard et donc ne pouvaient pas être régulièrement présents dans les données d'apprentissage. Les critiques contenant un grand nombre de fautes d'orthographe auraient des traits significativement réduites et ainsi fournissent très peu d'informations pour la classification. Nous avons constaté que les fautes d'orthographe isolées et communes ne touchent pas beaucoup le classement, mais les critiques qui contiennent relativement beaucoup de fautes d'orthographe ont tendance à être mal classés. Parfois, les fautes d'orthographe pourrait être volontaire pour exprimer une sorte de stress et de l'intonation tels que "énnnnorme". Le problème avec ce type de mots est qu'ils sont irréguliers dans le corpus. Par exemple, "énnnnorme" est très positif mais il n'estpas présent dans l'ensemble de traits, il est donc pas utile. Les critiques très mal orthographié ont tendance à être mal classés.

5.2.3 Association des traits avec les textes subjectifs

Cette étape cherche à associer toutes les textes subjectifs extraits dans l'étape 5.2.1, avec les traits extraits dans l'étape 5.2.2. Autrement dit, nous devons associer à chaque TS l'ensemble des traits de l'opinion, sous forme de couples (Fi, TSj).

5.2.4 Classification

Après l'extraction de tous les couples (fi, tsj), nous arrivons à la dernière étape de ce module qui est la catégorisation des opinions au niveau des documents qui consiste à attribuer une valeur positive ou négative à un document en attribuant la polarité dominante. Notons qu'à un mot positif est assigné à une classe qui est égale à 1 et un mot négatif est assigné à une classe qui est ègale à 0. Si le score final est positif, alors l'opinion sur F est positive. Si le score final est négatif, alors l'opinion sur F est négative.

5.2.5 Classificateur

Beaucoup de méthodes de classification supervisée existent et beaucoup d'entre elles ont été testées pour la classification d'opinions. On peut citer les arbres de décision, les réseaux de neurones, la régression logistique, les règles de décision ainsi que des méthodes combinant différents classificateurs comme les systèmes de votes ou les algorithmes de Boosting. Toutefois, les méthodes les plus présentes dans la littérature, et qui semblent également être les plus performantes sur les textes, sont les machines à vecteurs de supports (SVM) [32, 123, 62, 63, 134] et les classificateurs bayésiens naïfs (NB) [62, 63]. Nous avons utilisé les classificateurs bayésiens naïfs qui sont des classificateurs bi-classes. En regroupant les segments présentés précédemment, nous obtenons l'architecture de le classification d'opinion suivante :

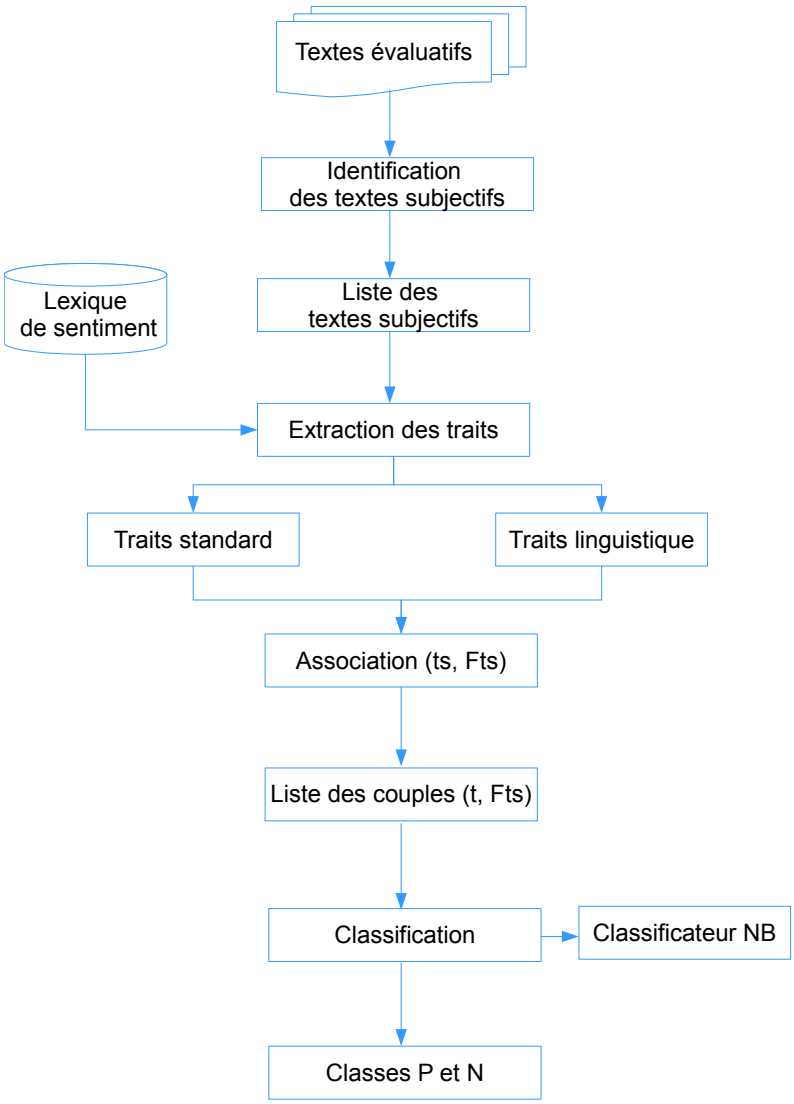


FIGURE 3.5 – Architecture d'analyse des sentiments

5.3 Évaluation des méthodes de classification

Comme nous l'avons mentionné dans le chapitre 2 de ce présent mémoire, pour mesurer l'efficacité d'un classificateur dans un problème à n classes, trois mesures principales sont utilisées : la précision, le rappel et le F-mesure.
Etant donné pour chaque classe i :

$$Précision_i = \frac{Objets\ correctement\ attribués\ à\ la\ classe_i}{Nombres\ d'objets\ attribués\ à\ la\ classe_i}$$

$$Rappel_i = \frac{Documents\ correctement\ attribués\ à\ la\ classe_i}{Nombre\ de\ documents\ attribués\ à\ la\ classe_i}$$

La précision indique le degré de vérité des résultats obtenus, tandis que le rappel indique leur pertinence. Il est intéressant de faire un compromis entre la précision et le rappel, c'est le rôle du f-mesure qui est le compromis équilibré entre la précision et le rappel.

$$F - mesure = \frac{2 \times Précision \times Rappel}{Précision + Rappel} \tag{3.1}$$

Habituellement, l'utilisation de ces mesures entre dans un processus de validation. Il y a deux types de validation : la validation par test et la validation croisée. Dans une validation par test, on apprend un modèle sur un corpus d'apprentissage et on test la classification sur un autre corpus. Dans une validation croisée, on ne considère pas précisément d'ensemble d'apprentissage et de test mais des sous-ensembles d'un corpus qui sont utilisés pour se valider les uns les autres. La validation croisée est considérée plus fiable qu'une simple validation par test.

5.4 Architecture générale

En regroupant les modules présentés précédemment, nous obtenons l'architecture générale de notre approche :

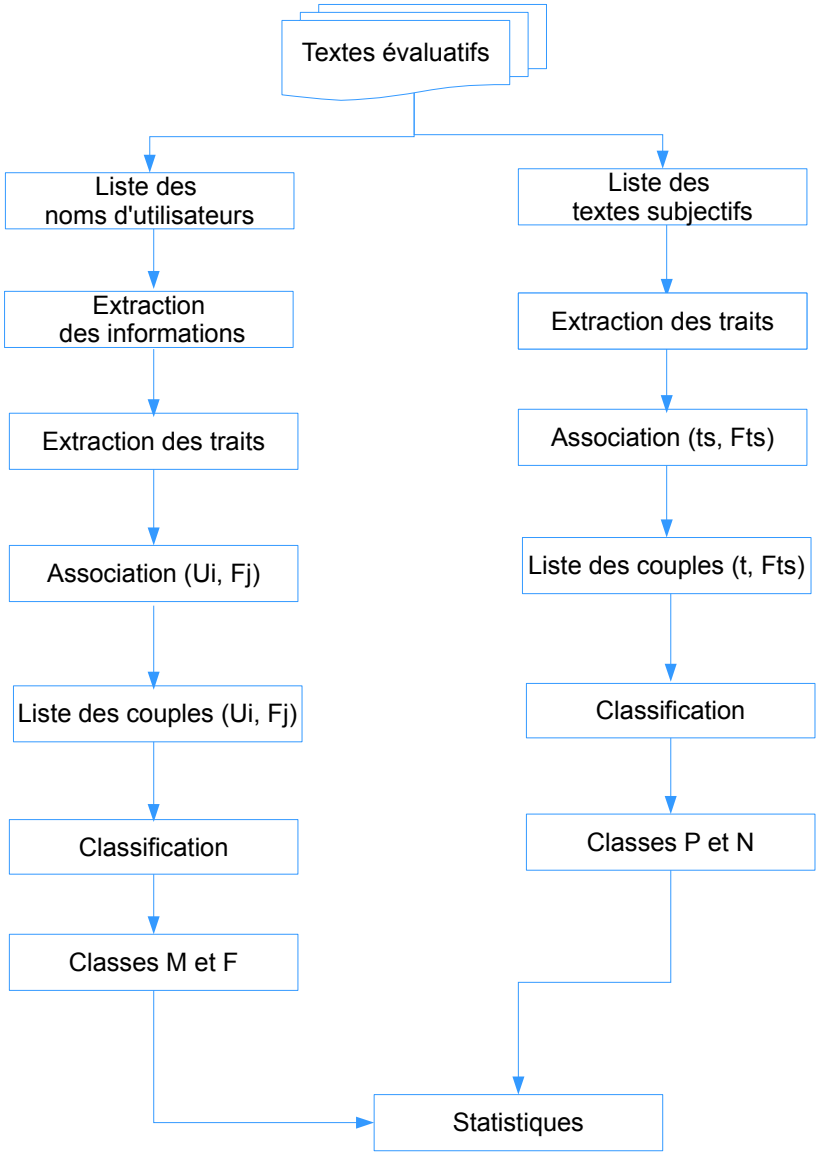


FIGURE 3.6 – Architecture d'analyse des sentiments

6 Conclusion

Ce chapitre présente une description générale de notre approche GenderSentiment. Elle se compose de deux modules principaux : détection du genre à partir des profils d'utilisateurs et analyse des sentiments à partir des textes évaluatifs en se basant sur un lexique des sentiments qui seront ensuite classifiée en faisant appel aux techniques les plus répondues dans le domaine de la détection de genre et de la fouille d'opinions.

Dans le chapitre suivant, nous présentons l'implémentation et l'évaluation de notre approche, en testant sa performance, soutenue par une étude de cas.

Chapitre 4

Implémentation et évaluation

Sommaire

1 Introduction 87
2 Choix du domaine étudié 87
3 Lexiques de sentiments 87
4 Corpus 88
5 Environnement de développement 89
5.1 Langage de programmation JAVA 89
5.2 Environnement de développement NetBeans 89
6 Outils utilisés 89
6.1 Twitter et Twitter4j 90
6.2 MADAMIRA 90
6.3 RapidMiner 91
7 Expérimentation et évaluation 91
7.1 Expérimentation 91
7.2 Classification et Évaluation 96
8 Conclusion 103

1 Introduction

Dans ce chapitre, nous présentons l'implémentation de notre approche GenderSentiment présentée dans le chapitre 3, puis l'évaluation et la discussion des résultats obtenus.

Comme nous l'avons mentionné dans le chapitre 3 de ce mémoire, la mise en œuvre de notre approche nécessite comme entrées : une collection des profils utilisateurs, un lexique reflétant les sentiments des personnes, qui mène à extraire leurs opinions envers les objets commentés et critiqués, et un corpus de textes pour évaluer notre approche.

2 Choix du domaine étudié

Le choix d'un domaine social pour notre étude parmi plusieurs a été motivé par :

- La richesse des textes arabes disponibles sur le net en matière de terminologie qui regroupe les concepts du domaine, notamment les micro-blogs où les détenteurs d'opinions de différents genre peuvent librement critiquer ou donner leurs opinions.
- Les intérêts qui peuvent être apportés par la conceptualisation de ce domaine se diffèrent par rapport au genre.
- D'après notre connaissance, le domaine social n'a pas subi une importance remarquable par les chercheurs notamment ceux qui travaillent dans le domaine de la fouille d'opinions.

3 Lexiques de sentiments

Un lexique d'opinions (ou lexique de sentiments) est une ressource linguistique qui associé à chaque mot sa polarité sémantique et éventuellement une intensité mesurant le degré de positivité et négativité.

Dans le domaine de fouille d'opinions, chaque mot d'opinion est considéré comme unité d'information de base fournissant un indice pour la détection de subjectivité dans le document. Beaucoup de ressources lexicales sont libres à utiliser : *Bing Liu's Opinion Lexicon* [1], *MPQA Subjectivity Lexicon* [2], *SentiWordNet* [3], *Harvard General Inquirer* [4], *LIWC* [5]... etc.
Malheureusement, les lexiques de sentiment pour la langue arabe disponibles sur Internet ne couvrent pas tous les domaines et un déficit significatif a été observé comme dans notre étude de cas, où de nombreuses expressions d'opinion ne sont pas présentes.

1. http ://www.cs.uic.edu/ liub/
2. http ://mpqa.cs.pitt.edu/
3. //sentiwordnet.isti.cnr.it/
4. http ://www.wjh.harvard.edu/ inquirer/
5. http ://www.liwc.net/

D'après notre recherche sur le net, deux lexiques de sentiments pour la langue arabe publiquement disponible qui sont *LABR : A Large Scale Arabic Book Reviews Dataset* [6], et must-read-sentimentAnalysis [7], nous avons fait une combinaison des deux lexiques qui comportent des mots d'opinion de la langue arabe standard et la dialecte égyptienne qui est un peu plus proche de la dialecte saoudienne utilisé dans notre étude puis nous avons l'enrichit manuellement par des mots d'opinions saoudiens. Le lexique est divisé en 2 catégories selon la polarité des mots : positif et négatif distribués dans deux fichiers ***Neg.txt*** et ***Pos.txt***.

4 Corpus

Twitter est un réseau social et une plate-forme de microblogging dont les utilisateurs peuvent publier des messages courts ou tweets (140 caractères). À la fin de juin 2016 [8], il a été estimé que Twitter compte 313 millions d'utilisateurs enregistrés dans le monde entier.

Twitter est un lieu attrayant pour la recherche dans les médias sociaux en raison de son grand volume, la population diversifiée et multilingue, et la nature généreuse de ses conditions d'utilisation. 4376 tweets ont été recueillies à partir d'un hashtag(fermeture des magasins à 9 h soir). Les utilisateurs de Twitter peuvent utiliser les hashtags (séquences de caractères préfixés par «#») pour désigner le sujet de leur tweets, plusieurs fois, les hashtags identiques ou similaires sont utilisés par les utilisateurs de Twitter afin de faciliter la récupération et le revêtement de l'information sur un sujet particulier.

Le manque d'informations sur le genre des profils Twitter nécessite la classification manuelle des utilisateurs de Twitter (détenteurs de tweets recueillis). Les comptes ont été filtrés pour éliminer tous les utilisateurs, où le genre n'a pu être déterminée. Ces profils comprennent les comptes twitter faites pour les entreprises, les organisations et les annonceurs. Tout profil où l'utilisateur ne parle pas l'arabe dans tous les tweets a également été supprimée car cette étude est seulement pour identifier le genre et les opinions des arabophones. En plus des tweets, chaque utilisateur Twitter a un profil avec les champs de texte libre suivantes :

- Screen name (ex. hadjerkebaili)
- Full name (ex. Hadjer OULD KEBAILI)
- URL (ex. site web utilisateur, page Facebook, etc.)
- Description (ex. Etudiante)

Twitter permet également d'accéder à d'autres informations de base de l'utilisateur, tels que le nombre d'amis, tweets et réponses d'un utilisateur. En utilisant l'API Twitter nous avons pu récupéré les informations et les tweets de 1154 comptes femelle et 1960 comptes mâle.

6. https ://github.com/mahmoudnabil/labr/tree/master/data/
7. https ://github.com/hadyelsahar/must-read-sentimentAnalysis/tree/master/resources
8. http ://www.alexitauzin.com/2013/04/combien-dutilisateurs-de-facebook.html

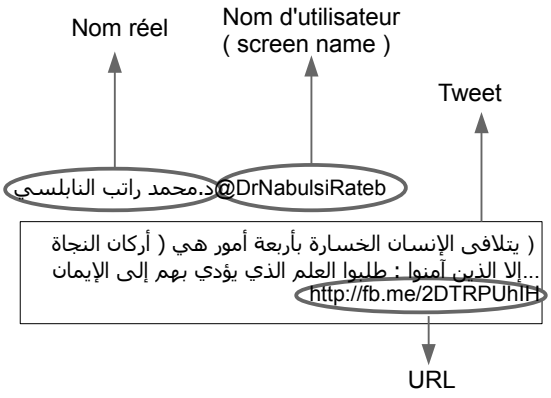


FIGURE 4.1 – Structure de base d'un tweet

5 Environnement de développement

Pour faire une expérimentation de notre approche de classification d'opinions par genre, nous avons choisi d'utiliser le langage de programmation suivant :

5.1 Langage de programmation JAVA

Java est un langage de programmation à usage général, évolué et orienté objet dont la syntaxe est proche du C. Ses caractéristiques ainsi que la richesse de son écosystème et de sa communauté lui ont permis d'être très largement utilisé pour le développement d'applications de types très disparates. Java est notamment largement utilisé pour le développement d'applications d'entreprises et mobiles.

Nous avons choisi d'utiliser le langage JAVA afin d'appliquer la librairie Twitter4j cité dans la section 6 implémenté en Java.

5.2 Environnement de développement NetBeans

Nous avons utilisé l'environnement NetBeans [9] 8.0.2 intégré (IDE). Il a été développé par IBM et considéré comme un IDE multi langages.

En plus de Java, NetBeans permet également de supporter différents autres langages, comme Python, C, C++, JavaScript, XML, Ruby, PHP et HTML. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web).

6 Outils utilisés

Afin de réaliser notre travail, nous avons utilisé les différents outils suivants :

9. https ://netbeans.org/

6.1 Twitter et Twitter4j

Twitter donne un accès gratuit au développeur à son REST API [10]. Cette API [11] permet aux développeurs d'accéder à la recherche et le téléchargement de texte tweet et métadonnées, ainsi que l'accès au flux Twitter, et l'accès à un flux de nouveaux tweets publics.

L'accès Twitter Firehose peut donner un accès utilisateur à acquérir un accès direct à tous les tweets publics sur Twitter, cette partie de l'API nécessite cependant un accès spécial généralement donnée par des gestionnaires tiers et des frais d'abonnement.

Lors de utilisation de l'API, Twitter demande que les développeurs suivent leurs " règles de développement " [12].

Twitter4J [13] est une bibliothèque de liaison Twitter API pour le langage Java sous la licence Apache 2.0. Avec Twitter4J, nous pouvons facilement intégrer notre application Java avec le service Twitter. Twitter4J est une bibliothèque non officielle.

La connexion a été mis en place avec la routine où ACCESS TOKEN et ACCESS SECRET sont reçus lors de l'exécution d'enregistrement pour l'accès des développeurs Twitter [14].

```
1 ConfigurationBuilder cb = new ConfigurationBuilder();
2 cb.setDebugEnabled(true)
3   .setOAuthConsumerKey("uWWKPjaVKqjx8v0hmGB2KD")
4   .setOAuthConsumerSecret("aYcUULSy03vP35nnNjhIBeX1nXvzij3N")
5   .setOAuthAccessToken("1332482792-dxQFbqeABQofeFFEd3rbsQeS")
6   .setOAuthAccessTokenSecret("tMbTG5LK9o04DLUrB6penT6fI58b");
```

Listing 4.1 – Connexion à Twitter API

6.2 MADAMIRA

MADAMIRA est la combinaison et le raffinement de deux outils précieux du Traitement du langage naturel arabe (NLP) : MADA et AMIRA. Il comprend également les fonctionnalités de MADA-ARZ, une version de MADA développé spécifiquement pour la dialecte égyptienne[135].

MADA est un système pour l'analyse morphologique et la disambiguation pour l'arabe. L'objectif principal de MADA est, étant donné un texte arabe, tirent autant d'informations linguistiques possibles sur chaque mot dans le texte, ce qui réduit ou élimine toute ambiguïté

10. https ://dev.twitter.com/overview/api
11. https ://dev.twitter.com/overview/api
12. https ://dev.twitter.com/terms/api-terms
13. http ://twitter4j.org/en/index.html
14. https ://apps.twitter.com/

entourant le mot.

AMIRA est un système pour la tokenization, l'étiquetage, Phrase à base Chunking (ang. Base Phrase Chunking) et la reconnaissance des entités nommées (ang. Named Entity Recognition). MADAMIRA aura finalement incorporer les composants BPC et NER de AMIRA, fournissant des informations qui peuvent être très utiles pour des tâches telles que l'analyse arabe.

MADAMIRA est conçu avec les objectifs d'être un remplacement fonctionnel pour MADA et AMIRA, étant indépendant de la plateforme, et offrant la possibilité de traiter le texte arabe à un rythme beaucoup plus rapide que les anciens outils. En outre, MADAMIRA fournit un support pour le fonctionnement du client-serveur qui permettra le traitement de l'arabe via HTTP. MADAMIRA est conçu pour traiter l'arabe moderne standard (MSA) et, dans cette version, la plupart de ses fonctionnalités est également disponible pour la dialecte égyptienne.

6.3 RapidMiner

RapidMiner [15] 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 classiffcation.

Nous avons utilisé ce dernier pour classer le genre de détenteur d'opinion puis pour classer les opinions elles mêmes.

7 Expérimentation et évaluation

Selon la structure modulaire de notre approche, nous avons évalué chaque étape indépendamment : détection du genre de détenteur d'opinion, analyse des sentiments, et enfin la classification des comptes twitter selon leurs genre (mâle, femelle) et la classification des tweets de sentiment selon la polarité associée en utilisant un classificateur NB.

7.1 Expérimentation

En se basant sur les solutions proposées lors de la présentation de l'approche dans le chapitre 3, nous avons développé une application qui est composée d'un certain nombre de classes et de méthodes Java, de 3114 tweets d'opinion, 3114 fichiers contenants des tweets et des informations sur des profils d'utilisateurs mâle et femelle. Nous détaillons par la suite chaque composante et sa fonctionnalité qui se base sur le diagramme de classe suivant :

15. https ://rapidminer.com/

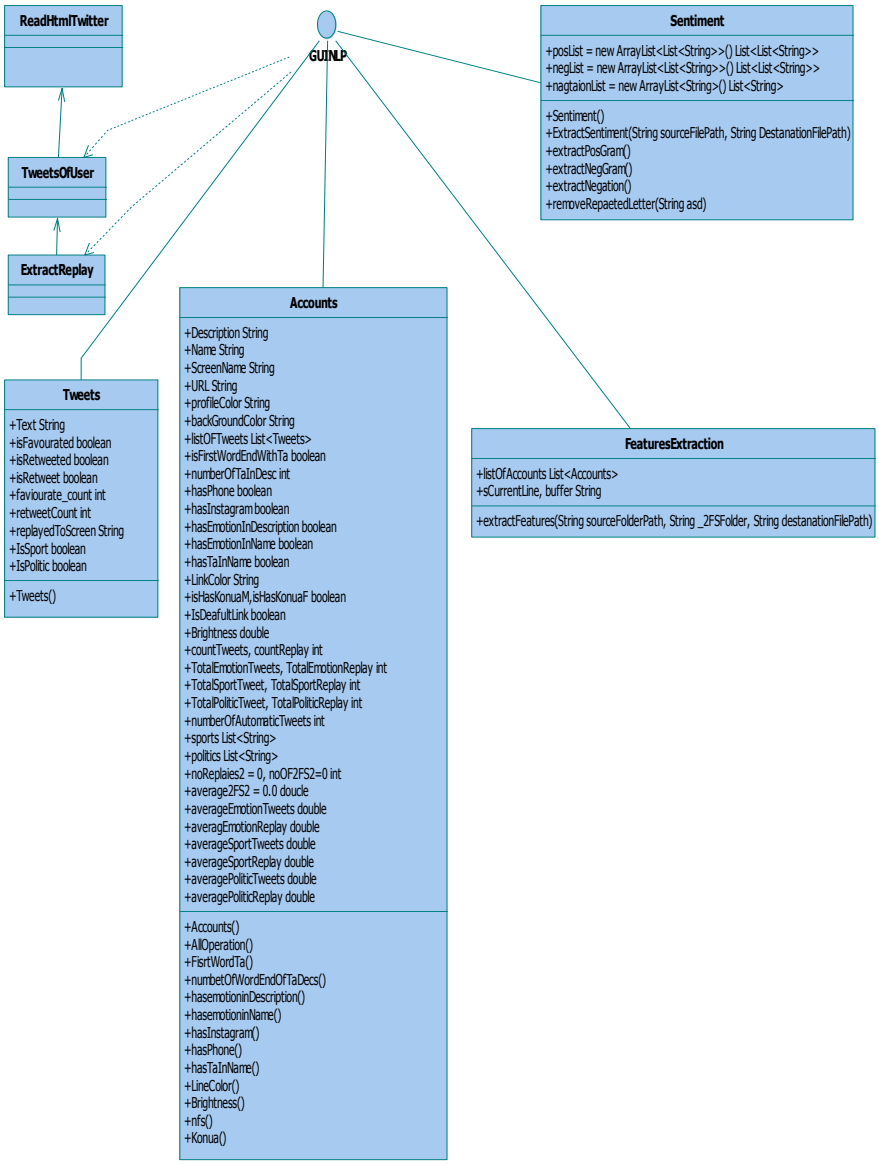


FIGURE 4.2 – diagramme de classe pour notre application

7.1.1 Interface de l'application

Afin de faciliter l'interaction avec notre application, nous avons mis en place une interface homme machine simple dont nous décrivons les principales fenêtres ci-dessous. Elle a pour but d'extraire les traits pour la détection du genre et pour l'analyse des sentiments afin de générer des fichiers contenants ces derniers. L'interface principale de notre application est présentée dans la figure suivante :



FIGURE 4.3 – Interface principale de notre application

Le bouton **« Browse »** du champ **« Choose The User Tweets Input Folder »** permet de choisir et charger le dossier contenant les informations et les tweets des utilisateurs.

Le bouton **« Browse »** du champ **« Choose The 2FS Folder »** permet de choisir et charger le dossier contenant le nombre de réponses d'une formule de destinataire féminine sur les tweets d'un utilisateur.

Le bouton **« Browse »** du champ **« Choose The User Tweets Output Folder »** permet de choisir le dossier qui va contenir les traits du genre des utilisateurs.

En cliquant sur le bouton **« Generate Gender Features File »**, le fichier des traits du genre sera créé, et comportera la liste des utilisateurs avec les traits correspondants afin de les classer en mâle et femelle.

Le bouton **« Browse »** du champ **« Choose The Sentiment Tweets Input Folder »** permettant de choisir et charger le dossier contenant les tweets d'opinions.

Après le choix du dossier qui va contenir les traits explicites d'opinions en cliquant sur le bouton **« Browse »** du champ **« Choose The Sentiment Tweets Output Folder »**. Nous devons cliquer sur le bouton **« Generate Sentiment Features File »** afin de créer et générer le fichier qui contient les traits explicites d'opinions et qui comportera une liste des tweets attaché à un ensemble de traits qui permet la classification des opinions selon leurs polarités.

7.1.2 Composantes de l'application

Dans le but d'exploiter aux mieux notre approche de détection du genre et d'analyse des sentiments, nous présentons ses composantes (Fichiers Excel, Fichiers Textes + Classes et Méthodes Java) :
Les fichiers Excel que contient l'application sont :

- ***sentimentTweets.xlsx***
  Ce fichier contient les Ids des détenteurs d'opinions et ses opinions.
- ***GenderFeatures.xlsx***
  Ce fichier contient les Ids des utilisateurs, les résultats des traits manipulés et les classes attribuées pour chaque genre.
- ***SentimentFeatures.xlsx***
  Ce fichier contient les tweets, les résultats des traits manipulés et la polarité attribuée pour chaque opinion.

Les fichiers textes que contient l'application sont :

- ***Male.txt***
  Ce fichier contient les Ids des comptes des hommes.
- ***Female.txt***
  Ce fichier contient les Ids des comptes des femmes.
  Ainsi qu'un dossier contenant des fichiers textes pour chaque utilisateurs et qui contiennent les informations et les tweets des utilisateurs.
- ***Pos.txt***
  Ce fichier regroupe 1185 mots d'opinion positifs.
- ***Neg.txt***
  Ce fichier regroupe 3937 mots d'opinion négatifs.
- ***Negation.txt***
  Ce fichier regroupe 31 mots de négations.
- ***GenderFeatures.txt***
  Ce fichier contient les traits des femmes ou des hommes.
- ***SentimentFeatures.txt***
  Ce fichier contient les traits des opinions positives ou négatives.

Les classes java que contient l'application sont :

- ***ReadHtmlTwitter.java***
  Elle permet de lire la page HTML Twitter et d'extraire les ID des détenteurs d'opinions avec ses tweets puis les écrire dans le fichier sentimentTweets.xlsx.
- ***TweetsOfUser.java***
  Elle permet d'extraire les informations sur les comptes Twitter extraient auparavant,ses tweets, ses retweets et les textes détiennent par l'utilisateur en utilisant l'API Twitter et les écrire dans des fichiers textes.

- ***ExtractReplay.java***
  Elle permet d'extraire les réponses des utilisateurs sur des tweets à partir des fichiers textes obtenus de la méthodes TweetsOfUser.java afin de les mieux exploiter.
- ***Tweets.java***
  Elle consiste à définir la structure d'un tweet.
- ***Accounts.java***
  Elle contient les méthodes que nous avons utilisées pour manipuler les traits d'utilisateur :
  - *hasemotioninName()* : Retourne *True* si le nom d'utilisateur contient une émotion ;
  - *hasemotioninDescription()* : Retourne *True* si la description d'utilisateur contient une émotion ;
  - *hasPhone()* : Retourne *True* si l'utilisateur a un numéro de téléphone ;
  - *hasInstagram()* : Retourne *True* si l'utilisateur a un compte Instagram ;
  - *LineColor()* : Retourne *True* si la couleur des liens est la couleur par défaut ;
  - *Brightness()* : Retourne le degré de luminosité des couleurs des liens.

  Elle contient les méthodes que nous avons utilisées pour manipuler les traits de comportement de messagerie :
  - *addTweets* : Retourne le nombre de tweets, le nombre total des tweets qui contiennent des émotions, le nombre de réponses, ainsi que le nombre total des réponses qui contiennent des émotions.

  Et contient les méthodes utilisées pour manipuler les traits linguistiques :
  - *nfs()* : Compte le nombre moyen des destinataires de la formule féminine par rapport au nombre total des réponses ;
  - *Konua()* : Retourne *True* si un compte femelle contient une appellation féminine et retourne *True* si un compte mâle contient une appellation masculin ;
  - *FisrtWordTa()* : Retourne *True* si le premier mot de la description se termine par "Ta" ;
  - *numberOfWordEndOfTaDecs()* : Retourne le nombre de mots qui se terminent par "Ta" ;
  - *hasTaInName()* : Retourne *True* si le nom d'utilisateur se termine par "Ta" ;
  - *isSportTweets(String text)* : Retourne *True* si le texte donné contient deux mots consécutifs qui parlent dans le domaine sportif ;
  - *isPoliticTweets(String text)* : Retourne *True* si le texte donné contient deux mots consécutifs qui parlent dans le domaine politique.

  Et enfin,
  - *AllOperation()* : La méthode qui fait appel a toute les méthodes précédentes et compte le nombre moyen des tweets et des réponses contiennent des émotions, le nombre moyen des tweets et des réponses contiennent des mots du domaine sportif et politique.

- ***featuresExtraction.java***
  C'est la classe principale (main) de la détection du genre, d'où nous créons des objets et appelons des différentes méthodes des autres classes.
  Nous décrivons par la suite la démarche suivie :
  - Lecture des fichiers contenant le nombre de destinataires d'une formule féminine ;
  - Lecture des dossiers contenant les fichiers textes des informations et des tweets ;
  - Initialisation des champs du fichier excel résultant ;
  - Écriture des résultats dans le fichier GenderFeatures.txt.
- ***sentiment.java***
  Cette classe permet l'extraction des traits explicites des opinions selon la démarche suivante :
  - Lecture du fichier contenant les tweets positifs ou négatifs ;
  - Initialisation du fichier qui va contenir les traits ;
  - Initialisation des champs du fichier résultant ;
  - Lecture des tweets ligne par ligne ;
  - Suppression des espaces multiples ;

  puis une séries de tests sera fait telles que :
  - *hasQuestionMark* : Retourne *True* si le tweet contient un point d'interrogation ;
  - *hasExclamationMark* : Retourne *True* si le tweet contient un point d'exclamation ;
  - *hasPic* : Retourne *True* si le tweet contient une image ;
  - *lentgthoftweets* : retourne la longueur du tweet ;
  - *extractPosGram()* : permet l'extraction des mots de sentiments positifs ;
  - *extractNegGram()* : permet l'extraction des mots de sentiments négatifs ;
  - *extractNegation()* : permet l'extraction des mots des négations ;
  - *removeRepeatedLetter(String asd)* : permet la suppression des lettres dupliquées.

  Enfin, nous avons besoin de faire :
  - Calcul du nombre de mots de sentiment positifs dans un tweet ;
  - Calcul du nombre de mots de sentiment négatifs dans un tweet ;
  - Calcul du nombre de négations dans un tweets.

7.2 Classification et Évaluation

7.2.1 Classification

Il existe 4 mesures de performance sont généralement utilisées pour évaluer la tâche de classification ; ces mesures sont calculées à base d'une matrice de confusion pour calculer la précision, le rappel, la pertinence et F-mesure :

Nous rappelons les formules [16] :

$$Précision_i = \frac{VP_i}{VP_i + FP_i} \tag{4.1}$$

$$Rappel_i = \frac{VP_i}{VP_i + FN_i} \tag{4.2}$$

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.3}$$

$$F - mesure = \frac{2 \times Précision \times Rappel}{Précision + Rappel} \tag{4.4}$$

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

Pour évaluer notre classificateur Naïf Bayes, nous avons utilisé l'outil RapidMiner cité dans la sous section 6.3.

En vue de démontrer l'efficacité d'utilisation des traits linguistiques dans la phase de détection du genre, nous avons effectué quatre test séparés, le premier concerne seulement les traits du profil, le deuxième concerne les traits de comportement de messagerie, le troisième concerne les traits linguistiques et enfin le quatrième test regroupe les trois tests précédents et qui classe les utilisateurs en mâle et femelle.
Après la préparation des données, nous avons appliqué le classificateur Naïf Bayes pour chaque test comme indiqué dans la figure suivante :

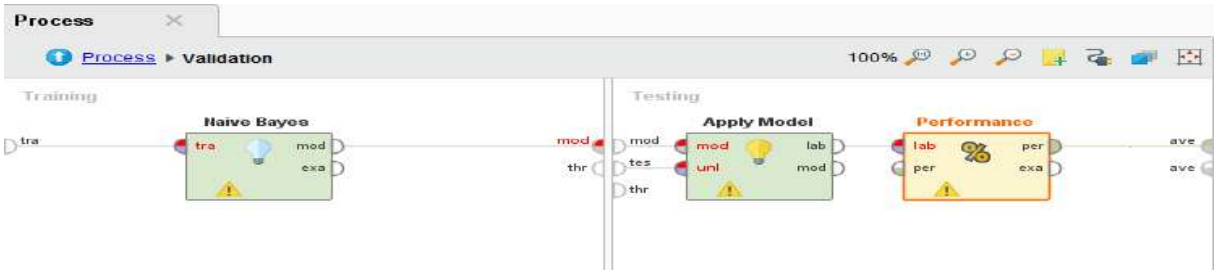


FIGURE 4.4 – Application du classificateur Naïf Bayes

16. **VP :**Vrai Positif, **FP :**Faux Positif, **VN :**Vrai Négatif, **FN :**Faux Négatif.

Et puis, nous avons validé notre modèle avec les données préparé afin de calculer les mesures de performance comme indiqué dans la figure suivante :

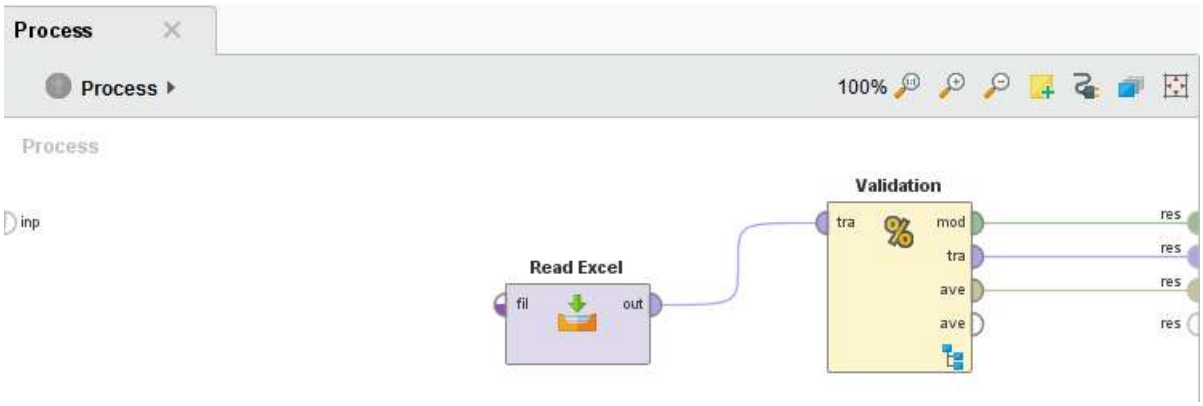


FIGURE 4.5 – Validation de test

- Classification selon les traits de profil

accuracy: 61.36% +/- 1.97% (mikro: 61.36%)

	true 1	true 0	class precision
pred. 1	1781	1025	63.47%
pred. 0	178	129	42.02%
class recall	90.91%	11.18%	

FIGURE 4.6 – Résultat de classification selon les traits de profil

- Classification selon les trais de comportement de messagerie

accuracy: 65.47% +/- 2.03% (mikro: 65.47%)

	true 1	true 0	class precision
pred. 1	1742	858	67.00%
pred. 0	217	296	57.70%
class recall	88.92%	25.65%	

FIGURE 4.7 – Résultat de classification selon les traits de comportement de messagerie

- Classification selon les traits linguistiques

accuracy: 69.90% +/- 1.86% (mikro: 69.90%)

	true 1	true 0	class precision
pred. 1	1770	748	70.29%
pred. 0	189	406	68.24%
class recall	90.35%	35.18%	

FIGURE 4.8 – Résultat de classification selon les traits linguistiques

- Classification générale

accuracy: 71.06% +/- 3.15% (mikro: 71.06%)

	true 1	true 0	class precision
pred. 1	1685	627	72.88%
pred. 0	274	527	65.79%
class recall	86.01%	45.67%	

FIGURE 4.9 – Résultat de classification générale

Les résultats du classificateur Naïf Bayes sont montrés dans le tableau suivant 7.1

	Précision	**Rappel**	**Pertinence**	**F-mesure**
Profil	**90.91%**	63.47%	61.36%	74.75%
Messagerie	88.92%	67.00%	65.47%	76.42%
Linguistique	90.35%	70.29%	69.90%	**79.07%**
Tout	86.01%	**72.88%**	**71.06%**	78.90%

TABLE 7.1 – Résultats des traits de détection du genre

Afin de classer les opinions en positifs et négatifs, nous avons effectué deux tests sur deux ensembles de textes, le premier regroupe tous les textes évaluatifs collectés et le deuxième regroupe que les textes subjectifs que nous avons déjà identifié :

- Classification selon tous les textes évaluatifs

accuracy: 77.96% +/- 1.64% (mikro: 77.96%)

	true 0	true 1	class precision
pred. 0	1794	386	82.29%
pred. 1	300	632	67.81%
class recall	85.67%	62.08%	

FIGURE 4.10 – Résultat de classification des tous les textes évaluatifs

- Classification selon les textes subjectifs

accuracy: 79.90% +/- 3.41% (mikro: 79.90%)

	true 0	true 1	class precision
pred. 0	1326	192	87.35%
pred. 1	308	662	68.25%
class recall	81.15%	77.52%	

FIGURE 4.11 – Résultat de classification des textes subjectifs

Les résultats du classificateur Naïf Bayes sont montrés dans le tableau suivant 7.2

	Précision	**Rappel**	**Pertinence**	**F-mesure**
Textes évaluatifs	**85.67%**	82.29%	77.96%	83.95%
Textes subjectifs	**81.15%**	87.35%	79.90%	84.14%

TABLE 7.2 – Résultats des traits de classification d'opinions

Afin de faire des statistiques sur les opinions mâle et femelle, nous avons effectué deux tests, le premier regroupe les opinions positives et négatives des hommes et le deuxième regroupe les opinions positives et négatives des femmes.

- Classification des opinions des hommes

accuracy: 81.22% +/- 2.77% (mikro: 81.23%)

	true 0	true 1	class precision
pred. 0	359	66	84.47%
pred. 1	124	463	78.88%
class recall	74.33%	87.52%	

FIGURE 4.12 – Résultat de classification des opinions des hommes

- Classification des opinions des femmes

accuracy: 87.51% +/- 3.27% (mikro: 87.50%)

	true 0	true 1	class precision
pred. 0	367	32	91.98%
pred. 1	32	81	71.68%
class recall	91.98%	71.68%	

FIGURE 4.13 – Résultat de classification des opinions des femmes

Les résultats du classificateur Naïf Bayes sont montrés dans le tableau suivant 7.2

	Précision	**Rappel**	**Pertinence**	**F-mesure**
Hommes	**74.33%**	84.47%	81.23%	76.04%
Femmes	**91.98%**	91.98%	87.50%	91.98%

TABLE 7.3 – Résultats de classification des opinions des hommes et des femmes

7.2.2 Évaluation

En vue de démontrer l'efficacité d'utilisation des traits linguistiques de la tâche d'identification du genre, nous avons effectué des tests séparés, le premier concerne seulement les traits du profil qui donne un résultat relativement bon, le deuxième concerne seulement les traits de comportement de messagerie qui donne un résultat bon par rapport au premier, le troisième test concerne les traits linguistiques qui améliore le résultat malgré la complexité morpho-syntaxique de la langue arabe qui rend difficile à repérer les expressions de féminisation dans les textes.

Après la combinaison des traits précédents, nous avons obtenu un très bon résultat ($f - mesure = 78.90\%$) par rapport à la complexité de la tâche d'identification du genre.

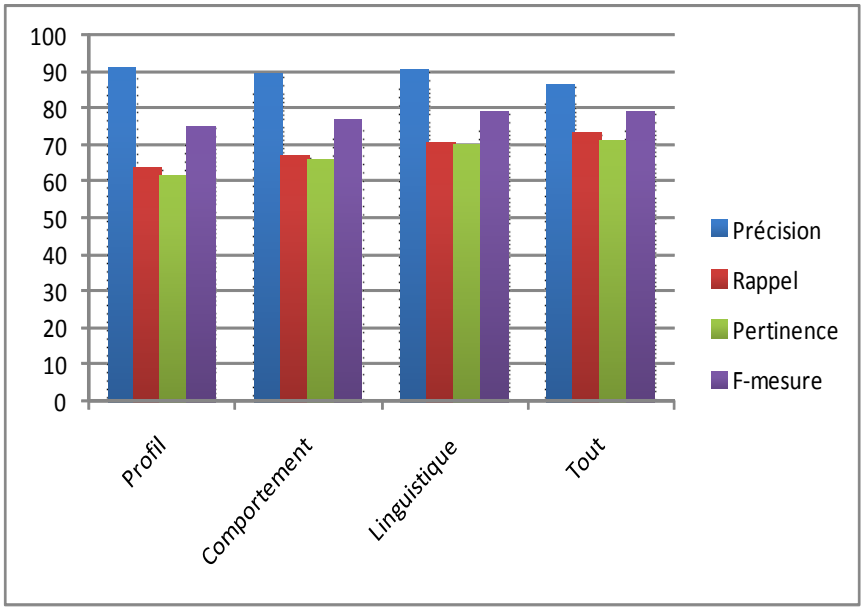


FIGURE 4.14 – Évaluation des différents traits de détection du genre

L'évaluation de deux ensembles de textes pour la classification d'opinions a montré un résultat très acceptable malgré le langage informel utilisé par les utilisateurs de Twitter de sorte que l'évaluation des textes subjectifs était meilleure par rapport à l'évaluation de tous les textes grâce aux expressions d'opinions explicites.

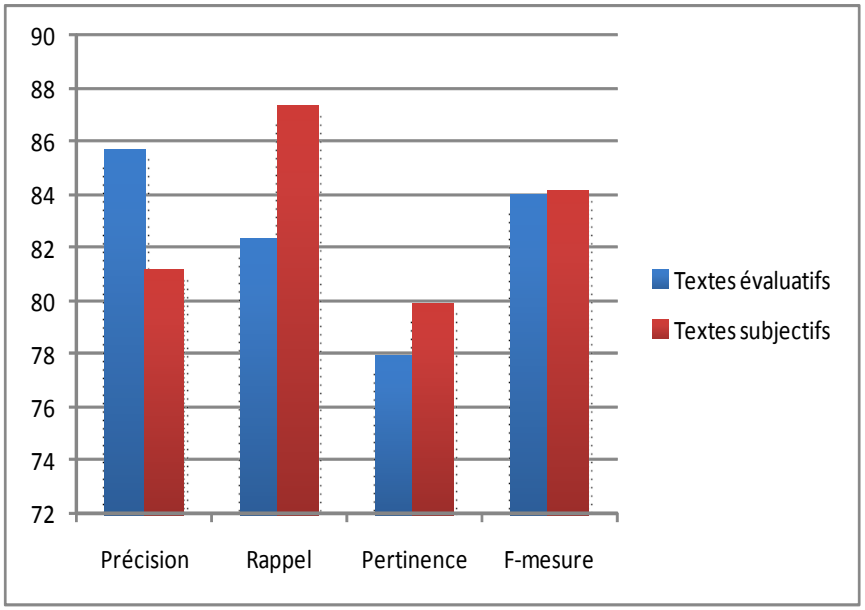


FIGURE 4.15 – Évaluation des différents traits de classification d'opinions

Afin de faire des statistiques sur les opinions et le genre de ses détenteurs, nous avons effectué une classification des opinions des hommes et des femmes, puis nous avons pris que les prédictions correctes pour atteindre le but des statistiques.

La figure suivante montre que les opinions se diffèrent entre les gens, surtout entre un

homme et une femme pour notre cas d'étude :

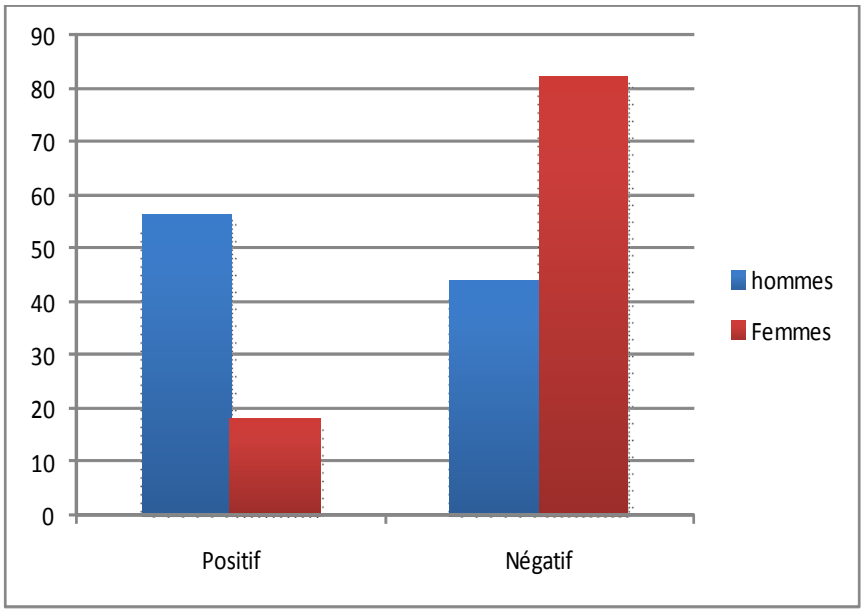


FIGURE 4.16 – statistiques d'opinions

8 Conclusion

Dans ce chapitre, nous avons présenté l'implémentation de notre approche d'identification du genre et de classification d'opinions, en utilisant le langage de programmation Java et d'autres outils. Nous avons pu effectuer encore une série de tests sur un ensemble d'utilisateurs et de textes évaluatifs.

Notre système d'identification du genre et de classification d'opinions a démontré sa performance lors de l'exploration linguistique malgré la complexité morphologique et syntaxique de certaines expressions arabes qui a diminué considérablement la performance des processus d'identification des textes d'opinion et des traits linguistiques.

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 lexique de sentiments 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 sont regroupés.

L'évaluation a également montré que le système atteint un très bon résultat qui peut être plus performant lorsque le lexique des sentiments est plus riche. Un tel système peut donc combiner plusieurs approches pour améliorer sa performance.

Conclusion et perspectives

Dans ce mémoire, nous avons présenté une approche de classification d'opinions par genre dans les textes arabes d'un domaine étudié, dont le but est de classer les détenteurs des opinions en mâle et femelle, de classer les opinions en positives et négatives et d'étudier la contribution des traits linguistiques et leurs influences sur la performance des processus d'identification de genre et de classification d'opinions.

Plusieurs ressources et techniques ont été employées pour extraire les traits du profil utilisateur, les traits d'opinion et les expressions d'opinion exprimées sur ces dernières : des profils utilisateurs, un corpus textuels et un lexique de sentiments spécifique au domaine étudié.

Pour extraire les traits du profil utilisateur, nous avons utilisé trois catégories de traits : traits du profil, traits de comportement de messagerie et les traits linguistiques en vue de construire la liste des couples utilisateur-traits.

Pour extraire les traits explicites d'opinions, nous avons identifié premièrement les textes subjectifs afin de minimiser le nombre de textes objectifs, puis nous avons utilisé les traits linguistiques en vue de construire la liste des couples opinion-traits.

Après avoir testée et évaluée notre approche GenderSentiment sur une partie de notre corpus textuel, nous avons constaté que l'exploration linguistique a amélioré considérablement la performance de notre système. L'exploration linguistique mène à bien classer les utilisateurs selon leur genre et à extraire des caractéristiques explicites sur lesquelles des opinions a été exprimées afin de les classer selon leurs polarité.

Lors de l'évaluation, nous avons constaté que trois facteurs principaux ont été derrière la diminution des performances de notre système :

- Les faux profils visant les problèmes et les insultes et qui ne représentent pas le visage public des tweets des hommes et des femmes ;

- Le nombre limités des tweets récupérés ;

- Les tweets publicitaires ;

- Le lexique de sentiments ne couvre pas toutes les expressions parce qu'il est construit comme un lexique général et n'est pas spécifique à notre domaine étudié (nous avons juste fait un simple enrichissement manuel) ;

- Certaines expressions arabes morphologiquement et syntaxiquement complexes ont diminué la performance d'extraction des expressions d'opinion et leurs traits.

Afin d'améliorer la performance de notre approche, nous proposons dans nos futurs travaux de recherches ce qui suit :

- Classification d'utilisateurs en trois classes : mâle, femelle et organisation en utilisant le classificateur SVM ;

- En vue de diminuer l'ambiguïté subjectives entre expressions d'opinion, nous proposons de séparer les lexiques de sentiment par domaine, parce qu'un mot d'opinion peut être positif dans un domaine et négatif dans autre et vice-versa ;

- L'enrichissement automatique des lexiques de sentiments peut aider à construire un lexique pour couvrir un maximum d'expressions, ce qui mène à améliorer considérablement la performance du système d'identification des textes subjectifs en particulier les expressions d'opinions ;

- Une analyse plus profonde des expressions arabes complexes peut faciliter à la fois l'identification du genre, la localisation des expressions d'opinion et leurs traits dans le texte.

- Classification d'opinions en trois classes : positif, négatif et neutre en utilisant le classificateur SVM.

- Utilisation de la classification d'opinion comme un trait qui améliore le système du détection du genre selon le domaine d'étude.

Bibliographie

[1] FJAP Mattosinho. Mining product opinions and reviews on the web. *Technische Universitat Dresden*, 2010.

[2] Yasuhisa Yoshida, Tsutomu Hirao, Tomoharu Iwata, Masaaki Nagata, and Yuji Matsumoto. Transfer learning for multiple-domain sentiment analysis—identifying domain dependent/independent word polarity. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[3] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2) :1–135, 2008.

[4] Damien Poirier. *Des Textes communautaires à la recommandation*. PhD thesis, Université d'Orléans, 2011.

[5] Jihene Jmal. Rests : Système de résumé automatique des textes d'opinions basé sur twitter et sentiwordnet. *JEP-TALN-RECITAL 2012*, page 233.

[6] Bing Liu. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2 :627–666, 2010.

[7] Gilles Hubert. *Recherche d'information et contexte*. PhD thesis, Université Paul Sabatier-Toulouse III, 2010.

[8] Romain Tursi. *Les défis de la gestion des connaissances en contexte interculturel*. PhD thesis, Université Laval, 2006.

[9] Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery : Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003.

[10] Nitin Jindal and Bing Liu. Mining comparative sentences and relations. In *AAAI*, volume 22, pages 1331–1336, 2006.

[11] Sanjiv Das and Mike Chen. Yahoo ! for amazon : Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific finance association annual conference (APFA)*, volume 35, page 43. Bangkok, Thailand, 2001.

[12] Peter D Turney. Thumbs up or thumbs down ? : semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[13] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up ? : sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[14] Malik Muhammad Saad Missen and Mohand Boughanem. Sentence-level opinion-topic association for opinion detection in blogs. In *Advanced Information Networking and Applications Workshops, 2009. WAINA'09. International Conference on*, pages 733–737. IEEE, 2009.

[15] K Sparck Jones, Steve Walker, and Stephen E. Robertson. A probabilistic model of information retrieval : development and comparative experiments : Part 2. *Information Processing & Management*, 36(6) :809–840, 2000.

[16] Sébastien Gillot. Fouille d'opinions. *Rapport de stage*, 2010.

[17] Sigrid Maurel, Paolo Curtoni, and Luca Dini. L'analyse des sentiments dans les forums. *Atelier Fouille des Données d'Opinion*, 2008.

[18] Ekaterina O.B Reffin J., Taras E.Z. "comparable domain dependency in sentiment analysis". *Journal of Siberian Federal University*, Vol. 3, :pp.764–775, (2010).

[19] Grzegorz Dziczkowski. *Analyse des sentiments : systeme autonome d'exploration des opinions exprimées dans les critiques cinématographiques*. PhD thesis, Thèse de doctorat, Ecole Nationale Supérieurres, 2008.

[20] Jean-Michel Adam. *La linguistique textuelle : introduction à l'analyse textuelle des discours*. Armand Colin, 2011.

[21] Damien Poirier, Françoise Fessant, Cécile Bothorel, Emilie Guimier De Neef, and Marc Boullé. Approches statistique et linguistique pour la classification de textes d'opinion portant sur les films. *Revue des Nouvelles Technologies de l'Information*, pages Pages–147, 2009.

[22] Faiza Belbachir. Expérimentation de fonctions pour la détection d'opinions dans les blogs.

[23] Lazhar FAREK. *Identification d'opinions dans les journaux arabes*. PhD thesis, Université Badji Mokhtar de Annaba, 2009.

[24] Ying Li, Zijian Zheng, and Honghua Kathy Dai. Kdd cup-2005 report : Facing a great challenge. *ACM SIGKDD Explorations Newsletter*, 7(2) :91–99, 2005.

[25] Bing Liu et al. Opinion mining and summarization-sentiment analysis. *Tutorial in the Proceedings of WWW*, 8, 2008.

[26] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.

[27] Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *EMNLP-CoNLL*, volume 7, pages 1065–1074. Citeseer, 2007.

[28] Benjamin Snyder and Regina Barzilay. Multiple aspect ranking using the good grief algorithm. In *HLT-NAACL*, pages 300–307, 2007.

[29] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3) :165–210, 2005.

[30] Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. Automatic extraction of opinion propositions and their holders. In *2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text*, volume 2224, 2004.

[31] Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics, 2005.

[32] Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. Just how mad are you ? finding strong and weak opinion clauses. In *aaai*, volume 4, pages 761–769, 2004.

[33] W Gerrod Parrott. *Emotions in social psychology : Essential readings*. Psychology Press, 2001.

[34] Janyce Wiebe and Theresa Wilson. Learning to disambiguate potentially subjective expressions. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–7. Association for Computational Linguistics, 2002.

[35] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Computational linguistics*, 30(3) :277–308, 2004.

[36] Ellen Riloff, Siddharth Patwardhan, and Janyce Wiebe. Feature subsumption for opinion analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 440–448. Association for Computational Linguistics, 2006.

[37] Janyce Wiebe. Learning subjective adjectives from corpora. In *AAAI/IAAI*, pages 735–740, 2000.

[38] Jaime Guillermo Carbonell. Subjective understanding : Computer models of belief systems. Technical report, DTIC Document, 1979.

[39] Yorick Wilks and Janusz Bien. Beliefs, points of view, and multiple environments*. *Cognitive Science*, 7(2) :95–119, 1983.

[40] Iadh Ounis, Craig Macdonald, and Ian Soboroff. Overview of the trec-2008 blog track. Technical report, DTIC Document, 2008.

[41] Cyril Grouin, Béatrice Arnulphy, Jean-Baptiste Berthelin, Sarra El Ayari, Anne García-Fernandez, Arnaud Grappy, Martine Hurault-Plantet, Patrick Paroubek, Isabelle Robba, and Pierre Zweigenbaum. Présentation de l'édition 2009 du défi fouille de textes (deft'09). *Actes de l'atelier de clôture de la cinquieme édition du DEfi Fouille de Textes*, pages 35–50, 2009.

[42] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1) :1–167, 2012.

[43] Peter D Turney and Michael L Littman. Measuring praise and criticism : Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4) :315–346, 2003.

[44] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2) :267–307, 2011.

[45] Xiaowen Ding, Bing Liu, and Philip S Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 231–240. ACM, 2008.

[46] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[47] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer : analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM, 2005.

[48] Alina Andreevskaia and Sabine Bergler. Mining wordnet for a fuzzy sentiment : Sentiment tag extraction from wordnet glosses. In *EACL*, volume 6, pages 209–216, 2006.

[49] Olfa Nasraoui. Web data mining : Exploring hyperlinks, contents, and usage data. *ACM SIGKDD Explorations Newsletter*, 10(2) :23–25, 2008.

[50] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions : Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics, 2003.

[51] Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of english words. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 183–190. Association for Computational Linguistics, 1993.

[52] Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics, 1998.

[53] Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics, 1997.

[54] Hiroshi Kanayama and Tetsuya Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355–363. Association for Computational Linguistics, 2006.

[55] Philippe Martin. Using the wordnet concept catalog and a relation hierarchy for knowledge acquisition. *Proc. Peirce*, 95, 1995.

[56] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet : An on-line lexical database*. *International journal of lexicography*, 3(4) :235–244, 1990.

[57] Saif Mohammad, Cody Dunne, and Bonnie Dorr. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 2-Volume 2*, pages 599–608. Association for Computational Linguistics, 2009.

[58] Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato Recupero, and Venkatramana S Subrahmanian. Sentiment analysis : Adjectives and adverbs are better than adjectives alone. In *ICWSM*. Citeseer, 2007.

[59] Guang Qiu, Xiaofei He, Feng Zhang, Yuan Shi, Jiajun Bu, and Chun Chen. Dasa : dissatisfaction-oriented advertising based on sentiment analysis. *Expert Systems with Applications*, 37(9) :6182–6191, 2010.

[60] Gerard Salton and Michael E Lesk. The smart automatic document retrieval systems—an illustration. *Communications of the ACM*, 8(6) :391–398, 1965.

[61] Kamal Nigam and Matthew Hurst. Towards a robust metric of polarity. In *Computing Attitude and Affect in Text : Theory and Applications*, pages 265–279. Springer, 2006.

[62] Bo Pang and Lillian Lee. A sentimental education : Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting*

on Association for Computational Linguistics, page 271. Association for Computational Linguistics, 2004.

[63] Michel Plantié, Mathieu Roche, Gérard Dray, and Pascal Poncelet. Is a voting approach accurate for opinion mining ? In *Data Warehousing and Knowledge Discovery*, pages 413–422. Springer, 2008.

[64] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[65] James H Martin and Daniel Jurafsky. Speech and language processing. *International Edition*, 2000.

[66] Adwait Ratnaparkhi et al. A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142. Philadelphia, USA, 1996.

[67] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

[68] Vasileios Hatzivassiloglou, Kathleen R McKeown, Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. Predicting the semantic orientation of adjectives. *ACM Transactions on Information Systems (TOIS)*, 21(4) :315–346, 2009.

[69] Janyce Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Computational Linguistics and Intelligent Text Processing*, pages 486–497. Springer, 2005.

[70] Ellen Riloff, Janyce Wiebe, and Theresa Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 25–32. Association for Computational Linguistics, 2003.

[71] Janyce M Wiebe, Rebecca F Bruce, and Thomas P O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 246–253. Association for Computational Linguistics, 1999.

[72] Michel Généreux and Marina Santini. Défi : Classification de textes français subjectifs. *Actes du troisième DÉfi Fouille de Textes*, page 95, 2007.

[73] Charu C Aggarwal. Towards meaningful high-dimensional nearest neighbor search by human-computer interaction. In *Data Engineering, 2002. Proceedings. 18th International Conference on*, pages 593–604. IEEE, 2002.

[74]

[75] Adam Bermingham and Alan F Smeaton. Classifying sentiment in microblogs : is brevity an advantage ? In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1833–1836. ACM, 2010.

[76] Thorsten Joachims. Making large scale svm learning practical. Technical report, Universität Dortmund, 1999.

[77] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3) :273–297, 1995.

[78] Theodore W Anderson and RR Bahadur. Classification into two multivariate normal distributions with different covariance matrices. *The annals of mathematical statistics*, pages 420–431, 1962.

[79] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4) :620, 1957.

[80] Christopher Manning. Maxent models and discriminative estimation. *CS 224N lecture notes, Spring*, 2005.

[81] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1 :12, 2009.

[82] Thomas M Mitchell et al. Machine learning, 1997.

[83] Norbert Fuhr, Stephan Hartmann, Gerhard Lustig, Michael Schwantner, Kostas Tzeras, and Gerhard Knorz. *AIR, X : a rule based multistage indexing system for large subject fields*. Citeseer, 1991.

[84] William W Cohen and Yoram Singer. Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems (TOIS)*, 17(2) :141–173, 1999.

[85] Yong H Li and Anil K Jain. Classification of text documents. *The Computer Journal*, 41(8) :537–546, 1998.

[86] J Ross Quinlan. *C4. 5 : programs for machine learning*. Elsevier, 2014.

[87] Chid Apte, Fred Damerau, Sholom Weiss, et al. *Text mining with decision rules and decision trees*. Citeseer, 1998.

[88] Robert E Schapire and Yoram Singer. Boostexter : A boosting-based system for text categorization. *Machine learning*, 39(2) :135–168, 2000.

[89] William W Cohen. Learning trees an ules with set-val ed features. 1996.

[90] Grzegorz Dziczkowski and Katarzyna Wegrzyn-Wolska. An autonomous system designed for automatic detection and rating of film reviews. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 847–850. IEEE Computer Society, 2008.

[91] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics : posters*, pages 241–249. Association for Computational Linguistics, 2010.

[92] Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger. Pulse : Mining customer opinions from free text. In *international symposium on intelligent data analysis*, pages 121–132. Springer, 2005.

[93] Lingwei Zeng and Fang Li. A classification-based approach for implicit feature identification. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 190–202. Springer, 2013.

[94] Wei Wang, Hua Xu, and Wei Wan. Implicit feature identification via hybrid association rule mining. *Expert Systems with Applications*, 40(9) :3518–3531, 2013.

[95] Qi Su, Kun Xiang, Houfeng Wang, Bin Sun, and Shiwen Yu. Using pointwise mutual information to identify implicit features in customer reviews. In *Computer Processing of Oriental Languages. Beyond the Orient : The Research Challenges Ahead*, pages 22–30. Springer, 2006.

[96] Christopher Scaffidi, Kevin Bierhoff, Eric Chang, Mikhael Felker, Herman Ng, and Chun Jin. Red opal : product-feature scoring from reviews. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 182–191. ACM, 2007.

[97] Roger J Kreuz and Sam Glucksberg. How to be sarcastic : The echoic reminder theory of verbal irony. *Journal of Experimental Psychology : General*, 118(4) :374, 1989.

[98] Kevin Duh, Akinori Fujino, and Masaaki Nagata. Is machine translation ripe for cross-lingual sentiment classification ? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers-Volume 2*, pages 429–433. Association for Computational Linguistics, 2011.

[99] Min Xiao and Yuhong Guo. A novel two-step method for cross language representation learning. In *Advances in Neural Information Processing Systems*, pages 1259–1267, 2013.

[100] Alexander Hogenboom, Malissa Bal, Flavius Frasincar, and Daniella Bal. Towards cross-language sentiment analysis through universal star ratings. In *7th International Conference on Knowledge Management in Organizations : Service and Cloud Computing*, pages 69–79. Springer, 2013.

[101] Huiwei Zhou, Long Chen, Fulin Shi, and Degen Huang. Learning bilingual sentiment word embeddings for cross-language sentiment classification.

[102]

[103] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. 2007.

[104] Rui Xia and Chengqing Zong. A pos-based ensemble model for cross-domain sentiment classification. In *IJCNLP*, pages 614–622. Citeseer, 2011.

[105] Yulan He, Chenghua Lin, and Harith Alani. Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, pages 123–131. Association for Computational Linguistics, 2011.

[106] Sheng Gao and Haizhou Li. A cross-domain adaptation method for sentiment classification using probabilistic latent analysis. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1047–1052. ACM, 2011.

[107] Abdel Rahman Hedar and M Doss. Mining social networks arabic slang comments. In *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2013.

[108] Muhammad Abdul-Mageed and Mona T Diab. Awatif : A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC*, pages 3907–3914, 2012.

[109] Mohamed Elarnaoty, Samir AbdelRahman, and Aly Fahmy. A machine learning approach for opinion holder extraction in arabic language. *arXiv preprint arXiv :1206.1011*, 2012.

[110] Alaa El-Halees. Arabic opinion mining using combined classification approach. 2011.

[111] Hossam S Ibrahim, Sherif M Abdou, and Mervat Gheith. Sentiment analysis for modern standard arabic and colloquial. *arXiv preprint arXiv :1505.03105*, 2015.

[112] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010.

[113] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

[114] William Deitrick, Zachary Miller, Benjamin Valyou, Brian Dickinson, Timothy Munson, and Wei Hu. Gender identification on twitter using the modified balanced winnow. 2012.

[115] Wendy Liu and Derek Ruths. What's in a name ? using first names as features for gender inference in twitter. In *AAAI spring symposium : Analyzing microtext*, volume 13, page 01, 2013.

[116] Wendy Liu, Faiyaz Al Zamal, and Derek Ruths. Using social media to infer gender composition of commuter populations. In *Proceedings of the when the city meets the citizen workshop, the international conference on weblogs and social media*, 2012.

[117] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. Homophily and latent attribute inference : Inferring latent attributes of twitter users from neighbors. *ICWSM*, 270, 2012.

[118] Clayton Fink, Jonathon Kopecky, and Maksym Morawski. Inferring gender from the content of tweets : A region specific example. In *ICWSM*, 2012.

[119] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender in twitter : Styles, stances, and social networks. *CoRR abs/1210.4567*, 2012.

[120] Anaïs Cadilhac, Farah Benamara, and Nathalie Aussenac-Gilles. Ontolexical resources for feature based opinion mining : a case-study. In *23rd International conference on computational linguistics*, page 77. Citeseer, 2010.

[121] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[122] Bruno Ohana and Brendan Tierney. Sentiment classification of reviews using sentiwordnet. In *9th. IT & T Conference*, page 13, 2009.

[123] Ali Harb, Gérard Dray, Michel Plantié, Pascal Poncelet, Mathieu Roche, and François Trousset. Détection d'opinion : Apprenons les bons adjectifs ! In *INFORSID'08 : INFormatique des Organisations et Systèmes d'Information et de Décision-Atelier FODOP'08*, pages 59–66, 2008.

[124] Matthieu Vernier, Laura Monceaux, and Béatrice Daille. Identifier la cible d'un passage d'opinion dans un corpus multithématique. In *TALN 2011*, 2011.

[125] Olga Feiguina. *Résumé automatique des commentaires de Consommateurs*. PhD thesis, Université de Montréal, 2006.

[126] Ana-Maria Popescu and Orena Etzioni. Extracting product features and opinions from reviews. In *Natural language processing and text mining*, pages 9–28. Springer, 2007.

[127] Giuseppe Carenini, Raymond T Ng, and Ed Zwart. Extracting knowledge from evaluative text. In *Proceedings of the 3rd international conference on Knowledge capture*, pages 11–18. ACM, 2005.

[128] Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A Reis, and Jeff Reynar. Building a sentiment summarizer for local service reviews. In *WWW workshop on NLP in the information explosion era*, volume 14, pages 339–348, 2008.

[129] Lili Zhao and Chunping Li. Ontology based opinion mining for movie reviews. In *International Conference on Knowledge Science, Engineering and Management*, pages 204–214. Springer, 2009.

[130] Xiwen Cheng and Feiyu Xu. Fine-grained opinion topic and polarity identification. In *LREC*, 2008.

[131] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. *ICWSM*, 11(1) :281–288, 2011.

[132] Vasileios Hatzivassiloglou and Janyce M Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 299–305. Association for Computational Linguistics, 2000.

[133] Jalal S Alowibdi, Ugo A Buy, and Philip Yu. Language independent gender classification on twitter. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 739–743. IEEE, 2013.

[134] Anh-Phuc Trinh. Classification de texte et estimation probabiliste par machine à vecteurs de support. *Actes du troisième DÉfi Fouille de Textes*, page 77, 2007.

[135] Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. Madamira : A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *LREC*, volume 14, pages 1094–1101, 2014.

ملخص

مع النمو السريع لتقنيات الشبكات الاجتماعية على شبكة الإنترنت في السنوات الأخيرة، تحليل المشاعر و التعرف على جنس المستخدم أثبت فائدة على نحو متزايد. التنبؤ الدقيق للسمات الديموغرافية لوسائل الإعلام الاجتماعية و غيرها من المحتوى غير الرسمي على الإنترنت له قيمة كبيرة للتسويق، التخصيص، علم الاجتماع و التحقيقات القانونية. يقدم تحليل الأبوة معلمات عن أصحاب الرأي التي تشمل في كثير من الأحيان جنس صاحب الرأي. الرجال و النساء يعبرون عن آرائهم بطريقة مختلفة بشكل ملحوظ.

تتناول هذه المذكرة مهمة تصنيف المستخدم العربي في وسائل الإعلام الاجتماعية مع تطبيق لتويتر. نستخرج تلقائيا قيم سمات المستخدم مثل التوجه السياسي أو الرياضي مع الاستفادة من المعلومات التي يمكن ملاحظتها مثل : المعلومات الشخصية للمستخدم، سلوكه و المحتوى اللغوي لمستخدم تويتر. نحن نستخدم منهج التلقين الذاتي بناءا على مجموعة شاملة من الميزات المستمدة من المعلومات التي تبين أن السمات اللغوية الغنية واعدة جدا لهذا النوع من متطلبات التصنيف.

هذه المذكرة تدرس أيضا مشكلة تصنيف الآراء في النصوص العربية الموحدة و الجدلية. تقدم الأطروحة أيضا تصنيفا لتغريدات هاشتاج إلى إيجابية و سلبية على أساس التلقين الذاتي باستخدام مجموعة مختلفة من الميزات الرقمية (الموحدة) و اللغوية من خلال تطبيق مصنف ثنائي.

تصف المذكرة أيضا مجموعة من الإحصائيات التي تبين الفرق بين آراء الذكور و الإناث.

الكلمات الرئيسية: تحليل الآراء، التعرف على الجنس، تصنيف الآراء، التلقين الذاتي، اللغة العربية.

Abstract

With the rapid growth of web-based social networking technologies in recent years, gender identification and sentiment analysis have proven increasingly useful. Accurate prediction of demographic attributes from social media and other informal online content is valuable for marketing, personalization, sociology, and legal investigation. Authorship analysis provides information about an opinion holder, often including the opinion's gender. Men and women express opinions in distinctly different ways.

This dissertation addresses the task of arabic user classification in social media, with an application to Twitter. We automatically infer the values of user attributes such as political orientation or athletic by leveraging observable information such as the user profile, behavior, and the linguistic content of the user's Twitter feed. We employ a machine learning approach which relies on a comprehensive set of features derived from such user information witch show great promise for gender classification needs.

This dissertation studies also the problem of opinion classification in Arabic and colloq-uial Twitter texts. Our approach introduce a machine learning approach for tweets classification as either positive, negative. We used a set of various features which are mainly linguistic and numeric (standard) with applying naive Bayesian classifier.

The present dissertation also describes a set of statistics that shows the difference between male and female opinions.

Key words: Sentiment analysis, gender identification, opinion classification, machine learning, arabic language

d'opinion et les textes subjectifs, où chaque mot de ce lexique peut être un adjectif, un verbe ou un adverbe éventuellement combiné avec certains modificateurs tels que la négation et la confirmation, et pour chaque mot de ce lexique, une polarité positive ou négative est associée.

• Contexte du travail

Le premier axe de notre recherche est la construction d'un corpus d'apprentissage arabe sur un domaine donné et qui regroupe les opinions et ses détenteurs.

Le deuxième axe est la conception d'une approche permettant la détection du genre des détenteurs d'opinion arabophones basée sur l'apprentissage automatique et qui néglige un ensemble de trait de profil utilisateur (numéro de téléphone, instagram, URL...) un ensemble de traits de comportement de messagerie et enfin un ensemble qui est le plus riche composé des traits linguistiques afin de classer les détenteurs en mâle et femelle.

Le troisième axe est l'approche qui permet l'identification des textes subjectifs à partir de l'ensemble des textes évaluatifs en se basant sur un lexique de sentiments.

Le lexique comporte des milliers de mots de sentiments, des émotions et des états psychologiques, tels que des noms comme حب (amour) ou خوف (peur), des verbes comme يؤيد (supporter), يعارض (opposer), et des adjectifs comme جميل (beau) ou غبي (édiot). Dans ce contexte nous proposons une classification dans laquelle ces mots sont en plusieurs classes, chaque classe est nommée par le sentiment ou l'état psychologique décrit, comme la classe négative, qui contient les mots relatifs à un sentiment négatif (غباء- peur,خشية - crainte, ذعر - frayeur, يرعب - effrayer, مرعب - effrayant, etc.). Nous distinguons deux classes de mots :

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

Puis la classification des opinions en positives et négatives.

Le quatrième axe consiste à faire des statistiques qui évaluent les résultats des opinions selon le genre des détenteurs.

Dans ces dictionnaires, une polarité est associée a priori à chacun des mots. Quel que soit le contexte dans lequel il sera inséré, le mot devrait ainsi avoir toujours la même polarité. Nous donnons ensuite au document un score d'opinion en fonction de la présence de mots issus de ces dictionnaires dans le texte.

3.4.2.1 Initialisation du dictionnaire

Dans notre approche, nous proposons de construire un lexique de sentiments pour notre domaine étudié. Nous croyons que la séparation des lexiques par domaine peut diminuer considérablement l'ambiguïté sémantique des mots d'opinion.

Par exemple le mot (صغير) (petit) peut être positif dans un domaine et négatif dans un autre.

Exemple :

(الهاتف صغير) (le mobile est petit), (صغير) (petit) est positif dans le domaine d'électronique.
(الغرفة صغيرة) (la chambre est petite), (صغير) (petit) est négatif dans le domaine d'urbanisme.

Une idée inspirée des travaux de Harb et al. (2008) [121] consiste à construire un lexique avec les mots de sentiments les plus connus dans le domaine étudié. Pour cela, nous considérons deux ensembles P et N de mots germes classiquement utilisés dans la littérature dont les orientations sémantiques sont respectivement positives et négatives.

Exemple :

P = {جيد، جميل، ممتاز ،ايجابي ،صحيح، محبوب} {aimable, correct, positive, excellent, beau, bien};

N = {سيء، قبيح ،رديء ،خاطئ، مكروه} {désagréable, faux, mauvais, laid, mauvais} ;

Le dictionnaire 3.1 aura alors pour chaque entrée : le mot de sentiment et sa polarité positive ou négative.

Mots de sentiments	Polarité
جيد	+
جميل	+
سيء	-
غبي	-
…etc.	

TABLE 3.1 – Structure simplifiée d'un dictionnaire des sentiments

Nous avons regroupé notre dictionnaire en deux fichiers l'un contient les mots de sentiments positifs et l'autre contient les mots de sentiments négatifs.

3.5.1.3 Extraction des traits linguistiques

L'utilisation lexicale des utilisateurs et le style linguistique peuvent fournir des conseils utiles pour classer les utilisateurs tels que :

• **Formule féminine :** montre la féminine de la personne.

Exemple :

(Ingénieur) (مهندس) : l'absence de la lettre " ة " ou " ه " indique que cette personne est mâle.

(Étudiante) (طالبة) : la présence de la lettre " ة " nous montre que la personne est femelle.

(Professeur) (أستاذه) : la présence de la lettre " ه " montre la féminine de la personne.

• **Destinataire de la formule de féminisation :** nous avons utilisé la formule de féminisation afin de savoir le genre d'un destinataire en utilisant les réponses qu'il reçoit.

Exemple :

للمرة القادمة انشري أذكى رئيس و أغبى رئيس في العالم ,يا أختي خديجة البراميل الفارغة تحدث أصواتا كبيرة
La lettre " ي " à la fin du mot " أختي_" et "انشري" preuve que le destinataire est femelle.

• **Mots prototypes :** ces mots nous aident à différencier entre les intérêts d'un homme et d'une femme sachant que les femmes s'intéressent beaucoup à la mode, au shopping ...etc., et les hommes sont généralement intéressé au sport et à la politique... etc.

Exemple :

رياضة(sport), فريق(équipe), مدرب(entraineur)…etc.
حرب(guerre), انقلاب عسكري (coup d'état)…etc.
موضة (mode), التسوق (shopping)…etc.

Nous avons aussi trouvé que les arabophones ont des appellations qui peuvent faire une distinction entre les genres.

Exemple :
(خالد بن الوليد) :indique que la personne est mâle car " بن " (fils) est indice du masculin.
(أبو حامد الغزالي) :indique que la personne est mâle car " أبو " (père) est indice du masculin.
(أم المؤمنين) :indique que la personne est femelle car " أم " (mère) est indice du féminine.
(زينب بنت جحش) :indique que la personne est femelle car " بنت " (fille) est indice du féminine.

3.5.2 Analyse des sentiments

Un texte évaluatif est composé de un ou plusieurs segments, chaque segment peut être subjectif ou objectif, chaque segment subjectif contient au moins une expression d'opinion (EO), notons qu'une EO est une expression explicite exprimée sur un objet, elle est composée d'un nom, un adjectif, ou un verbe avec certains modificateurs comme les négations et les adverbes.
Exemple :
(Très belle vue) منظر جميل جدا
Ce simple exemple contient une expression d'opinion composée d'un adjectif جميل (belle) et un modificateur جدا (très).
Notre approche qui s'intéresse à la classification d'opinion, a besoin de deux composants de base :
— Un lexique de sentiments : comme nous l'avons vu précédemment, il s'agit d'une ressource lexicale L des expressions d'opinion propres à un domaine étudié ;
— Un ensemble de textes évaluatifs T collectés d'Internet.
Nous détaillons dans ce qui suit les différents segments de ce module.

3.5.2.1 Identification des textes subjectifs

Nous rappelons que l'EO est la plus petite unité au sein d'un segment subjectif (SS). Elle est composée d'un seul mot d'opinion (un nom, un adjectif ou un verbe), éventuellement associé à certains modificateurs comme les mots de négation et les adverbes. Par exemple : حقا ليس جيدا (vraiment pas bon) est une EO composée d'un modificateur حقا (vraiment, d'une négation ليس (pas), et d'un adjectif جيد (bon). Une EO peut aussi être simplement un adverbe comme dans متحمسا (enthousiaste). Enfin, nous croyons également que l'extraction des expressions de recommandation, telles que : اشتر هذا الكتاب، سوف لن تندم (achète ce livre, tu ne vas pas regretter), qui sont très fréquentes dans les commentaires, et qui peuvent améliorer la performance de processus d'identification des textes subjectifs.

Nous croyons qu'il est difficile, voire impossible de créer un lexique regroupant tous les mots de sentiments propres a un domaine étudié, et le problème qui se pose dans ce cas est comment identifier les autres mots de sentiments dans les textes évaluatifs sans leur présence dans le dictionnaire ?

Dans notre approche, et pour extraire les textes porteurs d'opinions explicites, nous avons proposé de faire une projection simple du dictionnaire de sentiments sur les textes évaluatifs, et nous avons fait un enrichissement du dictionnaire manuellement qui repose sur l'expertise humaine afin d'avoir le maximum de textes subjectifs. La figure suivante illustre le processus d'identification des textes subjectifs avec un enrichissement du lexique.