**R**épublique **A**lgérienne **D**émocratique et **P**opulaire
**M**inistère de l'**E**nseignement **S**upérieur et de la **R**echerche **S**cientifique

**U**NIVERSITE **M**OULOUD **M**AMMERI DE **T**IZI **O**UZOU

**F**ACULTE DU **G**ENIE **E**LECTRIQUE ET D'INFORMATIQUE
**DEPARTEMENT D'INFORMATIQUE**

# Mémoire de Fin d'Etude de MASTER ACADEMIOUE

## Filière : Informatique
## Spécialité : conduite de projet informatique

***Présenté par : AICHOUCHE Sadia*** ***SEMAI Sabrina***

**Thème**

# Implémentation d'un système de classification de textes courts

*D*evant le jury composé de :
**Président : Mr HAMMACHE Arezki MCB**
**Examinateur : Mr SADOU Samir MAA**
**Examinateur : Mme AIT YAKOUB Zina MAB**
**Encadreur : Mr SAIDANI FaycalRedha MAB**

**Promotion 2018/2019**

# Remerciements

*En premier lieu,*
*Nous tenons à remercier **ALLAH** qui nous a aidés et nous a donné la patience et la force à accomplir ce travail.*

*Nos vifs sincères remerciements*
*S'adressent spécialement à notre promoteur,*
**M^r^SaidanifayçalRédha**

*Dont nous avons eu la chance de l'avoir comme Encadreur pour sa confiance, ses encouragements continuels, et son suivi de prés de nos travaux durant la réalisation de notre mémoire.*

*Nous adressons nos remerciements les plus sincères aux membres du jury qui nous font l'honneur de juger notre travail.*

*Notre profonde gratitude et sincères remerciements vont à tous les Enseignants qui nous ont suivis durant notre parcours d'étude.*

*Enfin, nous adressons nos plus sincères remerciements à tous nos proches et amis, qui nous ont toujours encouragés au cours de la réalisation de ce mémoire.*

***Merci à tous et à toutes.***

# Dédicace

*Je dédie ce  modeste travail à mes chers parents pour leurs sacrifices et leurs encouragements que Dieu les protège, à mes chers frères (Sofiane, Slimane, Cherif, Walid, Hamid et mon frère gémeaux Karim), à l'homme de ma vie Bilal , à  toute ma famille et à mes amis.*

***Sabrina***

*Je dédie ce travail à mes chers parents pour l'éducation qu'ils m'ont prodigué, avec tous les moyens et au prix de tous les sacrifices qu'ils ont consentis à mon égard, pour le sens du devoir qu'ils mon enseigné depuis mon enfance. Que Dieu les gardes et les protège.*

*A mes chères et adorables sœurs ( Lynda, Hayet).*

*A mes chers frères (Karim ,Farid).*

*A ma nièce(Ayla).*

*A ma collègue(Sabrina).*

*A tous ceux qui me connaissent de près ou de loin.*

***Sadia***

# Table des matières

# Résumé

L'analyse des sentiments ou l'analyse des opinions est l'étude computationnelle des opinions, des sentiments, des attitudes et des émotions des personnes exprimées dans le langage écrit. L'analyse des sentiments est l'un des domaines de recherche les plus actifs dans le traitement du langage naturel et l'analyse des textes ces dernières années. Sa popularité est principalement due à la gamme large d'applications, car les opinions sont au cœur de presque toutes les activités humaines, en particulier dans les médias sociaux. Il n'est donc pas surprenant que la création et la croissance rapide du domaine coïncident avec celles des médias sociaux sur le Web. En fait, la recherche s'est étendue aux sciences de gestion et aux sciences sociales en dehors des médias sociaux et de l'informatique en raison de son importance pour les entreprises et la société dans son ensemble. Nous présentons une méthode d'analyse de sentiments basée sur le lexique à partir d'un corpus de messages issus de l'application Twitter. Nous insistant sur la phase de prétraitement des messages afin de déterminer la classification de ces tweets.

---

**Mots-clés** : Analyse d'opinion,Fouille d'Opinion, opinion mining, méthodes symbolique, statistique,Analyse de Sentiments, classification de polarité ,classification des opinions, Opinion, polarité.

# Liste des Figures

# Liste des tableaux

# Liste des abbreviations

**CRM** Customer RelationshipManagement

**NLP** Natural Language Processing

**DASA** Dissatisfaction-oriented Advertising based on Sentiment Analysis

**SVM** Support Vector Machine

**RI** Recherche d'information

**TREC** TextRetrievalConference

**NTCIR** NII Test Collection for Information Retrieval System

**NIST** National Institute of Standard and Technology

**NII** L'Institut National en Informatique

**DEFT** Digital Evidence &ForensicsToolkit

**LRI** Laboratoire de Recherche en Informatique

**LIRMM** Laboratoire d'informatique, de robotique et de microélectronique de Montpellier

**LIMSI** Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur

**ELRA** European Language Resources Association

**DUC** Document Understanding confernce

**POS** Part Of Speech

**TAC** Text Analysis Conference

**TAL** Traitement automatique des langues

**NB** Naïfs Bayes

**TALN** Traitement automatique du langage naturel

**OS** L'orientation sémantique

**SO** Sémentic orientation

**PMI** PointwiseMutual Information

**Pos** Positive

**Neg** Negative

**URL** Uniform Resource Locator

**HTTP** HyperText Transfer Protocol.

# Introduction général

De nos jours, l'Internet se révèle plus que jamais un outil indispensable de partage d'informations, des outils, des fichiers multimédias, etc. Il offre aux internautes un lieu d'échanges de leurs opinions et avis sur divers sujets. À titre d'exemple des médias sociaux on trouve Twitter, avec près de 326 millions d'utilisateurs et plus que 500 millions message par jour d'après les statistiques twitter en 2019. Cela permet de rassembler des identités sociales telles que des individus, des organisations et des entreprises qui leur permet de connaître les points forts et les points faibles de leurs produits, d'estimer la perception du produit par les clients, afin d'améliorer leur produit.

Il permet également aux clients de, consulter les avis d'autres utilisateurs sur les produits qui leur intéressent, acheter des produits, lire les commentaires avant de choisir le film à voir au cinéma, voir des propositions d'autres personnes avant de choisir un article, etc.

Un nouveau domaine est né, l'analyse des sentiments et l'opinion Mining. Ce domaine est le descendant de son fameux ancêtre : fouille de données, plus connu sous le nom du Data Manning. Ce dernier, vise à rechercher les points de vue et les opinions des internautes.

Parmi les différentes approches existantes pour l'analyse du sentiment, certaines se basent sur des lexiques de termes subjectifs et d'autres sur l'apprentissage automatique.

Dans le cadre de ce mémoire nous nous sommes intéressés uniquement à la classification d'opinion, et pour ce faire on a choisi de travailler avec Les approches basées sur le lexique, qui consiste à utiliser un lexique de mots qui contiennent un sentiment. Ce lexique est soit externe c'est-à-dire construit indépendamment de tout corpus, il peut être général (SWN,SentiWordNet ,SUBJ lexique, General Inquiry, Wilson lexicon, Afinn) ou construit manuellement, soit généré automatiquement à partir du corpus (les mots qui contiennent une opinion sont extraits directement du corpus). À chaque mot du lexique est associé un score d'opinion. Ce score est traité différemment par les différentes approches pour le calcul du score d'opinion d'un document. La méthode la plus simple est de donner à un document un score d'opinion égale au nombre total de score de mots qui contiennent une opinion présente dans le document.

# Problématique

L'analyse de sentiment sur des textes court issue des médias sociaux, dans notre cas (Twitter) est confrontée à certaines problématiques dues à :

- ✓ la taille réduite limitée à 280 caractères par tweet.
- ✓ le type de langage utilisé, qui est très éloigné des normes du langage traditionnel avec ses conventions (telles que les hashtags, les retweet, etc.).
- ✓ utilisation d'un lexique particulier est souvent grossière et contient d'abréviations, des émoticônes, des acronymes.
- ✓ la syntaxe qui est parcellaire dans le meilleur des cas. Les données extraites de Twitter sont hautement bruitées, non structurées,

C'est ce que nous a amené à opérer notre système en prenant en compte toute ces problèmes, en utilisant les outils de classifications issues des approches basées sur des lexiques.

# Organisation du mémoire

Ce mémoire va être organisé de la façon suivante : un premier chapitre préliminaire, pour définir la fouille d'opinion. Un état de l'art va être étalé au cours de chapitre deux, le troisième chapitre, fera l'objet d'une présentation et d'implémentation de notre approche.

Le premier chapitre, introduit les concepts préliminaires et les difficultés de l'analyse d'opinion. Nous avons énuméré les principales tâches liées à l'analyse d'opinion ainsi son processus. En finale nous avons élaboré les deux approches de détections d'opinion.

Le deuxième chapitre, présente un état de l'art des deux approches de classification de polarité, celles basées sur le lexique et celles basées sur l'apprentissage automatique ainsi les principaux travaux liés à ces derniers .Ensuite on finira par déterminer les méthodes d'évaluation de la performance de classification.

Le troisième chapitre est composé de deux parties : la première partie présente les outils et les concepts utilisés pour l'implémentation et la réalisation de notre approche. La deuxième partie présente les tests d'évaluation et la discussion des résultats obtenus.

Finalement, nous clôturons ce mémoire par une conclusion générale

# Chapitre 01 : Description du domaine de la Fouille D’opinion

# 1.1. Introduction

Avec l'émergence du Web 2.0, de plus en plus de documents textuels, contenant des informations exprimant des opinions ou des sentiments sont mis à disposition. Cette omniprésence d'informations, sous diverses formes, a suscité un certain nombre de problématiques liées au traitement de ces données, parmi lesquelles on retrouve, l'analyse d'opinion. Cette dernière a connu un intérêt croissant depuis une quinzaine d'années et suscite un intérêt à la fois dans le monde académique et professionnel. Pour les chercheurs en TAL g(Traitement Automatique des Langues) ; l'analyse d'opinions également connue sous l'angélisme « *Opinin mining* » ou « *Sentiment Analysis* », constitue un des sous domaines de recherche ayant enregistré le plus de travaux durant cette dernière décennie.

Dans le domaine professionnel, l'analyse d'opinion a trouvé de nombreuses applications dans le domaine de la prédiction et de la supervision notamment avec l'essor des médias sociaux, des sites de e-commerce, du CRM, qui sont autant de lieux où s'expriment librement toutes sortes d'opinions et d'avis sur les produits, les marques, les entreprises, les personnalités, etc.

Dans ce chapitre, nous allons aborder les différents concepts et définitions liées à l'analyse d'opinion. Dans un deuxième temps, nous présenterons les trois grandes approches utilisées en détection d'opinion:(Approche symbolique ou linguistique, approche statistique ou approche basée sur l'apprentissage automatique et l'approche hybrid.

# 1.2. Définition et composantes d'une opinion

De par sa définition, une opinion est le jugement qu'un individu ou un groupe d'individus émet.Elle représente ses idées, ses convictions, ses appréciations, et ses évaluations sur un sujet ou un élément donné. [1]

Dans[2], les auteurs rivent l'importance qu'ont les opinions sur les individus, les gouvernements, les communautés sociales, ainsi que leur impact sur les processus décisionnels. A partir de là, les auteurs ont émis une définition formelle de l'opinion sous la forme d'un quintuple, qui fait actuellement, office de référence pour divers travaux en l'analyse d'opinions.

Bing [2] a proposé une représentation standardisée de l'opinion qui se note ainsi :

($O_j$,$f_{jk}$, $ooi_{jkl}$, $h_i$, $t_l$) Où :

$O_j$ : est un objet,

$f_{jk}$ : est une caractéristique de l'objet $O_j$,

$oo_{ijkl}$ : est l'orientation ou la polarité de l'opinion sur la caractéristique $f_{jk}$de l'objet $o_j$

$h_i$ : est le détenteur de l'opinion ou la source de l'opinion.

$t_l$ : le temps où l'opinion a été exprimée par $h_i$.

Cette formule de l'opinion décrite par Bing, permet de fixer les principaux axes de recherche liés à l'analyse d'opinion. Néanmoins elle est utile pour fixer le cahier des charges d'une application qui veut réaliser de l'opinion mining et du sentiment analysis. Car l'absence d'un de ces éléments rend l'analyse particulièrement superficielle.

A savoir le premier quintuplé concerne l'objet sur lequel porte l'opinion Oj ; qui est en général défini clairement au sein de textes, il peut s'agir d'un produit, service, personne… ; ainsi le deuxième quintuplé consiste aux caractéristiques de l'opinion fjk ; qui sont plus complexes à déduire de faite que le problème consiste d'abord à trouver le lien entre ses caractéristiques et l'objet sur lequel elles portent.

On a aussi les deux autres aspects, qui concernent le temps ou l'opinion a été exprimé et le l'entité qui a exprimé cette opinion. En effet, la temporalité est importante pour juger l'évolution des opinions vu que dans certains cas une opinion qui date trois ans n'est pas forcément la même avec celle d'aujourd'hui.

Enfin, le troisième quintuplé, concerne la polarité de l'opinion qui peut être positive, négative ou neutre, ou peut être exprimée par des niveaux d'intensité.

# 1.3. Les taches liées à l'analyse d'opinion

Les cinq composantes d'opinionprésentées par Bing ont permis de faire émerger les principales tâches liées à l'analyse d'opinion. Ainsi on trouve cinq grandes catégories de travaux qui peuvent être mise en valeur.

- **Classification de subjectivité**

La détection de la subjectivité est la tâche d'identification des mots, expressions et phrases subjectifs.

D'après [3] ; identifier la subjectivité permet de séparer les opinions des faits.

La classification de subjectivité reflète l'avis d'une personne à propos d'une entité, un objet spécifique, etc. son objectif est de pouvoir distinguer entre les propositions objectives et subjectives.

La classification de subjectivité est considérée comme une étape préliminaire à la classification de polarité d'opinion. D'abord, on repère les documents porteurs d'opinion (classement objectif/subjectif), pour ensuite, attribuer aux documents subjectifs une polarité (positive ou négative).[4]

Dans notre cas on suppose que les documents sont subjectifs.

- **Classification de polarité**

Classification de polarité est une des tâches les plus étudiés en analyse d'opinion, elle consiste à déterminer si un document ou un texte subjectif est soit positif soit négatif. Par exemple, évaluer un produit ou un service, pour comprendre les opinions du public sur des événements sociaux ou sur des mouvements politiques ou aussi savoir l'avis des gens sur un film spécifique. Nous allons aborder cette tâche plus en détail dans le chapitre suivant.

- **Classification des émotions**

La recherche sur les émotions a récemment attiré une attention accrue de la communauté NLP ; elle consiste à détecter les expressions d'émotions dans les textes ou les documents ; elle est considérée comme une tache détaillée de la tache de classification de polarité en identifiant de manière plus fine les expressions d'opinion.

Il existe plusieurs théories de l'émotion, à savoir, celle d'Ekman qui a proposé une liste pour six catégories d'émotions de base qui représentent les expressions faciales d'émotion distinctement identifiables « bonheur, tristesse, colère, dégoût, surprise et peur » [5].

En général, la classification de l'émotion est liée au domaine de l'informatique affective, qui est une étude et un développement de systèmes et d'appareils ayant les capacités de reconnaître, d'exprimer, de synthétiser et modéliser les émotions humaines.

- **Détection de la source d'opinion**

La détection d'opinion sert à identifier les sources directes et indirectes pour tous les types d'opinions, d'émotions, de sentiments et d'autres états privés qui sont exprimés dans le texte,

l'émetteur de l'opinion est simplement l'auteur de texte qui peut être un ou plusieurs personnes.[6]

- **Détection des cibles d'opinion**

La détection ou extraction de l'opinion peut être vue et réalisée grâce à un système composé de deux modules principaux, illustrés en figure suivante : d'une part, le sujet d'un texte est extrait, par exemple avec des techniques issues du domaine de l'extraction d'information. Ce sujet devient la cible de l'opinion. En parallèle, parfois en dépendance, le sentiment du texte est calculé et devient la polarité de l'opinion. [7]

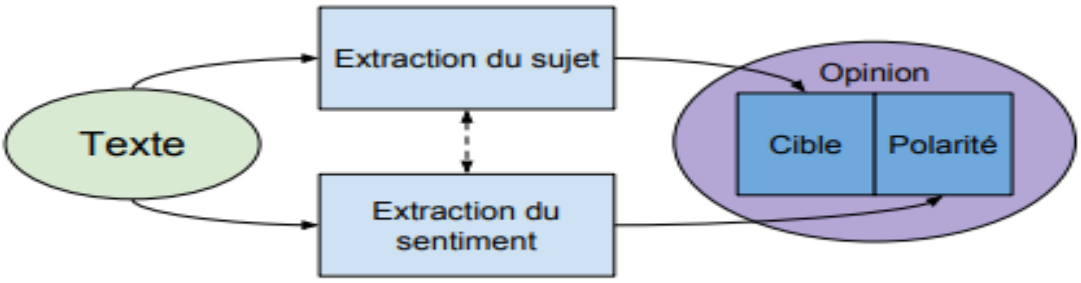


**Figure1.1 :** Extraction classique des opinions d'un document [7]

# 1.4. Difficultés de l'analyse d'opinion

L'opinion et le sentiment sont généralement exprimés par des verbes et des adjectifs qualificatifs. De plus, l'ajouter d'adverbes permet de modifier la force et l'intensité de l'adjectif ou du verbe. Ces dernières sont le plus souvent décrits par la polarité et sont en général, soit positive (opinion favorable), soit négative (opinion défavorable), soit neutre.

Nous montrons ci-dessous quelques difficultés liées à l'analyse d'opinions :

- **Difficulté due au contexte**

Un mot positif ou négatif peut avoir un sens inverse en fonction du contexte.

Par exemple « *J'ai fait un excellent travail* » peut-être interprété comme une affirmation positive. Cependant, dans « *mon fournisseur d'Internet fait un excellent travail quand il s'agit de me voler de l'argent* », faire un bon travail n'est plus une chose positive, basée sur le contexte (« *me voler de l'argent* »).

- **Difficulté due au Sarcasme**

L'un des traits les plus difficiles à interpréter est le sarcasme. C'est un sentiment d'ironie ou une raillerie est tournée en dérision envers une personne ou une situation donnée. Malheureusement, ce trait est encore mal interprété par les systèmes d'analyse d'opinions existants.

Par exemple, la phrase « *bien sûr, je suis contente que mon navigateur plante au milieu de mes cours* » est une déclaration sarcastique (négative), malgré la présence du terme positif « *contente* ». Aussi, cela dépend souvent du contexte, comme c'est le cas dans cette phrase ou nous savons que, le fait qu'un navigateur se bloque est négatif. [8]

- **Difficulté due à l'ambiguïté de sentiment**

Une phrase avec un mot positif ou négatif n'exprime pas nécessairement un sentiment. Par exemple, *« pouvez-vous recommander un bon outil que je pourrais utiliser ?* » N'exprime aucun sentiment, bien qu'il utilise le mot positif « *bon* ». De même, les phrases qui ne comportent aucun terme à caractère opiniâtre peuvent également exprimer un sentiment subjectif. Ainsi, la phrase « *ce navigateur utilise beaucoup de mémoire* » ne contient aucun terme subjectif malgré le fait qu'elle exprime clairement un sentiment négatif.

- **Difficulté due aux méthodes d'analyse utilisées**

De manière générale, la représentation en Sac de mots sert à décrire de manière compacte un document texte en recherche d'informations. Toutefois, cette représentation aveugle d'un document, prête parfois à confusion. Notamment, lorsqu'il est nécessaire d'effectuer une analyse approfondie ou l'aspect sémantique est important.

Ainsi, dans les deux phrases suivantes, on remarque qu'elles contiennent les mêmes sacs de mots, sans pour autant exprimer les mêmes sentiments. La première phrase contient un sentiment positif alors que dans la deuxième il est négatif : « *Je l'ai apprécié pas seulement à cause de ... », « Je l'ai pas apprécié seulement à cause de ...* ». [9]

- **Difficulté due à l'emploi d'une thématique**

Une même thématique peut être utilisée dans différentes classes et peut exprimer une toute autre signification.

- **Difficulté due au langage qu'utilisent les internautes**

La plupart des internautes utilisent un langage spécifique à eux pour s'exprimer. Des mots spécifiques sont utilisés tels que : « *ha haha* », «*Goood* », « *super* ».Les ponctuations non plussont pas forcément utilisées pour marquer les fins de phrases.

- **Difficulté due aux variations régionales**

Un mot peut changer le sentiment et la signification en fonction de la langue utilisée. Ceci est souvent vu dans l'argot, les dialectes et les variations de langue. Un exemple est le mot «sick», qui peut changer le sens en fonction du contexte, du ton et de la langue.

Example: *«that is a sick song!* »versus *«I'm not feeling well at all, I might be sick").*

Un autre exemple de variation régionale peut être trouvé entre l'anglais Britannique et américain pour des mots comme « *quite* », « *rather*», « *pretty* ». En anglais britannique ces mots prennent le sens de «*fairly* », tandis qu'en anglais américain ils prennent le sens de "*very*". Cela peut parfois être mal compris dans les conversations quotidiennes. Il n'est donc pas étonnant que les outils d'analyse puissent trouver cela comme étant problématique.

- **Difficulté due au vocabulaire**

Le vocabulaire qu'on utilise pour exprimer une opinion diffère d'une personne à une autre. Par exemple, un anglo-saxon lorsqu'il exprime ses sentiments, utilise souvent des mots bien représentatifs de ce qu'il ressent, contrairement aux personnes qui ne connaissent pas ou peu sa langue. Néanmoins, il existe des mots dont l'orientation qui peut changer selon le contexte dans lequel ils sont employés. Il peut s'agir de mots polysémiques ou bien d'homonymes ayant des axiologies différentes. C'est le cas du "navet" qui est un légume tout à fait ordinaire en cuisine mais un film à éviter dès lors que l'on parle de cinéma. La désambiguïsation sémantique (savoir quel sens est effectivement utilisé) s'appuie justement sur les mots du contexte. Les méthodes existantes utilisent des corpus, annotés ou non, ainsi que des dictionnaires inventoriant les sens existants. L'orientation d'un mot non polysémique peut également changer à l'intérieur d'un même domaine, selon l'objet qu'il évalue. Par exemple, pour un ordinateur portable, une batterie "large" est un inconvénient mais un écran "large" est un atout. L'orientation des mots peut aussi dépendre des préférences et de l'idéologie de l'auteur, c'est alors bien plus difficile à détecter. Les textes politiques sont notamment très sensibles à cela. Par exemple, le mot "bourgeois" est fondé sur une sémantique neutre mais quand il s'agit de préjugé ou d'opinion, ce qui est "bourgeois" est souvent mal vu.[10]

# 1.5. Domaines d'application de la fouille d'opinion

Comme le rappelle Pang Lee dans « *Opinion Mining and Sentiment Analysis* », le sentiment de « *ce que les autres pensent* » est régulièrement invoqué dans tout processus décisionnel. Que ce soit, en vue de l'achat d'un produit, dans le contexte d'une élection ou encore pour évaluer la réputation d'une marque.

La figure ci-dessous,présente quelques domaines d'application de l'analyse des sentiments.

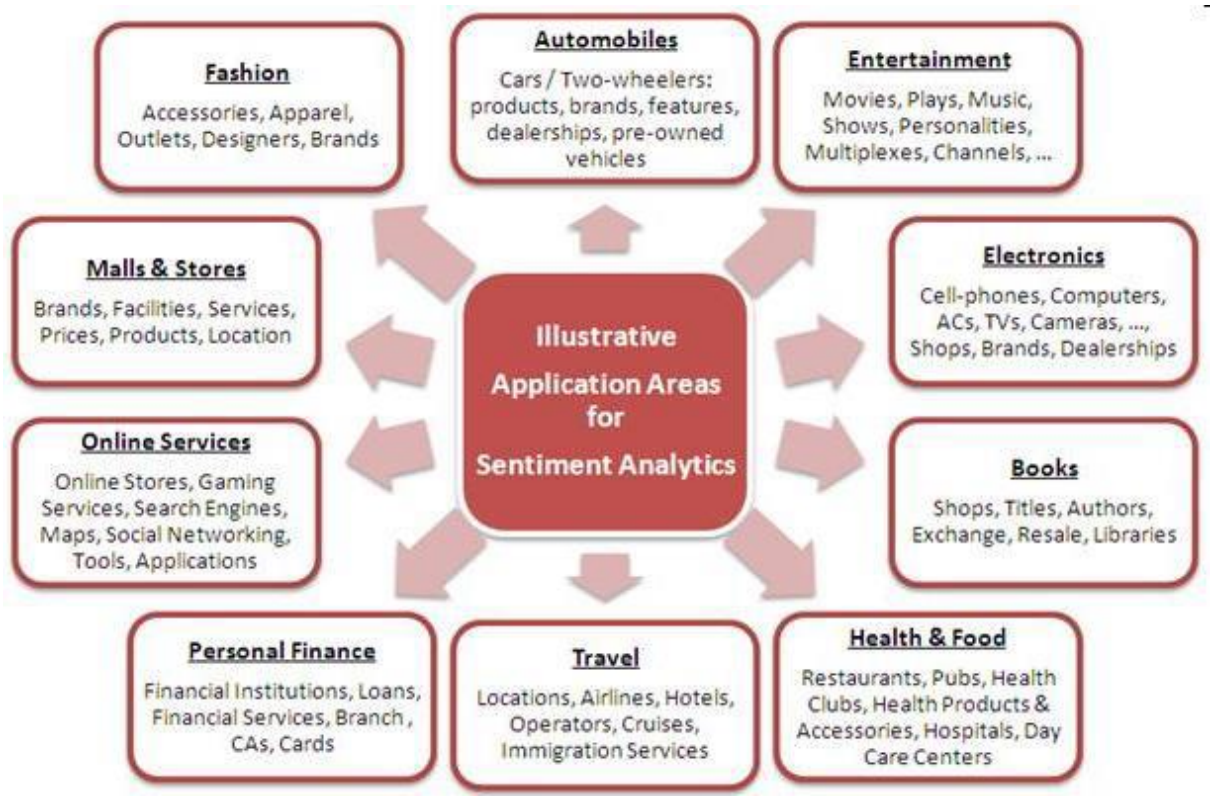


**Figure1.2 :** Domaines d'application d'analyse des sentiments. [11]

- **Domaine du Marketing**

Le marketing a rapidement compris l'intérêt de l'analyse des sentiments.

- du côté des entreprises, cela 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.

- du côté client, cela lui permet de 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. L'intérêt majeur d'un processus d'analyse d'opinion est de fournir à un client un aperçu sous forme de résumé (pourcentage positif/négatif), concernant un produit à acquérir sans avoir forcément besoin de lire l'ensemble des commentaires d'autres utilisateurs. [12]

- **Domaine du politique**

La publication croissante sur internet de textes à teneur politique (lois, rapports, billets de blogs politiques, etc.) est le constat que la politique ne se fait plus seulement dans des cercles réduitsmais aussi dans les débats en ligne. Cela, a conduit certains chercheurs à utiliser les techniques d'analyse des sentiments pour déterminer l'accord ou le désaccord des commentateurs avec les différentes propositions de loi. Les acteurs politiques suivent également cette tendance, en essayant de récolter l'avis des internautes avant de promulguer une nouvelle loi. L'analyse d'opinions permet également de connaître l'appréciation des potentiels électeurs sur homme politique durant une élection présidentielle. [13]

- **Domaine de la veille**

Le faite que les techniques d'analyse des sentiments permettent de classer de grandes quantités de textes, rapports, conversations informelles sur des produits ou des dirigeants d'entreprises, etc., peuvent être utilisées dans le domaine de la veille, au sujet du produit, qu'elle soit économique, technologique, stratégique ou institutionnelle. [14]

- **Domaine du la publicité en ligne**

Si la publicité en ligne, ciblée et contextuelle, s'est considérablement développée ces dernières années, elle pourrait bénéficier des recherches en sentiment analysis. Car si une annonce publicitaire est d'autant plus efficace qu'elle apparaît au bon endroit au bon moment, elle pourrait l'être encore plus si elle s'adaptait au ressenti des consommateurs vis-à-vis d'un produit ou d'un service. C'est ce que propose par exemple la stratégie DASA (*Dissatisfaction-orientedAdvertisingbased on Sentiment Analysis*), qui a pour but de détecter et de prendre en compte les points d'insatisfaction des consommateurs afin d'adapter encore mieux les annonces publicitaires à leurs cibles. [15]

- **Domain du la prédiction des marchés financiers**

(Bollen ,j. , Mao,H., and Zeng)[16]ont étudié la prédiction du marché en exploitant les opinons véhiculé sur Twitter à propos de données financières. L'analyse s'est faite à l'aide d'outils de suivi de l'humeur, à savoir, OpinionFinder qui mesure l'humeur positive et négative, ainsi que, Google Profile of Mood States qui permet de mesurer l'humeur sur une échelle de six valeurs. [17]

# 1.6. Approches de la détection d'opinion

Plusieurs méthodes ont été utilisées pour la détection d'opinions. Leur but est de réordonner les documents selon un score d'opinion. Ainsi, les documents qui contiennent le plus d'opinions sont classés parmi les premiers.

En général, il existe trois types d'approches pour la détection d'opinion, l'une basée sur un lexique, l'autre sur l'apprentissage machine et la dernière est une combinaison des deux approches précédentes.

## 1.6.1 L'approche lexicale

(Appelées aussi approche symbolique ou encore approches basées sur lexique). Les approches lexicales utilisent des dictionnaires de mots subjectifs exprimant une opinion, ces dictionnaires peuvent être généraux comme le General Inquirer, SentiWordNet, Opinion Finder, Wilson lexicon etc., ils peuvent également être construits en fonction des corpus (les mots qui contiennent une opinion sont extraits directement du corpus).

Dans ces dictionnaires, un score d'opinion est associé a priori à chacun des mots, ce score est traité différemment par les différentes approches pour le calcul du score d'opinion d'un document. La méthode la plus simple est de donner à un document un score d'opinion égal au nombre total de mots qui contiennent une opinion présente dans le document, ce qui se résume à une fréquence des termes au niveau du document. L'autrepossibilité consiste à considérer la similarité entre les documents d'opinions et non opinions, pour le calcul de leur score de subjectivité.

Ces dictionnaires ont été constitués de différentes façons :

- à la main ;
- à partir de corpus ;
- à partir de dictionnaires existants.

## 1.6.2. L'approche basée sur l'apprentissage automatique

(Appelées aussi classification supervisée, ou encore approches statistique). Ces approches utilisent des classifieurs. Des données représentent des phrases subjectives (ou des documents avec opinion) sont fournies aux classifieurs pour l'apprentissage. Le classifieur génère ensuite un modèle, qui sera utilisé dans la partie test.

D'après Pang et Lee dans « Opinion Mining and Sentiment Analysis » des « features » sont utilisées pour l'apprentissage tel que les bigrammes, les *n*-grammes, POS (étiquettes morphosyntaxiques) etc. Plusieurs types de classifieurs ont été utilisés : SVM, Naive Bayes, Multiples Classifieur, Naïfs de Bayes, ainsi que la régression logistique.

- **Naive Bayes** est une approche probabiliste, qui utilise une loi de Bayes, où les probabilités sont en fonction des mots contenus dans les documents :

$$P(c/d) = P(c) * \frac{P\,(d/c)}{P\,(d)}$$

Avec *d* est un document et *c* la classe du document. « *P(c/d)* » est déterminé par le classifieur Bayésien naif.

- **L'approche SVM** repose sur la notion d'hyperplan séparateur et de marge maximale. Un hyperplan séparateur entre deux ensembles de points (ensemble de documents de polarité positive et l'ensemble de documents de polarité négative) est la frontière entre ces deux ensembles. La marge représente la distance entre un de ces ensembles et cet hyperplan.

- **La régression logistique** est une méthode statique permettant de produire un modèle pour décrire des relations entre une variable catégorielle et un ensemble de variables de prédiction.

## 1.6.3. L'approche hybride :

(Appelées aussi classification semi-supervisées)**.** Ces approches combinent les points forts des approches symboliques et statistiques. Elles prennent en compte tout le traitement linguistique des approches symboliques avant de lancer le processus d'apprentissage comme dans les approches statistiques.[18]

La combinaison des approches symboliques et statistiques a donné des résultats plus précis que chacune des méthodes employées séparément. [8]

# 1.7. Processus de la fouille d'opinion

La fouille d'opinions est considérée comme un processus de traitement de données textuelles répartit en plusieurs étapes. Il prend en entrée un ensemble de documents et fournit en sortie un ré-ordonnancement des documents selon leur polarité. On peut ainsi découper le processus de fouille d'opinions en trois étapes principales, illustrées dans la Figure 1.3.

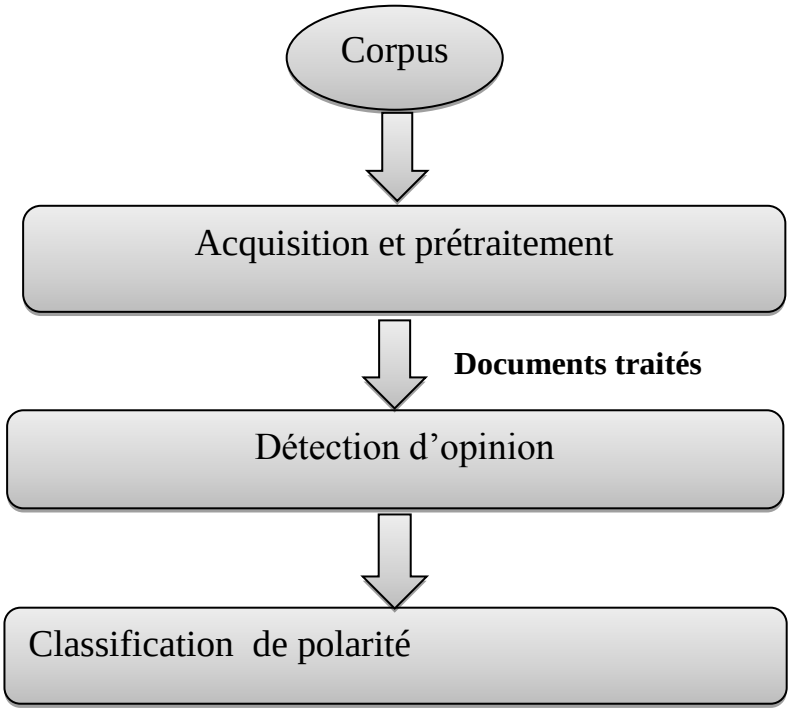


**Figure1.3 :** processus de la fouille d'opinion.

## 1.7.1 Acquisition et prétraitement des données

Cette étape porte sur l'accès aux données. Ces dernières peuvent provenir de divers sources et sous différentes formes. Elles peuvent être sous un aspect brut ou sous forme des données crawlée à partir des sites web tels que Facebook, Twitter, forum, blogs…etc. Ces informations sont généralement récoltées, en vue de construire des corpus de données spécialisés.

Dans la phase de prétraitement, les textes sont prétraités linguistiquement en éliminant les mots vides et les mots qui n'apportent aucune information importante. Ainsi, la suppression des mots fréquents et des mots rares au sein du document d'apporter un gain en termes de temps et de puissance lors de la phase d'analyse

Au cours de cette étape, on a souvent recours à un étiquetage grammatical afin de mettre en valeur des natures de mots susceptibles d'apporter une information opiniâtre telle que les adjectives, les adverbes, etc. On retrouve également, certains travaux ou les grammaires de dépendances sont utilisées pour structurer la phrase de manière hiérarchique.

### 1.7.2. La Phase de détection d'opinion

La détection d'opinions utilise plusieurs méthodes pour le but de déterminer la subjectivité des documents prétraités et les classer en deux oppositions (subjectif/objectif).

### 1.7.3. La classification de polarité

La classification de polarité est une sous tâche de la détection d'opinion, elle permet de déterminer si un document porte une opinion positive, négative ou neutre sur un sujet donné, puis de classer L'ensemble des documents du corpus selon leurs types de polarité. Il existe deux types de classifications : binaire ou multi-classes.

La classification binaire se définit sur une échelle : positive et négative. En contrepartie, la classification multi-classes peut définir plusieurs classes telles que : fortement positive, positive, mixte, négative, fortement négative.

Une fois l'ensemble des étapes du processus sont terminées, une évaluation de ces résultats sont généralement confrontés à la perception humaine. La comparaison est faite grâce à des mesures de similarité.

Plusieurs campagnes d'évaluation ont vu le jour, permettant aux chercheurs de présenter leurs travaux et les évaluer sur des collections test élaborées par ces campagnes.

## 1.8. Les campagnes d'évaluation

### 1.8.1 DEFT

DEFT ou Défi fouille de texte est une campagne d'évaluation scientifique francophone portant sur la fouille de textes. Il a été créé en 2005 par un groupe de chercheurs (Prince et al, 2007)[ 63][19],dans le but d'initier une série de campagne d'évaluation francophones sur des thématiques relevant de la fouille de textes.

Le défi est organisé depuis 2005 par des chercheurs du LRI(Laboratoire de Recherche en Informatique, Orsay) et du LIRMM (Laboratoire d'informatique, de robotique et de microélectronique de Montpellier), puis du LIMSI (Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur, CNRS), dès 2007 avec le soutien du consortium EuropeanLanguageResources Association (ELDA/ELRA),ce défi est proposé tous les ans sur une thématique différente. Chaque édition a rassemblé une dizaine d'équipes participantes, pour la plupart issues des laboratoires de recherche publics français. Les campagnes DEFT visent essentiellement à mettre à disposition à la fois des protocoles et des outils de mesure des systèmes d'analyse du langage, et des corpus d'expérimentation en français, sur le long terme.

Quatre corpus de sources différentes répondant à deux critères principaux : en premier lieu, la capacité d'accès aux données et en second lieu, la possibilité d'extraire du corpus les données de référence de la tâche. Les corpus sont : un corpus de critiques de films et de livres, un corpus de critiques de jeux vidéo comportant pour chaque critique à la fois un texte évaluatif et une note globale appréciative, un corpus de relectures d'articles de conférences comportant le texte évaluatif et la notification d'acceptation ou de rejet et enfin un corpus de débats parlementaires sur des projets de lois auquel nous avons pu associer à chaque intervenant dans les débats les méta-données de son vote pour ou contre le projet de loi.

- **DEFT 07** : discute détection de l'opinion exprimée dans un texte, quatre corpus, deux à trois classes (positif, neutre, négatif) par corpus.

- **DEFT 08** : classification automatique de textes en genre et en thèmes différents (art, économie, littérature, politique internationale, politique nationale, problèmes de sociétés, sciences, sports, télévision).
- **DEFT 13**: identification du niveau de difficulté de réalisation d'une recette, identification du type de plat préparé, appariement d'une recette avec son titre, identification des ingrédients d'une recette.
- **DEFT 15** : discute la fouille d'opinion, de sentiment et d'émotion dans des messages postés sur Twitter.

### 1.8.4. La campagne DUC/TAC

Depuis 2010, le NIST a organisé des compagnes d'évaluation de la performance des algorithmes de TAL, document Understandingconfernce (DUC). À partir de 2008, la compagne a changé d'appellation pour TextAnalysisConference(TAC).

Cette dernière campagne est beaucoup plus ambitieuse que la précédente. En 2008, la compagne TAC a organisé atelier portant sur trois volets : Summarization, Question &Anwering et RecognizingTextualEntailment. L'objectif des compagnes DUC/TAC est double : d'un côté, il s'agit de promouvoir les progrès réalisés dans le domaine du résumé automatique de documents et d'un autre, de permettre aux chercheurs de participer à des expérimentations de grande échelle, tant au point de vue du développement que de l'évaluation de leurs systèmes. [20]

## 1.9. Conclusion

Dans ce chapitre, nous avons présenté le domaine de la fouille d'opinion, en commençant par la définition de l'opinion ainsi quelques concepts de base de ce domaine comme les besoins et domaine d'application de ce dernier.

Nous avons montré un processus de la fouille d'opinion et ses différentes étapes. Nous avons également abordé quelques facteurs qui rendent difficile la fouille d'opinion. Nous avons aussi les trois grandes approches de détection d'opinion, pour finir avec une présentation des compagnes d'évaluations de la fouille d'opinions les plus populaires.

Dans le chapitre suivant, nous allons aborder la tache de la classification d'opinion et ses déférentes approches.

# Chapitre 02 : Etat de l’art

# 2.1. Introduction

La classification de polarité d'opinion est une des phases importantes en analyse des sentiments [21], elle s'agit de regrouper les opinions des utilisateurs envers un objet selon leurs orientations sémantiques en différentes classes de polarité, soit binaire « contient une opinion positive ou contient une opinion négative », soit multi-classe définie en trois classes « positive, négative, neutre », ou sur un axe plus large« fortement positive, positive, neutre, négative, fortement négative ».La plupart des travaux se sont focalisés sur la classification binaire, mais la classification multi-classes peut être aussi utile dans certaines applications telles que les systèmes de recommandation en e-commerce.

La classification de polarité apporte une information de plus que la détection d'opinions vue qu'elle classifie les opinions selon leur type de sentiments. 7ette tâche a attiré l'attention de plusieurs chercheurs en TAL, à titre d'exemple détection des points faibles et des points forts d'un objet commenté par les utilisateurs.

Les techniques de la classification de polarité sont répertoriées selon deux principales approches, la première est dite classification basée sur le lexique, la deuxième est dite classification basée sur l'apprentissage automatique.

Dans ce chapitre, nous présentons, la définition de quelques concepts essentiels, les deux approches de classification, ainsi les différents travaux réalisés sur ces approches. Enfin, les mesures de performances largement utilisées pour l'évaluation y seront présentées.

# 2.2. Approches de classification

Comme pour la détection d'opinion, il existe deux approches pour la classification de polarité, l'une basée sur le lexique, l'autre sur l'apprentissage machine. La première, fait d'abord une analyse du texte phrase par phrase, en extrait ensuite les relations qui véhiculent des sentiments, tandis que la deuxième traite les textes en une seule phase et attribue un sentiment global au texte entier à la fin du traitement. La figure ci-dessous illustre les méthodes de classification d'opinion.

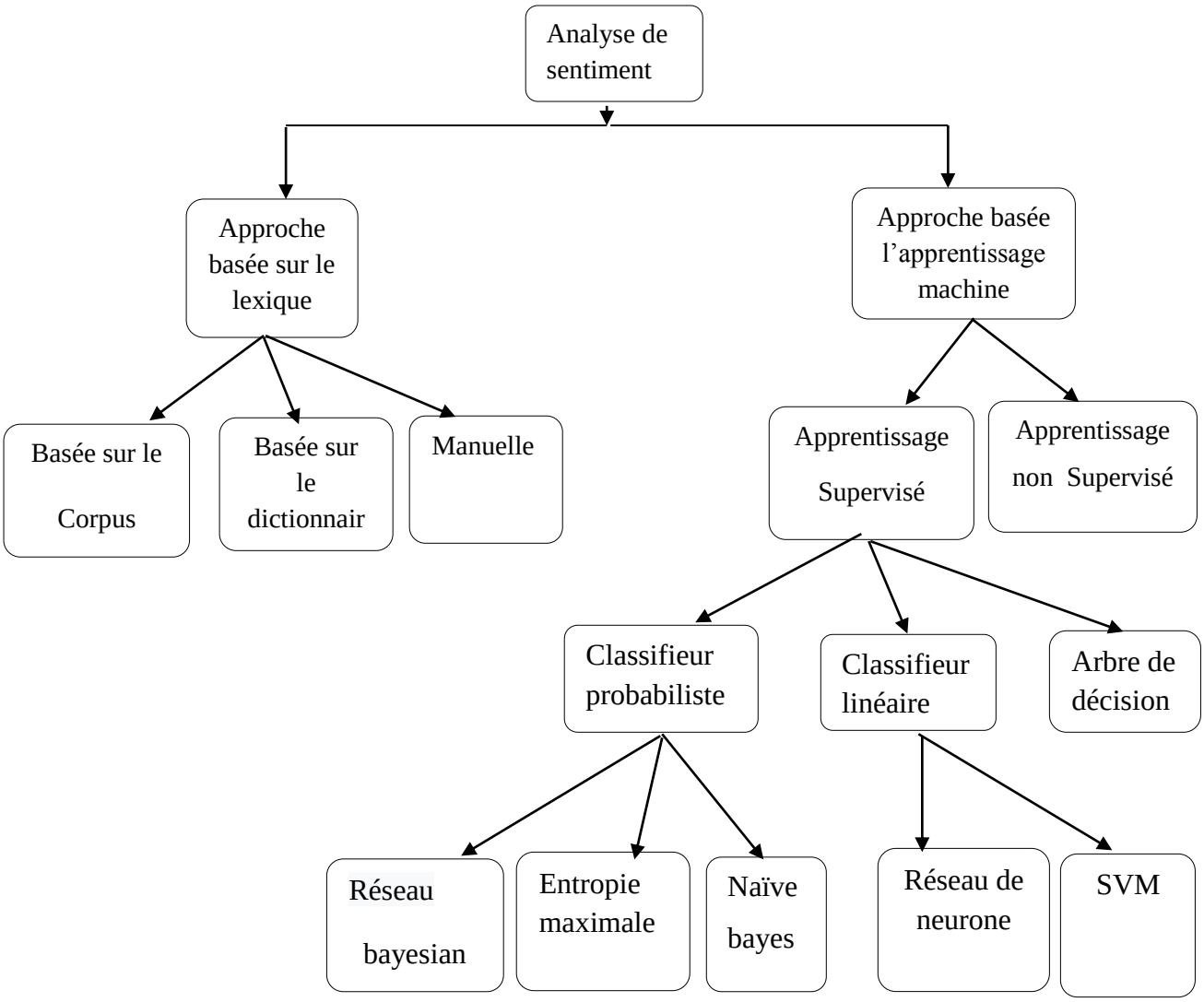


**Figure 2.1**: les méthodes de classification d'opinion

## 2.2.1. Approche basée sur le lexique

Cette approche est basée sur l'hypothèse que l'orientation de sentiment contextuelle est la somme de l'orientation de sentiment de chaque mot ou phrase.

Les méthodes basées sur le lexique pour lecalcul de polarité consistent à analyser syntaxiquement le texte et le découper en phrases, puis vérifier chaque phrase si elle contient des relations de sentiment en employant une grammaire spéciale.Les relations syntaxiques de base sont généralement employées, ainsi que d'autres relations plus complexes.

Cette méthode se base sur les mots du lexique de sentiments ayant reçu une annotation marquant le caractère positif ou négatif du mot. Il s'agit pour la plupart de verbes (aimer, apprécier, détester, ...) et d'adjectifs (magnifique, superbe, insupportable, ...). Puis, pour mesurer plus précisément la force de l'opinion exprimée dans une phrase, on a souvent recours à l'extraction des adverbes associés aux adjectifs [12]. Pour ce faire, Benamara et al. [22]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. 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 suivant présente quelques exemples de mots classés dans les catégories définies par Benamara et al.[22].

| 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 littele, almost, a bit, little, rather, nearly, barely, scarcely, weakly, slightly… |
| Négation | Not, never, less, no…. |

**Tableau 2.1: Exemple de catégories d'adverbes**

La négation est une caractéristique importante pour déterminer la polarité de l'opinion, Das et Chen [23] ont déterminé que la négation dans une phrase inverse le sens de la phrase. Ils ont discuté sur comment les mots tels que *« not », « never »* ou no sont utilisés pour inverser le sens d'une phrase.

Pour mettre en valeur la négation de la phrase, Ils proposent de rajouter une indication de négation *« NOT »* à des mots qui se trouvent près de la négation, de sorte que dans la phrase *« I do not likethismovie»* , les lemmes *« like »* , *« this »*, *« movie »* sont convertis en des nouveaux lemmes *« I do not NOT_likeNOT_thisNOT_movie ».*

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.

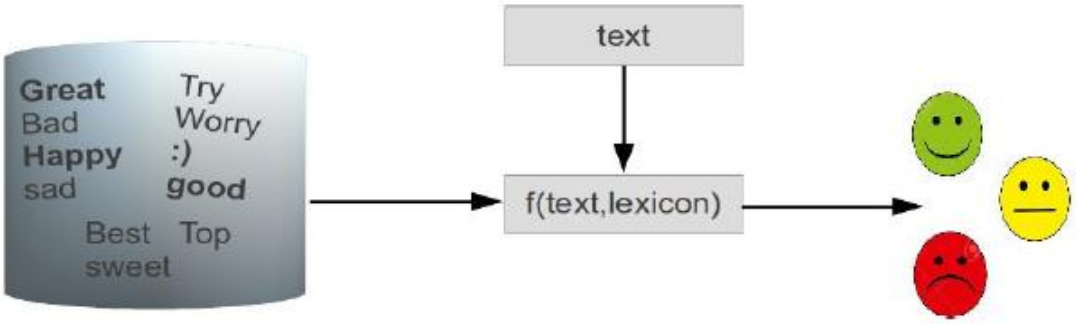


**Figure 2.2 :** Approche d'analyse de sentiment basée sur le lexique.[20]

### 2.2.1.1. Phase 1 : Construction des lexiques d'opinion

Quelle que soit la méthode utilisée pour le classement de polarité, il est avant tout important de disposer d'un ou plusieurs lexiques d'opinion permettant de précéder à l'analyse du texte, ce lexique peut être créé manuellement [24][25] ou en expansion automatiquement à partir d'une graine de mots [26].

- **La méthode manuelle**

L'idée est de construire manuellement un lexique et d'ajouter des mots comme positif ou négatif. Cette méthode consiste à remplir le lexique de mots d'opinion sans l'aide d'outil particulier, c'est les experts qui font la sélection de mots et expressions porteurs d'opinion ainsi que le choix de leur polarité.

Cet ensemble de mots est appelé graine ou germe *« seedwords »*et permet de construire une première liste de mots et d'expressions, qui sera utilisée par la suite afin de trouver, répertorier et classer d'autres mots et expressions porteurs d'opinion.

**Inconvénients :**

- Cette approche prend beaucoup de temps et n'est donc généralement pas utilisée.
- Une part non négligeable de subjectivité de la part des experts peut entrer en jeu et peut entraîner certaines erreurs de classification.

- **Méthode basée sur le corpus**

Cette approche repose sur l'affirmation suivante : « Un document doit être positif (resp.Négatif) s'il contient de nombreux mots positifs (resp.Négatifs).En contrepartie, un mot doit être positif (resp.Négatif) s'il apparaît dans de nombreux documents positifs (resp.négatifs)».Ces méthodes génèrent des mots d'opinion avec une précision relativement élevée. La plupart deces méthodes à base de corpus ont besoin d'un grand nombre de données d'entraînement. L'avantage de ces méthodes, c'est le fait de générer des mots d'opinion spécifiques au domaine ainsi que leurs orientations.

Elles contribuent à enrichir le dictionnaire de différentes manières :

- En cherchant les "co-occurrences" de mots dans le corpus, en comptantle nombre de fois ou ces mots ou expressions apparaissent dans le corpus à coté de mots ou expressions présents dans la liste de germes, pour déterminer la polarité de mots ou expressions non classées. Donc, un mot apparaissant plus souvent à coté de mots positifs sera systématiquement classé dans la catégorie positive et inversement.

- En exploitant un ensemble de règles linguistiques, pour identifier plus de mots d'opinion et leurs orientations au sein du corpus. Cette méthode consiste à utiliser les conjonctions de coordination présentes entre un mot déjà classé et un mot non classé. [ 27],[28]

  L'une des règles concerne la conjonction *« AND »*, qui dit que les adjectifs conjoints ont généralement la même orientation. À l'inverse, si la conjonction BUT sépare un mot classé positif et un mot non classé, alors ce dernier sera considéré comme étant négatif. Des règles ont également été conçues pour d'autres connecteurs, à savoir, *« OR »*,*« EITHER- OR »*, et *« NEITHER–NOR »*.

**Inconvénients :**

L'utilisation de l'approche basée sur un corpus pour identifier tous les mots d'opinions, n'est pas aussi efficace que l'approche par dictionnaire, car il est difficile de préparer un énorme corpus pour couvrir tous les mots.

- **Approche basée sur le dictionnaire**

Elle consiste à utiliser des dictionnaires de synonymes et antonymes existants, afin de déterminer l'orientation sémantique de nouveaux mots. L'idée est d'abord, de collecter

manuellement un petit ensemble de mots d'opinion avec une orientations connues en tant que graine, puis de développer et d'élargir cet ensemble en recherchant dans les dictionnaires connus tels que SentiWordNet, AFINN, Loughran McDonald, CNRC-Hashtag, General Inquirer et Lexicon3, ainsi que d'autres ressources moins connues telles que HowNet, Wodnet. Les mots récemment trouvés sont ajoutés à la liste de germes.Puis, une nouvelle itération est lancée. Leprocessus itératif s'arrête quand il n'y a plus de nouveaux mots trouvés **[29].**En fin de compte, une vérification manuelle est effectuée pour supprimer les erreurs.

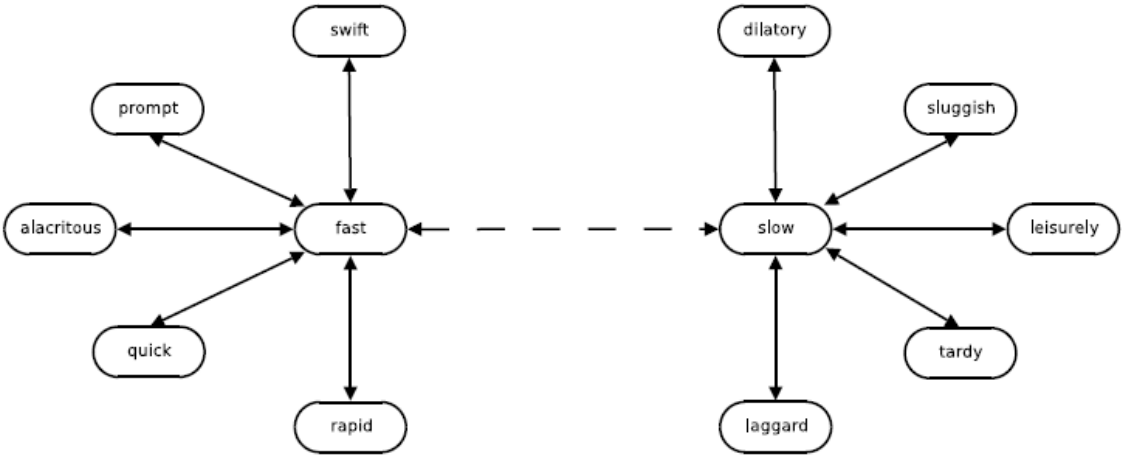


**Figure 2.3** : Exemple d'arbre de synonymes et antonymes présents dans WordNet (flèche pleine =
synonymes, flèche hachurée = antonymes)[30]

**Inconvénients :**

- L'inconvénient majeur de cette approche est qu'elle est incapable de trouver des mots d'opinion spécifiques au domaine et au contexte.

## 2.2.1.2. Quelques ressources lexicales

Nous citons dans ce qui suit les ressources les plus utilisées pour la classification des sentiments.

**WordNet**

WordNet est un dictionnaire anglais développé à l'Université de Princeton disponibleen format électronique [31]. Son organisation est différente de la plupart des dictionnaires.

Il a été développé dans un laboratoire de sciences cognitives comme un modèle de représentation humaine de connaissances lexicales, ce modèle est basé sur des études psycholinguistiques. L'initiative de EuroWordNet a développé des dictionnaires semblables pour d'autres langues, mais celui pour l'anglais reste toujours le plus complet.

Les concepts synonymes sont misent dans des synsets (contraction pour synonym sets), par exemple, {car, auto, automobile, machine, motocar}. Chaque synset a une définition, comme celles des dictionnaires courants.

**SentiWordNet**

SentiWordNet est une ressource lexicale pour l'extraction d'opinionspécifique à l'opinion mining qui se base sur le dictionnaire lexical WordNet.

Dans ce dictionnaire, les adjectifs, les noms, les verbes et autres normes grammaticales sont regroupés dans des ensembles de synonymes appelés synsets. Trois scores entre l'intervalle -1 et 1 sont associés par SentiWordNet à des synsets pour identifier le sentiment du texte donné comme étant positif, négatif ou neutre.[30][32]

**AFINN**

Le lexique AFINN est une liste de termes anglais manuellement classés pour la valence avec un nombre entier compris entre -5 (négatif) et +5 (positif) [33]. Il est distribué sous la licence Open Database License (ODbL) v1.0.

Le lexique AFINN est peut-être l'un des lexiques les plus simples et les plus populaires pouvant être largement utilisés pour l'analyse des sentiments. La version actuelle du lexique contient plus de 3 300 mots avec un score de polarité associé à chaque mot.

### 2.2.1.3. Phase 2 :Classification des textes grâce aux lexiques

Une fois que les mots porteurs d'opinion sont répertoriés dans les lexiques, la dernière étape, consiste à calculer l'opinion globale du texte à base des sentiments positifs, négatifs retenus pour chaque phrase qui sont mise en relation pour donner un sentiment global du texte entier.

Une valeur appelée orientation sémantique a été créée pour démontrer la polarité des mots. Elle varie en deux grandeurs : positive et négative et peut avoir différents niveaux d'intensité. Il existe plusieurs méthodes de calcul de l'orientation sémantique d'un document.La méthode la plus simple consiste à attribuer à un document un score d'opinion égale au nombre total

de score des mots qui contiennent une opinion présente dans le document ce qui se résume à une fréquence des termes au niveau du document.

Dans les approches basées sur les dictionnaires, l'orientation sémantique des termes présents dans la phrase opiniâtre est obtenue, en comparant le nombre de termes à orientation sémantique positive avec le nombre de termes à orientation sémantique négative. Si le nombre de positifs est plus grand, la polarité de l'opinion est considérée positive et vice-versa.

## 2.2.2. Les approches basée sur l'apprentissage automatique

Le but de cette approche est de créer de la connaissance de manière automatique à partir de données brutes et générer de nouveaux modèles pour la classification des documents non structurée. Cette connaissance (modèle) peut alors être exploitée pour prendre des décisions.

On distingue deux types de méthodes d'apprentissage, qui sont l'apprentissage supervisé et l'apprentissage non supervisé**.**

- **Apprentissage supervisé(Classification)**

Ce type de méthodes consiste, à construire un modèle basé sur un jeu d'apprentissage et des labels (nom des catégories ou des classes) et à l'utiliser pour classer de nouvelles données.[34] Le but de cette méthode est que l'algorithme puisse « apprendre» en comparant sa sortie réelle avec les sorties «enseignées» pour trouver des erreurs et modifier le modèle en conséquence. Ensuite, ces modèles sont utilisés pour prédire les valeurs d'étiquettes sur des données non étiquetées supplémentaires

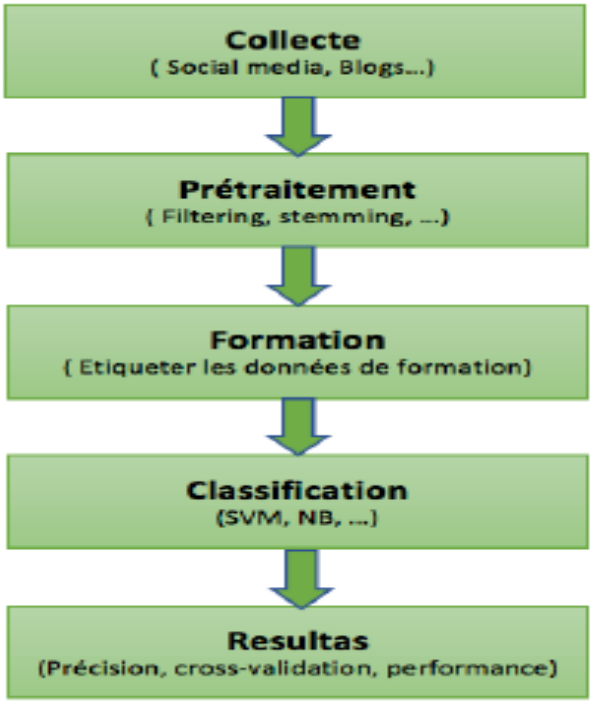


**Figure 2.4 :** Etapes d'une approche supervisée.[34]

- **Apprentissage non supervisé**

L'apprentissage non supervisé (en anglais *Clustering*) vise à construire des groupes (clusters) d'objets similaires, à partir d'un ensemble hétérogène d'objets. [35]

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

L'objectif de l'apprentissage non supervisé peut être aussi simple que de découvrir des modèles cachés dans un ensemble de données, mais il peut aussi avoir un objectif d'apprentissage des caractéristiques, qui permet aux algorithmes de découvrir automatiquement les représentations nécessaires pour classer les données brutes

### 2.2.2.1. Les principaux classifieurs :

Les méthodes de classification ont pour but, d'identifier les classes auxquelles appartiennent des objets à partir de certains paramètres descriptifs.

Les méthodes utilisées pour la classification sont nombreuses, citons : la méthode des SVM, Naïve bayes, les Réseaux de Neurones, les arbres de décision, etc. Nous présentons brièvement dans la suite certains des classifieurs les plus fréquemment utilisés en analyse d'opinion :

- **Classifieur naïve Bayésienne (NB)** :

C'est une méthode de classification statistique la plus utilisée en analyse d'opinion, elle repose sur l'hypothèse que les attributs sont fortement indépendants. Elle calcule la probabilité d'une classe en fonction de la distribution des mots dans les documents de la classe, et pour se faire elle se base sur le théorème de Bayes. Ce théorème est donné par :

$$P(x\ /y)\ = p\ (y/x)\ \times p(x)\ /p(y)\ ;$$

Où $P\ (x/y)$ est la probabilité conditionnelle d'un événement $x$ sachant qu'un autre événement $y$ de probabilité non nulle s'est réalisé.

- **Machine à vecteurs de support (SVM)** :

Les machines à vecteurs de support appelés aussi séparateurs à vaste marge sont des techniques d'apprentissage supervisées qui exploitent les concepts relatifs à la théorie de l'apprentissage statistique et à la théorie des bornes de Vapnik et Chervonenkis .La justification intuitive de cette méthode d'apprentissage est la suivante : si l'échantillon d'apprentissage est linéairement séparable, il semble naturel de séparer parfaitement les éléments des deux classes de telle sorte qu'ils soient le plus loin possibles de la frontière choisie. En effet, l'idée principale de SVM est de reconsidérer le problème dans un espace de dimension supérieure, éventuellement de dimension infinie. Dans ce nouvel espace, il est alors probable qu'il existe un hyperplan séparateur linéaire. Si c'est le cas, les SVM cherchent parmi l'infinité des hyperplans séparateurs celui qui maximise la marge entre les classes.

Cette technique est une méthode de classification à deux classes qui tente de séparer les exemples positifs des exemples négatifs dans l'ensemble des exemples. La méthode cherche alors l'hyperplan qui sépare les exemples positifs des exemples négatifs, en garantissant que la marge entre le plus proche des positifs et des négatifs soit maximale. [36]

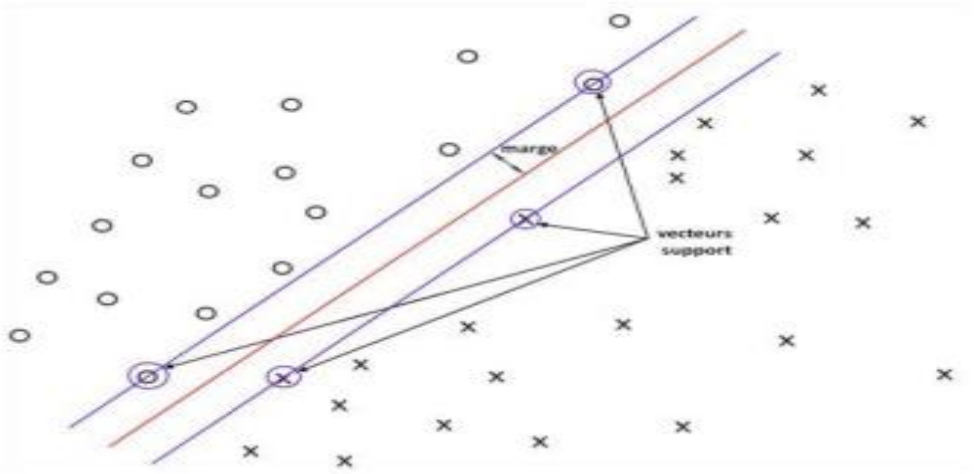


**Figure 2.5** : Principe du Séparateur à Vaste Marge (SVM). [36]

- **L'arbre de décision**

Les arbres de décision ont été utilisés pour permettre la catégorisation des documents dans un certain nombre de classes prédéfinies. L'apprentissage des probabilités d'attribution d'un texte à une classe est réalisé sur des textes étiquetés manuellement. L'arbre de décision généré, sur l'ensemble des documents du corpus d'apprentissage, permet de décider à quelle classe appartient chaque nouveau document du corpus de test. Chaque feuille de l'arbre contient la probabilité d'appartenance à l'une ou l'autre des classes. Suivant les réponses aux questions posées au document à classer, celui-ci est « dirigé » vers telle ou telle feuille de l'arbre. Le document est alors attribué à la classe de plus forte probabilité. [37]

- **Réseaux de neurones**

Ce classificateur est constitué d'un grand ensemble d'unités (ou neurones), chacune de ces unités d'entrée représentent les fréquences des termes dans le document, l'unité de sortie représente la classe ou les catégories d'intérêts, et le poids sur les bords reliant les unité reprisent les relations de dépendance.

Les entrées d'un neurone sont soit les entrées du réseau global, soit les sorties d'autres neurones. Les connexions entre les neurones qui composent le réseau décrivent la topologie du modèle. Les paramètres les plus importants de ce modèle sont les coefficients synaptiques (c'est-à-dire les poids), le seuil, et la façon de les ajuster lors de l'apprentissage. Ce sont eux qui construisent le modèle de résolution en fonction des informations données au réseau. Il faut donc trouver un mécanisme qui permet de les calculer à partir des grandeurs que l'on peut acquérir du problème. C'est le principe fondamental de l'apprentissage. [38]

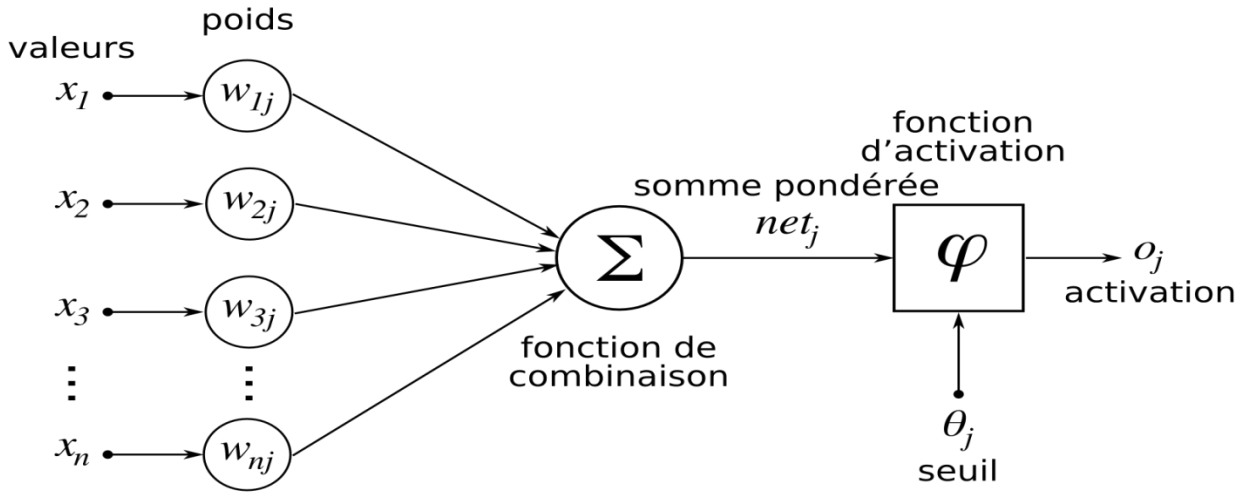


**Figure 2.6 :** structure d'un réseau de neurone artificiel. [38]

# 2.3. Les travaux sur la classification de polarité

Beaucoup de travaux importants se sont intéressés à la classification d'opinions. Nous les avons répartis en deux catégories :

## 2.3.1. Approches basées sur le lexique

De nombreux travaux ont été réalisés pour analyser et classer les sentiments en utilisant des approches basées sur le lexique. Citons notamment quelques travaux des auteurs dans ce domaine.

**Hatzivassiloglou et McKeown**[39]constituent probablement la première tentative importante de classification des mots en fonction de leur polarité. Au lieu d'Internet, ils ont utilisé le corpus du Wall Street Journal et se sont uniquement intéressés à la question de savoir si un mot était positif ou négatif. Pour cela, ils ont utilisé des coordinateurs et d'autres caractéristiques morpho-syntaxiques locales pour prédire la polarité des adjectifs, ils ont décrit une méthode d'apprentissage non supervisée pour obtenir des adjectifs orientés positivement et négativement avec une précision supérieure à 90%, et démontré que cette orientation sémantique, ou polarité, est une propriété lexicale cohérente avec un accord élevé entre les évaluateurs.

**Turney et al.** [40]ont proposé un algorithme prenant en compte un commentaire écrit et produisant une classification en tant que résultat selon une approche en trois étapes: la première étape consiste à utiliser un tagueur de partie du discours (POS)pour identifier les phrases d'un commentaire contenant des adjectifs ou des adverbes.la seconde étape consiste à estimer l'orientation sémantique (SO) de chaque phrase extraite recommandée ou non recommandée, sur la base de l'orientation sémantique moyenne des phrases extraites. Si la moyenne est positive, alors l'élément est supposé recommandé, sinon l'élément n'est pas recommandé. Dans la dernière étape, un algorithme de récupération d'informations mutuelles point par point est utilisé pour mesurer la similarité de paires de mots ou de phrases afin d'estimer l'orientation sémantique d'une phrase.

**Taboada et al.** [41] utilisent, quant à eux, un lexique pour extraire les mots véhiculant des sentiments (y compris les adjectifs, verbes, noms, et adverbes) dans un texte en combinant l'utilisation de corpus et de dictionnaires en calculant l'orientation sémantique (SO-CAL). Cette dernière s'appuie sur des dictionnaires de mots annotés avec leur orientation sémantique et intègre l'intensité et la négation.

**Almas et Ahmad** [42]ont utilisé une approche symbolique basée sur les grammaires locales pour analyser le sentiment dans les articles de presse financiers, leur étude était spécifique àce domaine

**Zhang et Liu**[40][43]ont montré que les syntagmes nominaux et le substantif peuvent aussi renfermer des opinions. Ils comptent le nombre de phrases positives et négatives pour chaque fonctionnalité du produit en utilisant le lexique d'opinion préparé par [44]. Leur approche permet d'atteindre une précision moyenne d'environ 0,44.

**Hu et Liu** [45]utilisent seulement les adjectifs pour la détection des opinions. Ils construisent manuellement une liste d'adjectifs qu'ils utilisent pour prédire l'orientation de la phrase et utilisent WordNet pour alimenter la liste par les synonymes et les antonymes des adjectifs dont on connaît la polarité. Ils assignent 1 à chaque adjectif positif et 0 à chaque adjectif négatif.

**Ohana et al.** [43][46] ont proposé une technique de classification de polarité en utilisant des fonctionnalités construites à partir de la base de données SentiWordNet qui contient des

scores de polarité des termes. Leur approche consiste à compter les scores des termes positifs et négatifs pour déterminer l'orientation sémantique. Ils ont mis en œuvre un algorithme de détection de la négation pour ajuster les scores SentiWordNet.

## 2.3.2. Approches basées sur l'apprentissage automatique

De nombreux travaux ont été réalisés pour analyser et classer les sentiments en utilisant des approches basées sur le lexique. Citons notamment quelques travaux des auteurs dans ce domaine.

**Pang et Lee** [47]ont été les premiers à avoir exploité l'approche d'apprentissage automatique pour l'analyse d'opinion sur des critiques cinématographique. Ils ont considéré le problème de la classification des documents non pas par sujet, mais par le sentiment général. En utilisant les critiques de films comme données, ils ont constaté que les techniques classiques d'apprentissage automatique surpassent largement les niveaux de base produits par d'autres approches. Cependant, parmi les trois méthodes d'apprentissage automatique ayant été utilisées (Naïve Bayes, classification d'entropie maximale et machines à vecteurs de support SVM), on remarque que seule SVM a donné les meilleurs résultats avec une précision de 82,9 % en utilisant les Unigrams. Les auteurs sont également montrés que l'utilisation des techniques d'analyse du discours et de coréférence a aidé à l'amélioration de la précision.

**Rushedi Saleh et al.** [48]ont réalisé plusieurs expériences avec différentes caractéristiques en analyse d'opinion en utilisant des machines à vecteurs de support (SVM) pour tester différents domaines d'ensembles de données et en utilisant plusieurs schémas de pondération. Pour cela il sont basés sur trois corpus toutefois, Les meilleurs taux de précision obtenus pour la classification étaient respectivement pour chaque corpus de 85,35%, 73,25% et 91,51%.

**Tan et al.** [49]ont introduit une approche automatique pour déduire les règles de modèle de polarité afin de détecter la polarité de sentiment au niveau de la phrase, ils prennent en compte les effets des relations plus complexes trouvées entre les mots dans la classification de polarité de sentiment.ils ont utiliser des règles séquentielles de classe (CSR) pour apprendre automatiquement les modèles de dépendance typés et comparer les performances de ce dernier à une méthode heuristique, ces modèles associes en outre les dépendances au relations grammaticales, telle que le sujet ou l'objet indirect.

Le tableau suivant représente certains travaux réalisés dans cette catégorie

| L'article | méthodes | Caractéristique | Corpus | Type de classification |
|---|---|---|---|---|
| [50] | NaiveBayes,SVM , entropie maximum | Unigrame, bigrame, POS | Corpus de tweets | Supervisée |
| [51] | Naive Bayes, entropie maximum, SVM | Emoticons | Corpus de tweets | Supervisée |
| [52] | NaiveBayes,SVM | Emoticons | Corpus de tweets | Supervisée |
| [53] | Naive Bayes | Emoticons, POS , N-grame | Corpus de tweets | Supervisée |
| [54] | Naive Bayes, Maximum d'entropie, SVM | N-grame, bigrame, Sac de mots | critiques cinématographique | Supervisée |

**Tableau** 2.2: synthèse des travaux basée sur l'apprentissage automatique

# 2.4. Méthodes d'évaluation de la performance de classification

Il existe plusieurs indicateurs permettant de mesurer les performances d'un système de classification, En classification de polarité on retrouve principalement :

- Les Matrices de confusion
- La F1-mesure
- La Validation croisée

## 2.4.1. Matrice de confusion

Une matrice de confusion ou tableau de contingence est l'un des outils les plus couramment utilisés pour évaluer la précision et la qualité d'un système de classification.

Une matrice de confusion est applicable à la fois aux classifications binaire et multi-classes, elle est obtenue en comparant les données classées avec des données de référence. Elle est construite en mettant respectivement sur les lignes et sur les colonnes les données de référence et la classification.

Prenons l'exemple d'une classification binaire, permettant de prédire deux classes notées classe positive et négative. Pour mesurer les performances de cette classification, il est d'usage de distinguer 4 types d'éléments classés pour la classe voulue :

- Vrai positif (VP): prédiction positive correcte
- Faux positif (FP): prédiction positive incorrecte
- Vrai négatif (VN): prédiction négative correcte
- Faux négatif (FN): prédiction négative incorrecte

Ces informations peuvent être rassemblées et visualisées sous forme de tableau dans une matrice de confusion. Dans le cas d'une classification binaire, on obtient :

| | **Classe prédite** | |
|---|---|---|
| **Classe réelle** | Positive | Négative |
| Positive | VP | FN |
| Négative | FP | VN |

**Tableau 2.3 :** la matrice de confusion.

En particulier, si la matrice de confusion est diagonale, la classification est parfaite.
Notons que la matrice de confusion est aussi généralisable lorsqu'il y a k > 2 classes à prédire.

## 2.4.2. Indicateurs de f-score (f-mesure) :

Il est possible de calculer plusieurs indicateurs résumant la matrice de confusion. Par exemple si nous souhaitons rendre compte de la qualité de la prédiction sur la classe positive, on définit:

- **Précision** :la proportion d'éléments bien classés pour une classe donnée :

$$\text{Précision} = VP\ /VP+FP$$

$$\text{Précision}_i = \frac{\text{Nombre de documents correctement attribués à la classe i}}{\text{Nombre de documents attribués à la classe i}}$$

- **Rappel** : la Proportion d'éléments bien classés par rapport au nombre d'éléments de la classe à prédire :

$$Rappel = VP / VP+FN$$

$$Rappel_i = \frac{\text{Nombre de documents correctement attribués à la classe i}}{\text{Nombre de documents appartenant à la classe i}}$$

- **F-mesure** : c'est la mesure la plus utilisée en classification d'opinion. Son évaluation permet de tenir compte à la fois de la précision ainsi que du rappel. Il se mesure à l'aide de la formule suivante :

$$F\text{-}mesure_i = \frac{2 \times (Précision_i \times Rappel_i)}{Précision_i + Rappel_i}$$

### 2.4.3. La validation croisée :

La validation croisée désigne le processus qui permet de tester la précision prédictive d'un modèle dans un échantillon test (parfois aussi appelé échantillon de validation croisée) par rapport à la précision prédictive de l'échantillon d'apprentissage.

La version la plus simple de la validation croisée consiste à entrainer l'approche sur un premier sous-ensemble de données. Ensuite, comparer les indicateurs de performance en appliquant l'algorithme sur le premier et le second ensemble de données.

Si les indicateurs trouvés pour l'ensemble d'entraînement sont bien supérieurs à ceux trouvés sur l'ensemble de test, alors il faut retoucher le modèle pour améliorer les performances sur l'ensemble de test.

## 2.5. Conclusion

Dans ce chapitre nous avons abordé un état de l'art sur la classification d'opinion, nous avons présenté les deux grandes approches de classification de polarité : l'approche basée sur le lexique et l'approche basée sur l'apprentissage automatique, puis nous avons énuméré quelques travaux en relation avec les deux approches. Enfin nous avons cité les mesures de performance utilisées pour l'évaluation de classification de polarité.

Dans le prochain chapitre, nous allons implémenter une approche qui se porte sur le domaine de classifications d'opinion sur tweets extraites à partir Twitter en s'appuyant sur l'approche lexicale basée sur les dictionnaires, et en s'inspirants des différents travaux réalisés sur cette approche. Tout d'abord, on récupère des tweets à partir d'un corpus de tweets initialement

annotés, puis, nous allons effectuer une série de prétraitement linguistique afin de faciliter la tâche de classification. Enfin, nous allons déterminer la classe positive, négative ou encore neutre pour chaque tweet

# Chapitre 03 :

# Implémentation et réalisation

# 3.1. Introduction

Les réseaux sociaux sont devenus une partie intégrante de notre quotidien. Leur objectif principal est de faciliter la communication entre individus. Le service de microblogging, offre la possibilité de publier des messages courts et permet aux réseaux sociaux de prendre une nouvelle dimension (publication par les utilisateurs de leurs pensées, sentiments ou avis d'une manière courte).

Ouvert à tout le monde, Twitter est devenu, ces dernières années, le numéro un dans le domaine du microblog. Il peut être vu, comme un indicateur pour connaître les réactions de ses utilisateurs sur plusieurs sujets sociaux, politiques, économiques, etc. Par conséquent, on peut l'utiliser pour extraire les émotions, les sentiments ou les opinions de ses utilisateurs.

Dans notre travail nous avons utilisé Twitter pour extraire des tweets sur lesquels on applique l'approche lexicale afin de déduire leurs polarités.

Ce chapitre permet de discuter les différentes étapes de notre approche proposée pour la classification des tweets à travers une approche linguistique basée sur les dictionnaires. Ensuite nous allons définir les outils utilisés pour la réalisation de la partie pratique de nos travaux avec une présentation générale des résultats obtenus en discutant les différentes comparaisons appliquer entre les différentes techniques utilisées et proposées durant notre travail.

# 3.2. Présentation

Twitter est un réseau de microblogging géré par l'entreprise Twitter Inc. Il permet à un utilisateur d'envoyer gratuitement de brefs messages, appelés tweets, sur internet, par messagerie instantanée ou par SMS. Ces messages sont limités à 280 caractères (140 caractères jusqu'en novembre 2017).

Twitter a été créé le 21mars2006 par Jack Dorsey, Evan Williams, Biz Stone et Noah Glass, et lancé en juillet de la même année. Le service est rapidement devenu populaire, jusqu'à réunir plus de 500 millions d'utilisateurs dans le monde fin février 2012. En 2019, Twitter compte 326 millions d'utilisateurs actifs par mois avec 500 millions de tweets envoyés par jour et est disponible en plus de 40 langues. [55]

**Figure 3.1** : Logo de Twitter.

### 3.2.1. Caractéristiques d'un tweet

Nous présentons rapidement dans cette sous-partie les principales caractéristiques d'un tweet

- **Longueur :**

Twitter a doublé la longueur maximale d'un message posté sur celui-ci, passant d'une longueur de 140 caractères par tweets à 280 caractères, mais malgré cette amélioration la longueur de tweet reste très courte contrairement à celle utilisée dans d'autres corpus pour la classification de sentiments (comme les critiques de films).

On ne peut donc pas voir de longues publications. Par conséquent, chaque utilisateur doit choisir ses mots soigneusement et doit utiliser des abréviations pour les mots longs ou des initiales pour les mots composés.

- **Disponibilité des données et modèles du langage :**

Les sujets abordés sur Twitter sont très divers et l'API Twitter permet de récolter des millions de messages. En effet, le nombre de tweets postés chaque jour est immense. Les utilisateurs peuvent poster des messages depuis n'importe quel lieu et avec différents appareils. Il est à noter qu'un tweet peut contenir des fautes d'orthographe liées à l'utilisation de Smartphones et à la limitation de caractères. De plus, le registre de langue utilisé peut être familier.

## 3.3. Présentation du système proposé

Notre travail consiste en l'analyse et classification de sentiments à partir de données textuelles. Ces données sont, en majorité, bruités et non structurés, elles sont extraites à partir de Twitter et organisées sous forme de corpus.

# Chapitre03 : implémentation et réalisation

Les processus composant notre système sont présentés dans la figure 3.2., représentant le processus général de la méthodologie du système. Ce processus est réparti en deux phases importantes :

- Phase 01 : prétraitement et préparation des données.
- Phase 02 : classification de polarité

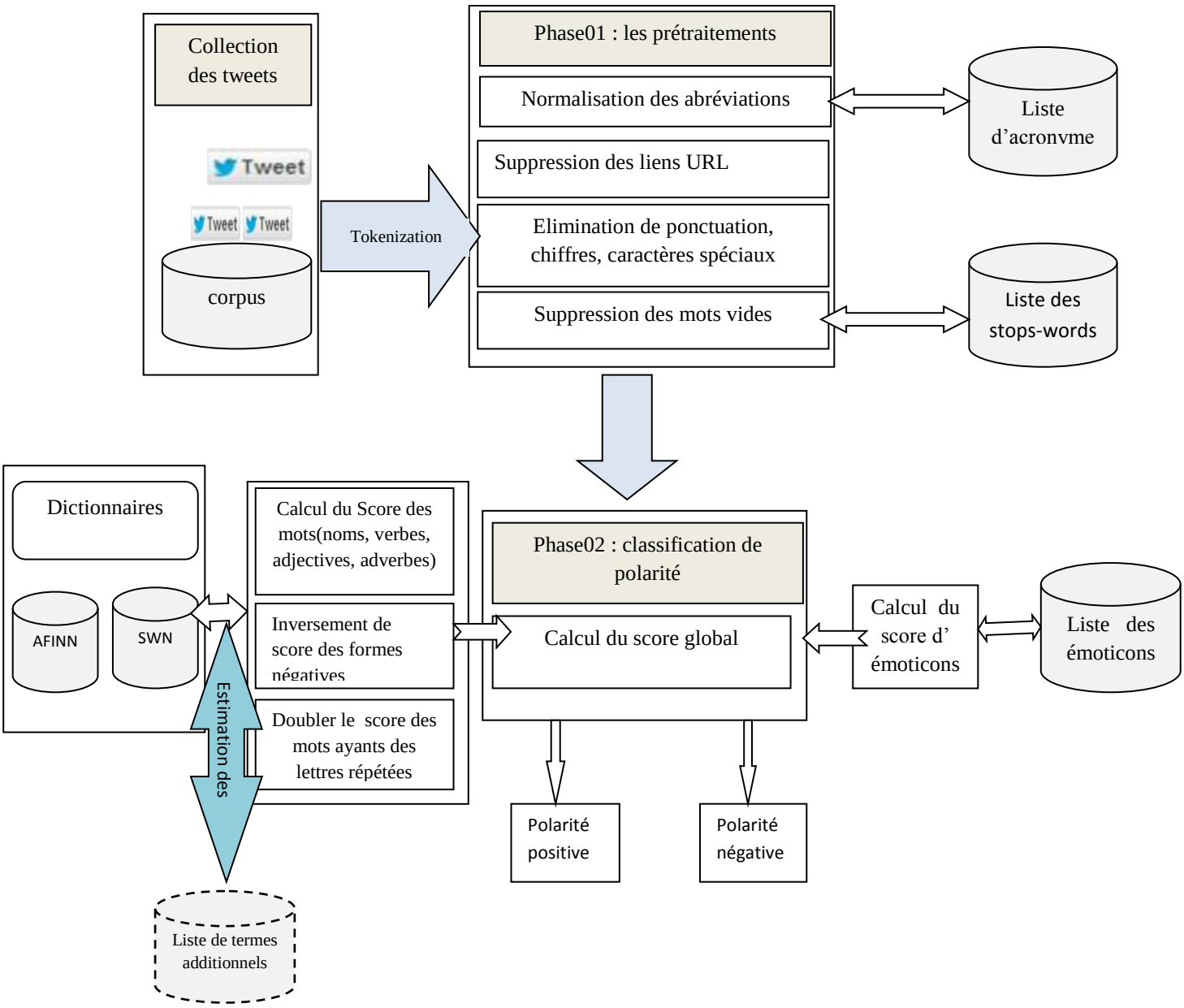


**Figure 3.2 :** Processus général de la méthodologie suivie pour la classification des tweets.

# 3.4. Environnements de travail

Nous présentons dans cette section, l'environnement matériel et logiciel en détaillant les différents outils utilisé pour la réalisation de ce travail.

## 3.4.1. Environnement matériel

Afin de mener notre expérimentation et évaluation, nous avons utilisé un PC marque HP avec un Système d'exploitation Windows7 64bits, équipé d'un processeur Intel(R) Pentium(R) cadencé par une horloge d'une fréquence de 2.16GHZ, avec 2,00 GO Octets de RAM, un disque dur d'une capacité de 200 Giga Octets.

## 3.4.2. Environnement logiciel

- **Les bibliothèques JAVA utilisées**

Ci-dessus une description des bibliothèques utilisées dans le cadre de notre travail :

- **StanfordCoreNLP**

StanfordCoreNLPest une bibliothèque d'analyse du langage naturel dont le code est écrit en Java et sous licence GNU General Public License (v3 ou ultérieure).StanfordCoreNLP intègre tous les outils de NLP(Naturel Langage Preprocessing), y compris l'étiqueteur de partie de discours POS(Part Of Speech), le programme de reconnaissance d'entité nommée (NER), l'analyseur, le système de résolution des problèmes de coréférence et les outils d'analyse des sentiments, et fournit des fichiers modèles pour l'analyse de l'anglais.

- **ConfusionMatrix**

Pour l'évaluation de résultats de l'approche, nous avons utilisé la classe ConfusionMatrix à partir de la bibliothèque com.github.confusionmatrix, cette classe permet de construire la matrice de confusion. Elle permet aussi de déterminer automatiquement les indicateurs de performance précision, rappel et F-score afin d'évaluer l'exactitude d'une classification.

# 3.5. Phase 01 : prétraitement et préparation des donnée

## 3.5.1. Corpus utilisés

**Sentiment140 :**

Pour notre étude, Nous avons utilisé le corpus d'opinions introduit par Go et al. [56], cet ensemble se présente sous forme d'un fichier d'extension (.csv) de taille 83.95 MB contenant p1600000 tweets extraits à l'aide de l'API Twitter. Cet ensemble comporte deux classes d'opinionspositive et négative) et chaque entrée de l'ensemble est structurée comme suit :

**target**: contient la polarité du tweet (0 = négative, 4 = positive).

**ids:** l'identifiant du tweet (2087).

**date:** la date du tweet (Sat May 16 23:58:44 UTC 2009).

**flag:** la requête (lyx). S'il n'y a pas de requête, alors cette valeur est NO_QUERY.

**utilisateur:** le nom de l'utilisateur qui a posté le tweet (Lyxis cool).

**text:** le texte du tweet (Lyx est cool).

| # 0<br>target | # 1467810369<br>id | Mon Apr 06 22:19:45 PD<br>date | NO_QUERY<br>flag | _TheSpecialOne_<br>user | @switchfoot http://twit<br>text |
|---|---|---|---|---|---|
| 0 | 1467810672 | Mon Apr 06 22:19:49 PDT 2009 | NO_QUERY | scotthamilton | is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah! |
| 0 | 1467810917 | Mon Apr 06 22:19:53 PDT 2009 | NO_QUERY | mattycus | @Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds |
| 0 | 1467811184 | Mon Apr 06 22:19:57 PDT 2009 | NO_QUERY | ElleCTF | my whole body feels itchy and like its on fire |

**Figure 3.3 :** description de corpus sentiemnt140

## 3.5.2. Prétraitements

Les performances d'une quelconque analyse sur Twitter découlent souvent de la nature des messages qui y sont publiés. En effet, les tweets sont totalement différents des documents classiques tels que les articles des journaux, les discours officiels, les pages web, etc. Parmi les particularités liées au tweet, on retrouve :

– La difficulté d'analyse due au style d'écriture informel. En effet, les utilisateurs utilisent un langage qui n'est pas formel, et un mélange entre le jargon, abréviations et plusieurs langues dans le même tweet.
– Les Tweets sont pleins d'erreurs d'orthographe, d'erreurs lexicales et d'erreurs syntaxiques.
– L'existence des liens et des identifiants compliquent l'opération d'analyse.

Ainsi, un prétraitement et un nettoyage de ces textes sont indispensables. Parmi les prétraitements opérés dans le cadre de ce travail, on retrouve principalement la suppression des liens, car dans notre contexte, un lien au sein d'un tweet n'a pas un poids sémantique pour le sujet suivi. On opère également une suppression les identifiants et des hashtags, vu leur inutilité par rapport à la polarité.

Un autre type de traitement consiste à nettoyer les mots qui contiennent uniquement un seul caractère, c'est une particularité que l'on retrouve souvent sur les réseaux sociaux et qui fait référence à des expressions humoristiques.

Nous présenterons dans ce qui suit les prétraitements suivi dans ce travail afin de rendre les tweets le plus proche possible d'un langage formel tout en réduisant l'espace d'attributs.

- **Normalisation des tweets :**

La tâche de la normalisation consiste à réécrire le texte dans la langue standard ou proche de la langue standard. Notre but n'est pas de faire la correction orthographique et syntaxique, mais de réécrire le texte en se basant sur les erreurs lexicales fréquentes dans les médias sociaux. Ainsi, pour assurer une correspondance entre le langage informel des abréviations et d'acronymes, nous avons créé une liste de référence, faisant office de lexique pour la suite de notre travail.

```
ATEOTD: At The End of the Day
ATB: All the best
ATM: At the moment
AWOL: Absent Without Official Leave
B2B: Business to Business
B2C: Business to Customer
B4: Before
BBIAB: Be back in a bit
BBQ: Barbecue
BBL: Be back later
BBS: Be back soon
BCNU: Be seein' you
BFF: Best Friends Forever
BFN: Bye For Now
BOFH: Bastard operator from hell
BRB: Be right back
BSOD: Blue Screen of Death
BTDT: Been there done that
```

**Figure 3.4:**La liste des Acronymes.

Le prétraitement opéré pour cette normalisation a été réalisés à l'aide d'expressions régulières appliquées pour filtrer notre texte des caractères indésirables et des sous-chaines qu'on souhaite retirer de façon à garder uniquement des données non bruitées. Nous créons, ainsi, un motif de traitement à l'aide des expressions régulières pour nettoyer notre ensemble de données textuel.Ainsi, nous avons :

- Remplacer les noms d'utilisateur par l'expression :<username>.

Exemple : @alydesignsiwas out most of the daysodidn'tgetmuchdone
Sera remplacépar :
<username>i was out most of the daysodidn'tgetmuchdone.

- Remplacementdes Hashtag(#)par l'expression <hashtag>

Exmple : @markhardy1974 Me too #itm.
Sera remplacé par :
@markhardy1974 Me too<hashtag>.

- SupprimerdesURLs

  Exemple :Crazy wind today = no birding http://ff.im/1XTTi.

  Sera remplacépar :

  Crazy wind today = no birding

- Suppression et élimination des ponctuations et caractères spéciaux

Les utilisateurs utilisent dans leurs tweets beaucoup de ponctuations et caractères spéciaux qui n'ont aucune importance pour la classification.

Exemple : [,.!_ ; :-«$^<» ;>|{#β€£≥ ?≠∞\¤%*+^<& ].

Et enfin, une liste de formes contractées a été utilisée pour rendre certains mots ou groupe de mot mieux manipulable pour la suite de l'analyse.

| Formes contractées | Formes non contractées |
|---|---|
| that's | that is |
| thats | that is |
| isnt | is not |
| cant | can not |
| isn't | is not |
| can't | can not |
| havn't | have not |
| havnt | have not |
| aren't | are not |
| arnt | are not |
| wouldn't | would not |
| wouldnt | would not |
| souldn't | should not |
| shouldnt | should not |
| Havn't | have not |
| havnt | have not |

**Figure 3.5:** liste de formes contractées.

## ➢ La suppression des mots vide

Cette étape consiste à éliminer tous les mots non significatifs à partir d'une liste de mots vides (Stop Word List). Pour chaque mot reconnu, nous le comparons avec un des éléments dans

l'anti-dictionnaire (figure 3.6). Si un mot en fait partie, il sera systématiquement ignoré et non pris en considération pour la suite du processus.

```
a
about
above
after
again
against
all
am
an
and
any
are
as
at
be
because
been
before
being
below
herself
him
himself
his
how
how's
i
i'd
i'll
i'm
i've
if
in
into
is
it
it's
its
itself
let's
of
off
on
once
only
or
other
ought
our
ours
ourselves
out
over
own
same
she
she'd
she'll
she's
should
```

**Figure 3.6** : liste des stops-words.

Le tableau suivant donne un aperçu de quelques tweets avant et après les opérations de prétraitements :

| Tweets avant Prétraitements | Tweets après Prétraitements |
|---|---|
| @switchfoothttp://twitpic.com/2y1zl - that's a bummer. You shoulda got David Carr of Third Day to do it. ;D" | bummer have got davidthird day ;d |
| is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah!" | upset can not update facebook texting might cry result school today also blah |
| spring break in plain city... it's snowing " | spring break plain city snowing |
| @Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds" | dived many times ball managed save rest go bounds |
| my whole body feels itchy and like its on fire " | whole body feels itchy like fire |

**Tableau 3.1 :** Tweets avant et après prétraitements.

## ➢ Etiquetage grammatical des mots

Afin d'étiqueter certains termes du corpus, nous avons utilisé le tagger « Part of Speech tagger » disponible dans la bibliothèque StanfordCoreNLP. Ce tagger permet d'étiqueter tous les mots du corpus selon leur fonction grammaticale. Nous gardons ensuite l'ensemble des adjectifs obtenus.

| Mot | Etiquette |
|---|---|
| As | RB |
| Often | RB |
| As | IN |
| Possible | NN |
| Drinking | VBG |
| Water | NN |
| Is | VBZ |
| GOOD | JJ |
| FOR | PRP$ |
| Internal | JJ |
| Organs | NN |

**Tableau 3.2 :** Exemple d'étiquettes grammaticales utilisées par StanfordCoreNLP tagger

La figure ci-dessous, indique la signification de chaque étiquette :


**Figure 3.7:** Signification des étiquettes grammaticales. [41]

- **Utilisation des Bi-grammes de mots avec catégorie et prise en compte de la négation au sein du texte :**

Les bi-grammes sont des séquences de deux items adjacents que l'on retrouve dans un texte. Les bi-grammes de mots peuvent être utiles pour créer des expressions permettant de faciliter la classification, ces bi-grammes permettent notamment de s'affranchir des problèmes de négation. A titre d'exemple, l'expression *« not great »*contribuera au score positif d'une phrase si elle avait été traitée uniquement avec une représentation en uni-gramme de mots. En effet, en utilisant cette dernière, l'expression *« not great »* sera séparée en deux uni-grammes de mots indépendants l'un de l'autre. En conséquence la polarité de l'expression *« not great »* sera biaisée vu que l'inversion de polarité ne sera pas prise en compte lors de l'analyse.

Ceci est d'autant plus intéressant lorsqu'on opère une analyse de la notion de négation. Cette dernière joue un rôle central en analyse des sentiments au sein du texte, car elle change souvent l'orientation sémantique d'une phrase. A ce propos, nous avons ignoré les opérateurs de négation au sein de la liste de Stop Words, afin de détecter les formes négatives.

Dans le cadre de ce travail, nous nous focalisons sur l'aspect négation des adjectifs. Pour ce faire, l'Api StanfordCoreNLP a été utilisé pour l'étiquetage grammatical. Puis, en s'aidant des

opérateurs de négation et d'une représentation en bi-grammes, nous avons généré une routine permettant de prendre en compte l'aspect d'inversion de polarité. Tout cela est détaillé dans la section suivante.

# 3.6. Phase 02 : classification de polarité

L'analyse de polarité est une étape cruciale dans tout système de détection d'opinions. Cette tâche consiste à déterminer la classe (positive, négative) d'un document. Dans le cas de Twitter, chaque tweet est composé d'un ensemble de termes (verbes, adjectifs, noms, adverbes…).Puis, pour chacun de ces termes, on recherche une orientation sémantique en procédant à l'utilisation de dictionnaire à savoir, SWN, AFINN. Cependant, l'analyse de polarité grâce aux lexiques se confronte à certaines problématiques. On retrouve par exemple, le problème lié aux termes dont la polarité est neutre. En effet, cette neutralité s'avère être pénalisante lorsque la tâche de classification d'opinion s'opère sur un axe binaire. Les changements de polarité dus aux formes négatives mais également aux formes d'intensification (répétions de lettres au sein d'un mot), sont aussi problématiques pour l'analyse de polarité.

Dans cette section, nous exposons l'ensemble des étapes suivis pour l'élaboration de notre système d'analyse. Nous y présentons notre contribution liée à l'enrichissement du lexique SWN.

## 3.6.1 Calcul du score des mots

De manière générale, chaque tweet est composé d'un ensemble de mots dont la plupart sont réparties au sein de lexiques. Ainsi, pour rechercher l'orientation sémantique de chaque mot on a recours divers lexiques. L'un des plus complets et des plus utilisés en anglais, est sans conteste, SentiWordNet (SWN) avec 100 000 termes répertoriés. On retrouve également AFINN, contenant 2477 mots et expressions. Ce dernier est une liste de mots en anglais dont la valeur des évaluations est comprise entre (- 05) pour un mot fortement négatif et (+ 05) pour un terme fortement positif

Dans le cadre de ce travail, nous avons opté pour une classification en prenant en compte le score dans l'intervalle [-5 ; -1](resp. [1 ; 5]), pour les termes négatifs (resp. positifs).

### 3.6.1.1. Contribution à l'enrichissement de lexique

Dans la majorité des approches basées sur les lexiques, la classification d'opinions s'effectue en calculant le score de polarité global (positif et négatif) de l'ensemble des mots contenus au sein du texte à analyser et ce, en se basant sur le score des termes extraits à partir d'un lexique spécialisé ou général. A titre d'exemple, si un texte contient plus de mots positifs que de mots négatifs, il sera classé comme étant de polarité positive. On retrouve plusieurs ressources lexicales disponibles pour ce type d'analyse à savoir, AFINN, OpinionFinder, SentiWordNet et SenticNet. Cependant, AFINN et SentiWordNet (SWN) restent parmi les lexiques d'opinions les plus utilisés en analyse d'opinions, en raison de leur énorme vocabulaire. Par exemple, SentiWordNet contient plus de 100 000 synsets. Cependant, plus de 90% d'entre eux sont classés comme étant des termes neutres. Cette neutralité s'avère être pénalisante lorsqu'une tâche de classification d'opinion opère sur un axe binaire dont la polarité est soit positive ou négative.

Aussi, tel que mentionnée ci-dessus, une analyse d'opinion basée sur Twitter se heurte à plusieurs difficultés en raison du langage informel utilisé. En effet, les tweets sont généralement courts et contiennent beaucoup d'argot, d'émoticônes, d'abréviations ou de mots mal orthographiés. Ainsi, la plupart des mots utilisés dans certains tweets, ne figurent pas forcément au sein des lexiques utilisés. De plus, ces mots non répertoriés où à polarité neutre peuvent éventuellement comporter une opinion implicite, notamment dans certains domaines particuliers. Il pourrait donc être préférable de prendre en considération ces termes lors du calcul du score final de polarité.

Justement, notre modeste contribution dans le cadre de ce mémoire consiste à réadapter la polarité de ces mots dits neutre, mais également attribuer aux termes manquant au sein des lexiques, un score de polarité adaptée en fonction du corpus utilisé. Une des manières possible d'opérer cette amélioration, consiste à estimer les scores de polarité des termes manquants en fonction de la polarité des tweets les incluant déjà au sein du corpus. A titre d'exemple, supposons que le terme anglais neutre *« gift »* apparaisse plus souvent dans les tweets positifs que dans les tweets négatifs. Ainsi, ce terme pourrait être reconsidéré comme étant un terme positif. En revanche, lorsqu'un terme ne figurant pas au sein des lexiques et qui apparaît plus souvent dans les tweets négatifs que dans les tweets neutres ou positifs, il peut être alors considéré comme un terme négatif.

# Chapitre03 : implémentation et réalisation

Dans le cadre de ce travail, nous proposant de construire une liste de termes additionnels au lexique utilisé. Cette liste couvrira les scores de polarité estimés de l'ensemble des termes neutres et manquants. Les scores attribués à ces derniers, sont estimés en supposant que les polarités des termes coïncidents avec la polarité des tweets inclus au sein du corpus utilisé (corpus annotés). Ceci semble raisonnable du fait que les textes utilisés (tweets) sont limités d'un point de vue contenu. En d'autres termes, si un mot apparaît fréquemment dans un nombre important de tweets (ensemble réduit de phrases) positifs (resp. négatifs), sa polarité risque fort probablement d'être à connotation positive (resp. négative).

La création de la liste additionnelle est réalisée en deux temps. Dans la première étape on associe, aux termes obtenus lors de la phase de prétraitement, un score SentiWordNet global (Score_terme) à l'aide de l'équation (1). Il est à noter que, SentiWordNet est utilisé ici simplement pour vérifier si le mot est à connotation neutre [Score_terme (wi) = 0] ou inexistant. Ces mots seront ensuite traités pour l'estimation de la polarité Score_estimation à l'aide de l'équation (2).

$$Score_{terme}(w_i) = Score\_terme_{pos}(w_i) - Score\_terme_{nég}(w_i) \qquad (1)$$

Où :

$Score\_terme_{pos}(w_i)$ : correspond au score positif du terme $(w_i)$ au sein de SentiWordNet

$Score\_terme_{nég}(w_i)$ : correspond au score négatif du terme $(w_i)$ au sein de SentiWordNet

$$Score\_estimation(w_i) = \begin{cases} Score_{pos}(w_i) \\ si Score_{pos}(w_i) > Score_{nég}(w_i) \\ \\ (-1)x Score_{nég}(w_i) \\ si Score_{pos}(w_i) < Score_{nég}(w_i) \end{cases} \qquad (2)$$

Où :

$$Score_{pos}(w_i) = \frac{P(pos|w_i)}{P(pos)}$$

$$Score_{nég}(w_i) = \frac{P(nég|w_i)}{P(nég)}$$

$$P(pos|w_i) = \frac{Fréquencedutermew_i danslestweetspositifs}{Fréquencedutermedansl'ensembleducorpus}$$

$$P(nég|w_i) = \frac{Fréquencedutermew_i danslestweetsnégatifs}{Fréquencedutermedansl'ensembleducorpus}$$

$$P(pos) = \frac{Nbretweetspositifs}{Totaltweetsauseinducorpus}$$

$$P(nég) = \frac{Nbre\ tweets\ négatifs}{Total\ tweets\ au\ sein\ du\ corpus}$$

Dans un second temps, l'intervalle de valeur obtenu via l'application de l'équation (2) sera réadapté afin de respecter les contraintes liées au format au sein des lexiques SWN et AFINN. En effet, les scores dans SentiWordNet étant compris entre -1 et 1, on se doit ainsi convertir les scores estimés $Score\_estimation(w_i)$ dans le même intervalle. Dans ce cas, nous utilisons une fonction sigmoïde bipolaire vu qu'elle renvoie une valeur comprise entre -1 et 1. Ceci permettra de respecter, en même temps, les contraintes de SWN et d'AFINN. La formule de conversion est indiquée dans l'équation (3). Cependant, le score obtenu doit également respecter une contrainte de fiabilité liée à la fréquence des termes.

Par exemple, lorsque la fréquence du terme est trop faible ou bien la marge entre les scores positifs et négatifs d'un terme au sein de SentiWordNet soit trop minime. Par conséquent, cela pourrait engendrer un biais pour la classification. Ainsi, pour remédier à cette contrainte, nous avons proposé d'utiliser deux seuils minimums (Seuil1 et Seuil2) pour chacune des contraintes ci-dessus. Par conséquent, Seuil1 correspondra au nombre minimum de termes dans l'ensemble de données et Seuil2 correspond à la différence minimale entre les scores de polarité positifs et négatifs du terme en question.

$$Score\_estimation(w_i) = sigmoid(Score\_estimation(w_i)) \qquad (3)$$

### 3.6.1.2. Inversement du score des formes négatives

Pour améliorer la classification, nous analysons morpho-syntaxiquement nos tweets afin de détecter les formes de négations. Nous recherchons les opérateurs de négation associés directement aux adjectifs (voir le tableau.3.3), puis nous réadaptons en conséquence le degré d'orientation sémantique de ces adjectifs.

| | |
|---|---|
| **Opérateurs de négation** | No,not,rather,couldn't,wasn't wouldn't , ,didn't, Shouldn't, weren't, don't, doesn't, haven't,hasn't,won't,hadn't,never,none,nobody,nothing, Neither,nor,isn't,can't,mustn't,mighthn't, Hardly,less, littel,rarely,scarcely,seldom |

**Tableau 3.3** : Liste des opérateurs de négation utilisée

**Exemple 01:**


**Figure 3.8 :** Exemple d'utilisation de la négation

Prenons l'exemple de la figure précédente, si nous essayons de déduire l'orientation sémantique de l'adjectif « *bad* » en tenant compte uniquement d'un traitement en sacs de mots, nous remarquons que celui-ci sera répertorié comme étant, à connotation négative. Ceci influera en conséquence sur l'orientation sémantique de la phrase, en lui attribuant une orientation négative. Cependant, en tenant compte des opérateurs impactant le sens d'un adjectif, nous pouvons facilement intervertir la polarité de celui-ci, évitant ainsi toute erreur d'affectation de polarité.

**Exemple 02:**

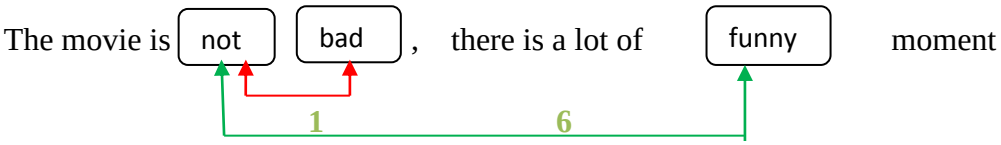


**Figure 3.11 :** Exemple d'utilisation de la forme négative

Dans la figure 3.11, l'opérateur de négation **«** *not* » est adjacent à l'adjectif « *bad* » (i.e. distance égale à un).Ainsi, en tenant compte de l'exemple précédent, on remarque que son orientation sémantique sera inversée. Cependant nous considérons que l'inversion de polarité n'affecte que les adjectifs adjacents aux opérateurs de négation. Dans l'exemple, l'adjectif « *funny* » se positionnant à une distance de 6 mots, ne verra pas sa polarité inversée en présence du not. Il conservera ainsi son orientation sémantique initiale.

Cette étape d'inversement du score est réalisée en deux phases :

- **Phase 01 : extraction des adjectifs porteurs d'opinion**

L'objectif de cette première phase est de tester si l'adjectif est présent dans chaque tweet de corpus. Pour cela, à partir de corpus collecté, on utilise l'étiqueteur grammatical POS qui attribue des catégories grammaticales pour chaque mot. Dans notre cas, on se contente uniquement à des termes portants l'étiquette (_JJ) correspondante aux adjectives.

- **Phase 02 : Détection de négation**

Afin de détecter les formes négatives spécifiques, nous cherchons à détecter la présence d'opérateurs aux abords des adjectifs précédemment répertoriés, tout en respectant l'aspect de distance et en vérifiant la juxtaposition des adjectifs aux opérateurs de négation.

### 3.6.1.3. Prise en compte de l'intensification de polarité grâce aux répétitions de lettres.

Nous avons vu plus haut, que les adjectifs sont fortement impactés par le contexte dans lequel ils apparaissent. Les collocatifs de la classe des adjectifs dits « intensifieurs » c'est-à-dire dont la fonction et non de modifier le sens global de la catégorie grammaticale (Adjectif) mais seulement approuver le trait sémantique de celui-ci, ont largement été exploités en analyse d'opinions.

Cependant, on remarque une pratique rédactionnelle propre aux réseaux sociaux et qui consiste en l'utilisation des lettres répétées. En effet, les internautes utilisent dans leur lexique des mots comportant des lettres répétées pour affirmer et assurer le sens de leur opinion. Ces mots permettent justement d'augmenter l'intensité de l'opinion. Il existe plusieurs travaux ayant exploité l'aspect de répétition. Cependant, très peu ont réellement étudié l'impact qu'à ce type de traitement sur l'analyse de polarité.

Dans le cadre de ce travail, nous proposons une normalisation des termes, en se basant sur un retranchement successif des lettres répétées jusqu'à ce que, celui-ci sera répertorié au sein des lexiques utilisés. Notre apport porte sur l'utilisation de la forme canonique, afin d'élaguer toutes les éventuelles erreurs d'orthographe pouvant s'immiscer lors du processus de suppression des lettres répétés. Un fois le terme en bonne et due forme repéré, celui-ci verra sa polarité doublée, afin de prendre en compte l'aspect d'intensification.

## 3.6.2. Calcul du score des émoticons

La contrainte de taille liée aux tweets encourage l'utilisation des émoticônes pour exprimer les opinions et les sentiments. Un émoticône est une courte figuration symbolique d'une émotion, d'un état d'esprit ou d'une intensité, utilisée dans un discours écrit. Dans le cadre de l'analyse de texte, ces émoticônes changent souvent la polarité de toute la phrase.

Pour notre part, Nous avons construit notre propre liste d'émoticônes réparties en positifs et négatifs (voir les figures suivantes 3.12 et 3.13). La prise en compte des émoticônes intervient principalement lors du calcul du score de polarité final du tweet. Si ce dernier comporte un émoticon apparaissant dans l'une des listes illustrée ci-dessous, nous procédons alors à l'incrémentation/décrémentation du score final en fonction du type d'émoticône utilisé.

```
:?)
:-]
:]
:-3
:3
:->
:>
8-)
8)
:-}
:}
:c)
:^)
=]
=]
:-))
;?)
;)
*-)
*)
;?]
;]
```

**Figure 3.12 :** Liste des émoticons positifs

```
:?(
:(
:?c
:c
:?<
:<
:?[
:[
:-||
>:[
:{
:{
>:(
:'?(
:'(
:?/
:/
:?.
>:\
>:/
:\
=/
```

**Figure 3.13 :** liste des émoticons négatifs

## 3.6.3. Calcule du score globale

Dans le cadre de ce travail, l'orientation positive ou négative d'un tweet sera calculée sur la base traitement ci-dessus. Ainsi, pour chaque tweet on comptabilisera l'ensemble des scores de polarité des termes et émoticône le constituant et ce, en tenant compte de l'ensemble des spécificités citées ci-dessus. La polarité des termes a été récupérée sur la base du lexique AFINN. La polarité et la classe de chaque tweet seront déduites en fonction de ce score globale.

# 3.7. Expérimentations et résultats obtenus

Dans cette section, nous présentons les différents tests que nous avons réalisés pour évaluer et approuver nos propositions, puis discutons nos résultats obtenus.

## 3.7.1. Les mesures utilisées

Les mesures de performances utilisées sont : la précision (P), le rappel(R), F-mesure(Fm) dont le calcul a été réalisé grâce à l'Api ConfusionMatrix sur la base de notre matrice de confusion. Cette dernière est construite après obtention des polarités de chaque tweet et en le comparant par rapport aux polarités réelles du corpus étiqueté, selon l'algorithme suivant :

---

**Algorithme de construction de la matrice de confusion**

---

**Input :**
**-C_tweet** : classe de tweet prédite par notre approche ($C_p$_tweet : positive, $C_n$_tweet : négative).
**-C** : Classe réelle du tweet, fournie par le corpus des tweets ($C_P$ : positive,$C_n$:négative).
**Output :**
VP, FP, FN, VN
**Début**
    Répéter pour tous les tweets du corpus :
        ***Si*** $C_p$_tweet est correctement attribué à $C_p$
        ***Alors***VP=VP+1 ;
        ***Sinon*** FN=FN+1 ;
        ***Si*** $C_n$_tweet est correctement attribué à $C_n$
        ***Alors***VN=VN+1 ;
        ***Sinon***FP=FP+1
Fin de la boucle Répéter
**FIN**

---

Nous en déduisons une série d'indicateurs utilisées pour évaluer et tester les résultats de classification.

### Evaluation 01:

Dans cette première évaluation, nous allons effectuer plusieurs tests sur quelques phrases (figure 3.12), afin d'étudier l'impact d'émojis, des lettres répétées ainsi que le

traitement des négatives, cela en utilisant les deux dictionnaires (SWN et AFINN). Les résultats obtenus seront discutés durant les tests suivants.

| **Les phrases** | Polarité réelle |
|---|---|
| It was great to see all of you again | positive |
| It wasn't great to to see them | négative |
| she is sad | négative |
| she is not sad | positive |
| I am :) | positive |
| i am :[ | négative |
| it is sweeeeeeeeeeeeeet | positive |
| it was very badddddddddddd | négative |
| on its own , it's not very interesting . as a remake , it's a pale imitation . | négative |
| routine and rather silly . | Negative |
| it is messy , uncouth , incomprehensible , vicious and absurd . | Negative |
| everything is not fine . | Negative |
| a warm , funny , engaging film . | Positive |
| It was a fanny movie | positive |
| It wasnt fanny | Negative |

**Tableau 3 .4 :** les phrases utilisées pour l'évaluation 01.

- **Test 01:** traitement des formes négatives +traitement d'émojis + prise en compte de l'intensification de polarité grâce aux répétitions de lettres

Dans le premier test, nous allons étudier l'impact d'émojis, des lettres répétées ainsi que le traitement de la négation des adjectives à la fois. Cela en utilisant les deux dictionnaires (SWN et AFINN). Les résultats obtenus sont représentés dans la figure 3.14.

```
le nombre totale d'emojis dans le corpus est 2
le nombre totale d'adjectifs negatifs  dans le corpus est 4
--------------Resultats avec AFINN----------------
 ↓gold\pred→         neg         pos
         neg           8           1
         pos           0           6

P/R/Fm: neg=1.000/0.889/0.941 pos=0.857/1.000/0.923
{neg=1.0, pos=0.8571428571428571}
{neg=0.8888888888888888, pos=1.0}
Macro F-measure: 0.932, (CI at .95: 0.127), micro F-measure (acc): 0.933


--------------Resultats avec SWN--------------


 ↓gold\pred→         neg         pos
         neg           6           3
         pos           0           6

P/R/Fm: neg=1.000/0.667/0.800 pos=0.667/1.000/0.800
{neg=1.0, pos=0.6666666666666666}
{neg=0.6666666666666666, pos=1.0}
Macro F-measure: 0.800, (CI at .95: 0.202), micro F-measure (acc): 0.800
```

**Figure 3.14** : Capture sur les résultats obtenus dans l'évaluation 01 (T1)

- **Discussions :**

D'après la figure 3.14, en combinat le traitement de négation, traitement d'émojis et traitement des lettres répétés, il est clair que les résultats sont conformes à ce qu'on avait attendu. D'après les résultats obtenus par le biais des deux dictionnaires, nous remarquons que l'évaluation réalisée à l'aide de lexique AFINN présente les meilleurs résultats. A partir de cette constatation, nous avons pris la décision naïve, de poursuivre le reste des expérimentations à l'aide de celui-ci.

- **Test 02 :** calcul du score sans prise en compte le score d'émojis ni le traitement des formes négatives ni l'intensification de polarité des lettres répétées :

Dans ce deuxième test, nous avons implémenté seulement les étapes de prétraitement (Tokenisation, normalisation et suppression des mots vides..).

Dans ce test, le score d'un tweet est égal à la somme des scores de l'ensemble des mots le constituant. Les résultats obtenus sont illustrés dans la figure ci-dessous :

```
le nombre totale d'emojis dans le corpus est 2
le nombre totale d'adjectifs negatifs  dans le corpus est 4
--------------Resultats avec AFINN----------------
 ↓gold\pred→         neg         pos
         neg           3           6
         pos           1           5

P/R/Fm: neg=0.750/0.333/0.462 pos=0.455/0.833/0.588
{neg=0.75, pos=0.45454545454545453}
{neg=0.3333333333333333, pos=0.8333333333333334}
```

**Figure 3.15**:Capture sur les résultats obtenus dans l'évaluation 01 (T2).

- **Test 03 :** calcul du score des mots + traitement de négation des adjectives

Dans ce troisième test, nous avons pris en considération le traitement de négation comme une étape supplémentaire pour les étapes de prétraitement. Typiquement pour voir si cela affectait sur le taux de précision, rappel, F-mesure.

Dans ce cas, le score global d'un tweet est égal au score total de tous les mots qui le constitue qui seront récupérés à partir de dictionnaire en inversant le score des adjectifs précédés par un opérateur de négation. Les résultats aboutis sont montrés dans la figure 3.15.

```
le nombre totale d'emojis dans le corpus est 2
le nombre totale d'adjectifs negatifs  dans le corpus est 4
--------------Resultats avec AFINN----------------
 ↓gold\pred→         neg         pos
         neg           6           3
         pos           0           6

P/R/Fm: neg=1.000/0.667/0.800 pos=0.667/1.000/0.800
{neg=1.0, pos=0.6666666666666666}
{neg=0.6666666666666666, pos=1.0}
```

**Figure 3.16**: Capture sur les résultats obtenus dans l'évaluation 01(T3)

- **Test 04 :** calcul du score des termes + traitement d'émoticons

Ce quatrième test, consiste à prendre les émoticônes en considération comme une étape supplémentaire pour les étapes de prétraitement, en les prenant en compte dans le score total des tweets. Et cela, Pour voir ce qu'importe le traitement des émojis sur le taux de précision, rappel, F-mesure.

Dans ce test, le score global est égal au score total de tous les mots qui le constitue qui seront récupérés à partir de dictionnaire + le score des émojis.

$$\text{Score total} = \sum \text{score\_mots} + \sum \text{score\_émojis}$$

Les résultats obtenus sont illustrés dans la figure 3.16.

```
le nombre totale d'emojis dans le corpus est 2
le nombre totale d'adjectifs negatifs  dans le corpus est 4
--------------Resultats avec AFINN----------------
 ↓gold\pred→         neg         pos
         neg           4           5
         pos           1           5

P/R/Fm: neg=0.800/0.444/0.571 pos=0.500/0.833/0.625
{neg=0.8, pos=0.5}
{neg=0.4444444444444444, pos=0.8333333333333334}
```

**Figure 3.17**: Capture sur les résultats obtenus dans l'évaluation 01 (T4).

- **Test 05 :** calcul du score + une prise en compte de l'intensification de polarité grâce aux répétitions de lettres.

Les résultats obtenus sont illustrés dans la figure 3.17.

```
le nombre totale d'emojis dans le corpus est 2
le nombre totale d'adjectifs negatifs  dans le corpus est 4
--------------Resultats avec AFINN----------------
 ↓gold\pred→         neg         pos
         neg           4           5
         pos           1           5

P/R/Fm: neg=0.800/0.444/0.571 pos=0.500/0.833/0.625
{neg=0.8, pos=0.5}
{neg=0.4444444444444444, pos=0.8333333333333334}
```

**Figure 3.18**: Capture sur les résultats obtenus dans l'évaluation 01 (T5).

Le tableau ci-dessous résumera toutes les phases de tests que nous avons effectués jusqu'à présent.

| | | Teste 01 | Teste 02 | Teste 03 | Teste 04 | Teste 05 |
|---|---|---|---|---|---|---|
| **Positive** | **Précision** | 0.857 | 0.455 | 0.667 | 0.500 | 0.500 |
| | **Rappel** | 1.000 | 0.833 | 1.000 | 0.833 | 0.833 |
| | **F-mesure** | 0.923 | 0.588 | 0.800 | 0.625 | 0.625 |
| **Négative** | **Précision** | 1.000 | 0.750 | 1.000 | 0.800 | 0.800 |
| | **Rappel** | 0.889 | 0.333 | 0.667 | 0.444 | 0.444 |
| | **F-mesure** | 0.941 | 0.462 | 0.800 | 0.571 | 0.571 |

**Tableau 3.5:** tableau récapitulatif des tests d'évaluation 01

## ➢ Discussions :

Selon les résultats obtenus, nous avons constaté que les résultats de la précision, rappel et F-mesure ont été nettement améliorés. Toutefois, ces résultats restent en dessous de ceux escomptés, cela s'explique certainement par le faible pourcentage de juxtaposition des adverbes associés aux adjectifs. En effet, la non prise en compte des polarités liées à certains termes aurait certainement pu être impacté ces résultats. Ceci est notamment au facteur de neutralité des termes.

## Evaluation 02 :

Dans cette deuxième évaluation, nous allons implémenter l'approche sur le corpus Sentiment140 composé de 1600000 tweets. Nous avons effectué deux tests pour étudier l'impact de traitement de la négation des adjectifs, la prise en compte de l'intensification de polarité des lettres répétées, ainsi l'utilisation d'émojis dans ce corpus.

- **Test 01:** calcul de score sans considération de score d'émojis ni le traitement de la négation des adjectifs ni l'intensification de polarité des lettres répétées.

Les résultats sont illustrés dans la figure 3.

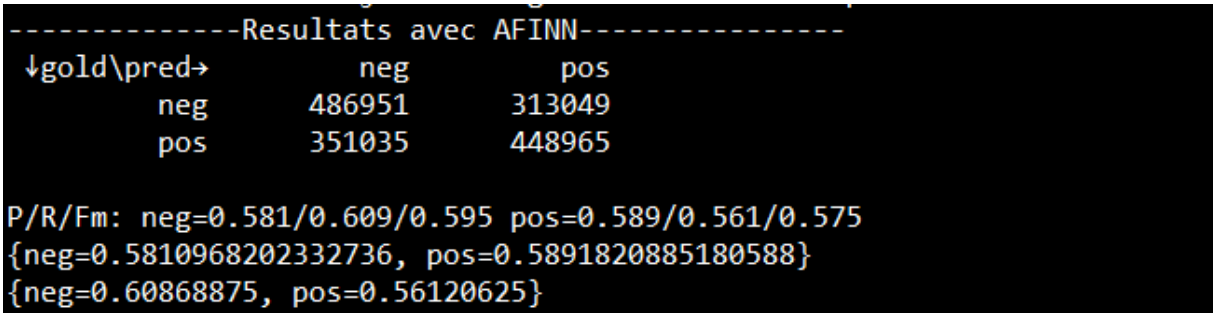


**Figure 3.19 :** une capture sur les résultats obtenus dans l'évaluation 02(T1)

- **Test 02:** tous les traitements sont effectués

Les résultats sont illustrés dans la figure 3.19

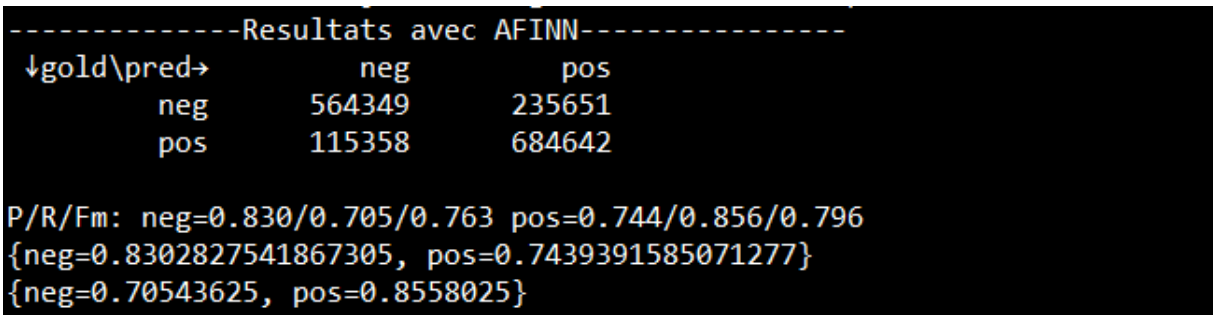


**Figure 3.20:**une capture sur les résultats obtenus dans l'évaluation 02 (T2)

Le tableau suivant résume les résultats obtenus dans le test01 et les résultats finaux obtenus dans le test 02

| AFINN | | Teste 01 | Teste 02 |
|---|---|---|---|
| Positive | Précision | 0.589 | 0.744 |
| | Rappel | 0.561 | 0.856 |
| | F-mesure | 0.575 | 0.796 |
| | Précision | 0.581 | 0.830 |

| | | | |
|---|---|---|---|
| Négative | Rappel | 0.609 | 0.705 |
| | F-mesure | 0.595 | 0.763 |

**Tableau 3.6**: tableau récapitulatif des tests d'évaluation 02 (T1) et (T2)

### ➢ Discussions :

Cette approche a bien classé 684642 tweets parmi 800000 tweets positifs et 564349 tweets parmi 800000 tweets négatifs.

D'après les résultats obtenus, on constate que l'approche a pu améliorer la classification de la classe positive et négative, elle a pu augmenter le taux de la précision, le rappel ainsi la F-mesure.

### Evaluation 03 :

Cette dernière expérience a été menée, afin d'étudier l'apport de notre proposition d'enrichissement de lexique. Tel que décrit dans la section 3.6.1.1, deux seuils (Seuil 1 et Seuil 2) ont été proposés afin d'élaguer, respectivement les termes à faible apparition au sein du corpus et ceux dont la marge, entre le score positif et négatif, est minimale.

Le choix de la valeur optimale pour le Seuil 1 a été établi sur la base des travaux présentés dans[60]. Les auteurs ont réalisé une série de tests dans laquelle ils étudient l'impact de divers facteurs (fréquentielles, syntaxiques et heuristiques) sur le taux de classification de l'ensemble de données SemEval 2013. L'étude montre que dans 5 cas sur 7, les termes ayant une fréquence d'apparition inférieure à 15% (relativement au terme le plus fréquent dans le corpus) présentent peu d'impact sur l'amélioration du taux de classification. Dans notre cas, cette valeur correspond au nombre de 5 apparitions du terme (Seuil1 = 5).

Tandis que, la valeur du Seuil 2 a été fixée en fonction de la marge maximale (=0,08) entre le score positif et négatif des termes neutres enregistrés sous SentiWordNet. Pour ce faire, nous avons simplement divisé cette marge maximale par 2 (Seuil2 = 0,04).

Au final, nous obtenons une liste additionnelle comprenant 312 nouveaux termes répartis en

deux catégories à savoir, positif et négatif (Tableau 3.8). Nous avons mené cette expérience sur le corpus utilisé dans l'évaluation 02°, celui-ci comporte au total 1600000 tweets. Le tableau 3.7 montre les résultats de la F-mesure obtenus en utilisant dans un premier temps, les lexiques AFINN et SWN séparément, puis en additionnant aux deux lexiques respectifs, la liste des termes générés par notre approche. La comparaison a été faite en fonction de l'évaluation 02°, vu qu'elles présentent relativement la même configuration (Négation, répétions de lettres et émoticônes). Ainsi, les résultats montrent une nette amélioration du taux de la F-mesure, celle-ci a été augmentée comparativement aux résultats de l'évaluation 02°. Certes, cette amélioration est conséquence, mais n'est malheureusement pas suffisante au vu des résultats obtenus aux travers d'autres travaux basés sur les lexiques.

| | | AFINN | AFINN+ Liste additionnelle | SWN | SWN+Liste additionnelle |
|---|---|---|---|---|---|
| classe Positive | Précision | 0.744 | 0.766 | 0.547 | 0.699 |
| | Rappel | 0.856 | 0.988 | 0.402 | 0.688 |
| | F-mesure | 0.796 | 0.862 | 0.463 | 0.693 |
| classe Négative | Précision | 0.830 | 0.894 | 0.527 | 0.624 |
| | Rappel | 0.705 | 0.853 | 0.639 | 0.691 |
| | F-mesure | 0.763 | 0.873 | 0.577 | 0.647 |

**Tableau 3.7** : Taux de classification (F-mesure) de l'évaluation 3.

| Termes positif | Score | Termes négatif | Score estimé |
|---|---|---|---|
| Thank | 0.8637 | Cancel | -0.9864 |
| Fun | 0.8628 | Damn | -0.9864 |
| Luck | 0.8560 | Niggas | -0.9690 |
| Yay | 0.8341 | Russia | -0.9039 |
| Yeah | 0.7999 | Cry | -0.9168 |
| Help | 0.8698 | Wait | -0.7434 |

**Tableau 3.8** : Exemple de termes avec score estimé

# 3.8. Conclusion

Dans ce chapitre, nous avons présenté l'essentiel de notre travail qui consiste à implémenter une approche basée sur le lexique pour l'analyse des sentiments dans les réseaux sociaux (Twitter).

Nous avons utilisé une méthode basée sur un dictionnaire, qui est une méthode simple et efficace utilisée depuis des générations, pour analyser les sentiments et les opinions d'un certain tweet. Nous avons aussi implémenté quelques techniques et certaines des principales étapes de prétraitement de notre système pour une meilleure précision et une meilleure performance.

# Conclusion générale

L’analyse des sentiments et l'opinion Mining est un domaine émergeant, durant ces dernières années. Plusieurs recherches s'intéressent à la manière d'analyser la masse de textes opiniâtres disponible sur le web, en particulier ceux issues des plateformes de micro-blogging.

Notre travail s'intègre justement dans cet axe de recherche. Nous avons proposé une approche de classification de polarité d’opinions binaires, l’étude a été réalisée sur des données issues de Twitter.

Notre système est loin d'être performant. Néanmoins, les résultats obtenus sont relativement satisfaisants. En guise de perspective, nous proposons l’utilisation d’approches hybrides pour étudier l’impact des approches statiques sur les propositions faites. Par exemple, le score de termes obtenu grâce à la proposition décrite dans la section 3.6.1.1 pourrait éventuellement être intégrés aux caractéristiques de classification SVM en tant qu'information supplémentaire et ce, de différentes manières. Soit en tant que pondération, soit en tant que facteur de normalisation.

Nous proposons ainsi d’élargir l’utilisation de ce travail vers d’autres objectifs comme l’analyse des tendances et l’extraction de connaissances à partir des réseaux sociaux

# Bibliographie


- [1] [URL :https://www.larousse.fr/dictionnaires/francais/opinion/56197]. Consulté le 24/03/2019.

- [2] Liu, B. «**Sentiment analysis and opinion mining. Synthesis lectures on human language technologies** » (p.1-167), (2012).

- [3] Riloff, E., Wiebe, J., & Wilson, T. «**Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL***» Association for Computational Linguistics, (p. 25-32), (2003, May).

- [4]Kim, S. M., &Hovy, E. «**Extracting opinions, opinion holders, and topics expressed in online news media text In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*** » (p.1-8) , Association for Computational Linguistics , (2006, July).

- [5]Ekman, P. « **An argument for basic emotions »**. *Cognition & emotion*, *6*(3-4), 169-200 , (1992).

- [6]Choi, Y., Cardie, C., Riloff, E., &Patwardhan, S. « **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*** » (p. 355-362) , Association for Computational Linguistics , (2005, October).

- [7] Gadek, G. « ***Détection d'opinions, d'acteurs-clés et de communautés thématiques dans les médias sociaux*** » , Doctoral dissertation , (2018).

- [8] [URL :https://www.adweek.com/digital/37705-sentiment-analysis/ ], consulté le 31/03/2019.

- [9] Maurel, Sigrid, Paolo Curtoni, and Luca Dini. «**L'analyse des sentiments dans les forums** » , *Atelier Fouille des Données d'Opinion* , (2008).

- [10] Navigli, R. « **Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*** », *41*(2), 10 , (2009) .

- [11] [URL : http://www.thebeaconservices.com/sentiment_analysis.php ], consulté le 07/04/2019.

- [12] Damien Poirier et al. «***Approches Statistique et Linguistique Pour la Classification de Textes d'Opinion Portant sur les Films***» , [en ligne] [URL: http:// hal.archivesouvertes.fr/docs/00/46/64/12/PDF/ rnti09-poirier_et_al.pdf ] , (2010).

- [13] FaizaBelbachir ,« ***Expérimentation de fonctions pour la détection d'opinions dans les blogs***». Mémoire de master. Université de Paul Sabatier, Toulouse , (2010).

- [14] Makroum, H., Belangour, A., &Azouazi, M. « **VERS UNE APPROCHE BIG DATA POUR LA PR EDICTION DE L'IMPACT DES RESEAUX SOCIAUX SUR LE PRIX DU MARCH E BOURSIER** ». *QUINZIÈME JOURNÉE DE MATHÉMATIQUES ET APPLICATIONS JMA17*, 84.

- [15 ].K. Liu *et al* ., « **Incorporate Sentiment Analysis in ContextualAdvertising** », p. 1-8 , *TROA2008*-WWW2008, , [en ligne] [URL : http://research.yahoo.com/workshops/troa-2008/papers/submission_4.pdf] , (2008).

- [18] H. Saggion, A. Funk ,**« Extracting Opinions and Facts for Business Intelligence** », *Revue des Nouvelles Technologies de l'Information (RNTI)*, (p. 119-146), (2009).

- [17] bollen ,j. , Mao,H., and Zeng, X , «**Twitter mood predicts the stock market.j.comput.Sci.2** », p. 1-8 , (2011).

- [18]Carol Hermann ,«***Entre Web 2.0 et 3.0: opinion mining*** », (2010).

- [19] Prince, V., Kodratoff, Y., Azé, J., & Roche, M. **« Défi Fouille de Textes: reconnaissance automatique des auteurs de discours-Campagne DEFT'05 »**

- [20] Juan-Manuel , Torres-Moreno , «**Résumé automatique de documents**» (p. 56), (22/09/2011) .

- [21] Pang, B., & Lee, L. «**Opinion mining and sentiment analysis**». *Foundations and Trends® in Information Retrieval*, *2*(1–2), (p. 1-135), (2008).

- [22] Benamara F., Cesarano C., Picariello A., Reforgiato D., Subrahmanian V., «**Sentiment analysis: Adjectives and adverbs are better than adjectives alone»,**In International Conference on Weblogs and Social Media (ICWSM), Boulder, Colorado,(U.S.A), (p. 203–206), AAAI Press. (2007).

- **[23]** Das, S., & Chen, M. Yahoo! for Amazon: «**Extracting market sentiment from stock message boards**». In Proceedings of the Asia Pacific finance association annual conference (APFA) (Vol. 35), (p. 43), (2001, July).

- **[24]**Taboada, M., Brooke, J., Tofiloski, M., Voll, K., &Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.

- **[25]**Tong, S., &Koller, D. **«Support vector machine active learning with applications to text classification**». *Journal of machine learning research*, *2*(Nov), (p. 45-66), (2001).

- **[26]**Turney, P. D., & Littman, M. L. **«Measuring praise and criticism: Inference of semantic orientation from association»,** *ACM Transactions on Information Systems (TOIS)*, *21*(4), 315-346, (2003).

- **[27]**Hatzivassiloglou, V., & McKeown, K. R. ,**«Predicting the semantic orientation of adjectives».** (p. 174-181), Association for Computational Linguistics , (1997, July).

- [**28]**Kanayama, H., &Nasukawa, T. **«Fully automatic lexicon expansion for domain-oriented sentiment analysis».** (p. 355-363). Association for Computational Linguistics, (2006, July).

- **[29]** B. Liu, «**Web Data Mining: Exploring Hyperlinks,Contents, and Usage Data. Springer**», (2006).

- **[30]** Poirier, D., Fessant, F., Bothorel, C., de Neef, E. G., &Boullé, M. « Approches statistique et linguistique pour la classification de textes d'opinion portant sur les films ». *Revue des Nouvelles Technologies de l'Information*,( P.147), (2009).

- **[31]** [URL :http ://wordnet.princeton.edu/]. consulté le 19/07/2019.

- **[32]** [URL :http://www.sentiwordnet.isti.cnr.it/] consulté le 19/07/2019.

- **[33]** NIELSEN, Finn Årup.« **A new ANEW: Evaluation of a word list for sentiment analysis in microblogs ».***arXiv preprint arXiv:1103.2903*, (2011)

- **[34]** [URL :http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010] consulté le 20/07/2019.

- **[35]** Ribeiro, C. S. , ***«Inductive infeTencefoT large scaletext classification.* berlinheidelberg: springer-verlag»**,(2010).

- **[36]** Hamza Cherif , «**Classification des tracées TocoGraphique (CTG) d'un fœtus a l'aide de classifieur multiples**»(chapitre 2) , Mémoire fin d'étude , Université de Tlemcen.

- **[37]** (Bigi& all, B.Bigi, R.De-Mori, M.El-Bèze, T.Spriet, **« A fuzzy decision strategy for topic identification and dynamic selection of language models », (**2000) .

- **[38]**Metomo JOSEPH BERTRAND RAPHAËL ,**«Machine Learning : Introduction à l'apprentissageautomatique**».

- **[39]**VasileiosHatzivassiloglou and Kathleen R. McKeownPredicting**«the Semantic Orientation of Adjectives»** Department of Computer Science 450 Computer Science Building Columbia University New York, N.Y. 10027, USA {vh, kathy)©cs, columbia, edu.

- **[40]** 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**». (P. 417-424). Philadelphia, US, (2002).

- **[41]**Taboada, M., Brooke, J., Tofiloski, M., Voll, K., &Stede, M. **«Lexicon-based methods for sentiment analysis».** *Computational linguistics*, *37*(2), (p.267-307), (2011).

- **[42]**Ahmad, K., Cheng, D., & Almas, Y. ,**«Multi-lingual sentiment analysis of financial news streams».** In *1st International Workshop on Grid Technology for Financial Modeling and Simulation* (Vol. 26, p. 001). SISSA Medialab, (2007, May).

- **[43]** Zhou, C., Wu, Y. L., Chen, G., Feng, J., Liu, X. Q., Wang, C., ... & Lu, S. , **«Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer (OPTIMAL, CTONG-0802): a multicentre, open-label, randomised, phase 3 study» ,** *The lancet oncology*, *12*(8), (p.735-742), (2011).

- **[44]** Ding, H., Richard, P., Nakayama, K., Sugawara, K., Arakane, T., Sekiba, Y., ...& Wang, Z. ,**«Observation of Fermi-surface–dependent nodeless superconducting gaps in Ba0. 6K0. 4Fe2As2».** *EPL (Europhysics Letters)*, *83*(4), 47001, (2008).

- **[45]** Hu, M., & Liu, B. (2006, July). Opinion extraction and summarization on the web. In *AAAI* (Vol. 7, pp. 1621-1624).

- **[46]** Bruno Ohana and Brendan Tierney, «**Sentiment Classification of reviews using SentiWordNet. In Proceeding of IT&T Conference**». (2009)

- **[47]** Pang, B., & Lee, L. **«A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*»** (p. 271). Association for Computational Linguistics, (2004, July).

- **[48]** Saleh, M. R., Martín-Valdivia, M. T., Montejo-Ráez, A., &Ureña-López, L. A. **«Experiments with SVM to classify opinions in different domains».** *Expert Systems with Applications, (p.* 14799-14804) , (2011).

- **[49]** Tan, L. K. W., Na, J. C., Theng, Y. L., & Chang, K. «**Phrase-level sentiment polarity classification using rule-based typed dependencies and additional complex phrases consideration.** ***Journal of Computer Science and Technology*»,** *27*(3), (p.650-666), (2012).

- **[50]**TapanSahni, ChinmayChandak, Naveen Reddy Chedeti, Manish Singh «**Efficient Twitter Sentiment Classification using Subjective Distant Supervision»,** 2017 IEEE 9th International Conference on Communication Systems and Networks (COMSNETS), 548-553.

- **[51]** Alec Go, RichaBhayani, and Lei Huang. «**Twitter Sentiment Classification using Distant Supervision**»CS224N Project Report, Stanford, (p. 1-12), (2009).

- **[52]**A.Pak and P. Paroubek., «**Twitter as a Corpus for Sentiment Analysis and Opinion Mining».** In Proceedings of the Seventh Conference on International Language Resources and Evaluation, , (p.1320-1326), (2010).

- **[53]**B. Pang, L. Lee, and S. Vaithyanathan, «**Thumbs up?: sentiment classification using machine learning techniques**»presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, (2002).

- **[54]** J. Read**.** « **Using emoticons to reduce dependency in machine learning techniques for sentiment classification**». In Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics. Association for Computational Linguistics,( 2005).

- **[55]** [URL :https://fr.wikipedia.org/wiki/Twitter ],consulté le 10/08/2019.

- **[56].**Go, A., Bhayani, R., Huang, L.:**« Twitter sentiment classification using distant supervision**». CS224N Project Report, Stanford (2009).

- [60] Polanyi, L. et A. Zaenen., **« Contextual valence shifters »,**Computing Attitude and AUect in Text: Theory and Applications, Springer Netherlands, (p. 1–10 ), (2006).

# Bibliographie

# Bibliographie

# Bibliographie