République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mouloud Mammeri de Tizi-Ouzou

Mémoire

De Fin D'étude
De Master Professionnel
Spécialité : Informatique
Option : Ingénierie des systèmes d'informations
Thème

Sélection De Caractéristiques Pour La Classification De Polarité D'Opinion

---

Présenté par :
Yasmine YOUSFI
Yasmine BELLAHOUES

Devant le jury composé de :
Président : Mr HAMMACHE Areski-MCB
Examinateur : Mr RADJA Hakim-MAA
Examinateur : Mme AIT YAKOUB Zina-MAB
Encadreur : Mr SAIDANI Fayçal Redha-MAB

Année universitaire : 2018/2019

Remerciements

D'abord, nous remercions le bon ***DIEU*** *de nous avoir donné santé et courage pour réaliser ce travail.*

Nous tenons à exprimer notre profonde gratitude à notre encadreur ***Mr SAIDANI Fayçal Redha****, pour nous avoir encadré et guidé et surtout pour ses judicieux conseils qui ont contribué à alimenter notre réflexion.*

Nous remercions chaleureusement les membres de jury pour l'honneur qu'ils nous ont fait en acceptant de juger notre travail.

Nos sincères sentiments vont à nos parents qui ont sacrifié jusqu'aujourd'hui et leurs encouragements tout le long de notre parcours.

Yasmine, Yasmine.

Dédicaces

Je dédie ce modeste travail : A mes très chers parents que dieu les protègent, pour leur aide et leur soutien tout au long de mes études,

A toute ma famille, à mes chers amis,

Enfin à tous ceux qui ont contribué de près ou de loin pour la réalisation de ce travail.

Yasmine, Yasmine.

Table des matières

Table des figures

Introduction générale

L'analyse d'opinion (Opinion Mining) également connue sous le nom détection de sentiment est un domaine de recherche entre le traitement automatique du langage naturel (NLP : Natural Language Processing) et la fouille de données. Le but de ce domaine est de pouvoir identifier et extraire des opinions, sentiments et attitudes présentes dans un texte ou dans un ensemble de documents. Sa relative nouveauté explique le fait que les termes techniques utilisés pour le décrire ne soient pas toujours normés. En effet , parmi les termes communément utilisés dans la littérature, on retrouve "Fouille d'opinion", "Analyse des sentiments" et "classification de polarité". Ainsi , tout au long de ce mémoire, nous utiliserons ces appellations de manière interchangeable pour exprimer le même concept.

L'étude des opinions est un axe de recherche qui s'est popularisé avec l'émergence du Web 2.0. En effet, L'émergence des plateformes de micro-blogging et des médias sociaux, ont offert aux internautes un lieu d'expression et de partage de leurs opinions et appréciations sur divers sujets. A titre d'exemple, Twitter, avec près de 650 millions d'utilisateurs et plus de 500 millions de messages par jour, est devenu une mine d'or pour les politiques et décideurs soucieux de la réputation et de l'image que porte le public à leurs sujets, marques etc. Cet intérêt porté à l'égard de cette mine informationnelle, motive les chercheurs afin de construire des modèles et approches pour analyser cette masse d'informations opiniâtre.

Il existe deux types d'approches en analyse d'opinion. La première se base sur des lexiques d'opinions pour détecter la polarité et la subjectivité d'un texte à un niveau de granularité très fin (habituellement au niveau des mots ou syntagmes). Quant à la deuxième approche basée sur l'apprentissage supervisé, elle s'opère à des niveaux de granularité divers (phrase, document, etc). Un modèle est appris sur la base d'un corpus de documents annotés, puis un score de polarité est attribué aux textes non annotés en se basant sur le modèle précédemment appris.

L'objectif principal de notre travail est de s'initier aux techniques de l'apprentissage automatique, à travers une approche supervisée de classification de polarité d'opinions. D'abord, nous décrivons le problème de la fouille d'opinions dans son ensemble. Ensuite, nous présentons un ensemble de prétraitements nécessaires pour appliquer efficacement les techniques automatiques de fouille d'opinions dans les textes. Finalement, nous décrivons quelques techniques pour l'analyse d'opinions en nous intéressant plus particulièrement à celle s'appuyant sur l'apprentissage supervisée.

La problématique de notre travail concerne principalement la sélection de caractéristique. le but étant d'extraire un vocabulaire de termes discriminant pour l'amélioration des résultats de classification, de manière générale, notre rôle est de créer une solution et de suivre une approche combinant outils linguistiques et outils de classification afin de déterminer si un texte exprime des opinions positives ou négatives , quel que soit le sujet du texte.

Pour cela nous avons scindé le présent mémoire en trois chapitres à savoir :

- ***Chapitre 1 : Contexte de l'analyse d'opinion.***
- ***Chapitre 2 : État de l'art sur l'analyse d'opinion et la sélection de caractéristiques.***

- ***Chapitre 3 : Description et présentation de la solution.***

Chapitre 1

Contexte de l'analyse d'opinion

Introduction

Avec l'avènement du web social et collaboratif, le nombre de documents opiniâtres croit de manière exponentielle. Ceci a poussé les chercheurs de différentes communautés (fouille de données, fouille de textes, TALN [1]) à s'intéresser à l'extraction automatique d'opinions sur le web. Certaines techniques d'extraction cherchent à déterminer les caractéristiques positives et négatives d'opinions à partir d'un ensembles d'apprentissages. Des experts sont d'abord mandatés pour constituer des corpus de référence, puis des techniques de classifications sont alors utilisées pour classer automatiquement ces documents extraits à partir du web.

Dans ce chapitre nous présentons un ensemble de généralités liées à l'analyse d'opinion. Nous définirons dans un premier temps, l'analyse d'opinions et la notion d'opinion selon les composantes de Bing Liu, puis nous présentons les tâches qui lui sont liées et les difficultés rencontrés. Enfin, nous terminerons par citer les domaines d'applications ainsi que quelques exemples d'application de l'analyse d'opinion.

1.1 Définition

En informatique, l'opinion mining (aussi appelé Sentiment Analysis) est l'analyse des sentiments à partir de sources textuelles dématérialisées sur de grandes quantités de données (Big Data).Ce procédé apparait au début des années 2000 et connait un succès grandissant dû à l'abondance de données provenant des réseaux sociaux, notamment celles fournies par Twitter. L'objectif de l'opinion mining est d'analyser une grande quantité de données afin d'en déduire les différents sentiments qui y sont exprimés. Les sentiments extraits peuvent ensuite faire l'objet de statistiques sur le ressenti général d'une communauté.

Avec le web 2.0, toute page web est susceptible d'être une source de données. Cependant Twitter présente des avantages intéressants comme la brièveté des tweets (140 caractères) ainsi que sa réactivité, de plus Twitter est ouvert et les textes qui y sont soumis sont accessibles à tous grâce à un service web ce qui facilite l'exploitation des données. Cependant plusieurs études ont été faites sur d'autres sources de données telles que des paroles de chansons ou des discours présidentiels. Les réseaux sociaux restent malgré tout une cible privilégiée, car ils représentent une source de donnée riche et assurent un renouvèlement des informations en temps réel.

Le but de l'analyse d'opinions est de déterminer si le sentiment dégagé par une phrase est positif ou négatif. La principale difficulté de l'analyse réside au cœur même de l'utilisation de la langue. Le sentiment dégagé par une phrase dépend directement du contexte dans laquelle elle est utilisée, du type de langage, ainsi que de la personne qui l'a écrite... En réalité, il existe une multitude de facteurs de plus ou moins grande influence qui altèrent le sentiment suscité par un propos.[1]

1. Sigle du traitement automatique du langage naturel, ou traitement automatique de la langue naturelle, ou encore traitement automatique des langues (abr. TAL), est un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle. Il vise à créer des outils de traitement de la langue naturelle pour diverses applications. Il ne doit pas être confondu avec la linguistique informatique, qui vise à comprendre les langues au moyen d'outils informatiques.

1.2 L'opinion selon Bing Liu

Le domaine d'analyse de sentiment cherche à identifier et à analyser du contenu subjectif, où s'expriment des opinions/sentiments/jugements sur une cible (entité nommée qui peut être une marque, une personne, une organisation, un objet, etc.). On peut opérer l'analyse de sentiment à deux niveaux : au niveau du document dans son intégralité ou au niveau de la phrase. Aujourd'hui, les machines sont capables de détecter une entité nommée et son environnement proche. Elles sont capables de trouver les relations entre ces mots à partir de modèles syntaxiques et grammaticaux. Mais il est nécessaire de préciser ce qui doit être pris en considération dans l'analyse, de quoi est composée sémantiquement une opinion. C'est ce que fait Liu avec sa proposition de quintuple qui fait désormais référence.

1.2.1 L'opinion selon le « quintuple» de Liu

Bing Liu, de l'Université Illinois de Chicago, a développé tout un vocabulaire décrivant les composantes d'une opinion et donne la définition du quintuple dans "Sentiment Analysis and Subjectivity".[1]

Le quintuple de Bing Liu(o_i,f_{jk},so_{ijk},h_i,t_l) où :

— o_i est l'entité

— f_{jk} est un des aspects de o_i

— so_{ijk} est le titulaire

— h_i est son orientation

— t_l est la date

Cette représentation, certes purement descriptive, est néanmoins utile pour fixer le cahier des charges d'une application qui veut réaliser de l'analyse d'opinion car l'absence d'un de ces éléments rend l'analyse particulièrement superficielle. En revanche, l'analyse d'opinion peut très bien s'intéresser avant tout à un des éléments du quintuple qui constituera son centre d'intérêt. Ainsi, ***à titre d'exemple,*** l'aspect de datation précise de l'opinion recueillie peut constituer le support majeur de l'analyse. En effet, l'aspect de répétition dans le temps permet de détecter des émergences, des disparités et des évolutions qui peuvent être intéressantes pour les besoins d'une analyse d'opinions.

La capacité à utiliser les références de l'auteur de l'opinion est souvent plus difficile lorsque l'on fait du traitement de masse des données d'opinion sur un site, blog, etc.,car elles sont des premières approches et doivent par exemple constituer un élément de pondération des opinions recueillies dans le corpus global. L'analyse sur les caractéristiques, sur les éléments qui composent l'objet,est déjà massivement mise en œuvre sur l'analyse de produits et constitue la véritable utilité opérationnelle de l'analyse de tonalité pour ceux qui réalisent du <social media monitoring> pour le suivi de produits car il s'agit de retours clients précis. C'est pourquoi Liu insiste sur cette analyse des caractéristiques qui suppose de descendre à un niveau d'analyse plus fin que le document ou la phrase.
D'autres aspects également importants est de pouvoir identifier l'objet de l'opinion (cible) ainsi que ces caractéristiques. En effet, classer les opinions, que ce soit au niveau du document ou

au niveau de la phrase, ne donne pas nécessairement des informations suffisantes sur la cible du sentiment exprimé. Les opinions, à ces niveaux, sont généralement trop vagues. ***Par exemple,*** *une opinion positive sur un objet peut être positive sur un aspect de l'objet, mais pas sur l'objet dans son ensemble.* C'est ce que propose « l'Analyse de sentiment basée sur les caractéristiques». Dans cette sous tâche, on détermine d'abord la cible de l'opinion dans la phrase, puis on détermine si celle-ci est positive, neutre ou négative. La cible peut être déclinée en attributs (pour un appareil photo, on trouvera par exemple, son prix, la taille de l'écran, la qualité de l'objet, etc.). On retrouve souvent ce cas dans les avis de consommateurs mais cela suppose donc une phase d'analyse plus approfondie.

Liu dans [2] présente un exemple d'avis proposé sur un iPhone. L'avis exprime différentes opinions (négatives ou positives) sur l'iPhone dans son ensemble ou sur des parties de l'iPhone. Le locuteur doit être identifié à chaque fois (parfois c'est la personne qui parle, parfois ce sont des paroles rapportées de la personne qui parle). L'analyse d'opinion doit être capable d'atteindre ce niveau de détail. Pour Liu, les « objets » (c-à-d les cibles de l'opinion) peuvent être un produit, une personne, un événement, etc. Cet « objet » peut avoir des « composants » ou des «attributs ». Chaque sous-partie peut elle-même avoir des sous-parties. Mais Liu explique que pour simplifier, il préfère parler de « caractéristiques » pour parler à la fois des « composants » et des «attributs». Il distingue alors «l'opinion général» pour faire référence à l'objet de manière général et «d'opinion spécifique» pour désigner les caractéristiques. Ces deux aspects ont été abondamment traités dans la littérature, soit séparément, soit de manière conjointe [5].

Enfin, le dernier quintuple concerne l'orientation (polarité) de l'opinion exprimée sur les caractéristiques et qui définit si l'opinion est favorable ou défavorable. La classification de polarité est probablement la sous-tâche la plus importante et la plus étudiée en analyse d'opinion. Elle consiste à déterminer si un texte exprime des opinions positives ou négatives. Ainsi, cette formalisation de l'opinion en quintuple de Liu nous permet de faire ressortir les différentes tâches sous-jacentes à l'analyse d'opinion et nous permet aussi de distinguer les plus importantes d'entre elles, en particulier celles relatives à la polarité.

1.3 Les niveaux de l'analyse d'opinion

Nombreuses sont les applications permettant de déterminer les sentiments véhiculés au sein de texte. Toutefois, l'échelle sur laquelle porte la recherche du sentiment diffère d'une application à une autre. D'après les travaux que nous avons eu l'occasion d'étudier, il en ressort quatre principaux niveaux de granularité en analyse d'opinions :

- **Niveau du document :** l'analyse à ce niveau vise à déterminer si l'ensemble du document exprime un sentiment positif ou négatif. Ce niveau suppose que chaque document concerne environ une entité précise, comme c'est le cas dans les critiques de films. Ainsi, une telle analyse ne peux s'appliquer si le document traite plusieurs entités.
- **Niveau de phrase :** l'analyse à ce niveau détermine si chaque phrase est positive, négative ou neutre. Peu de travaux ont été répertoriés à ce niveau vu la difficulté de la tâche(Liu, 2012).
- **Niveau Entité(Aspect) :** Ce niveau suppose une analyse fine. Tout d'abord, l'aspect doit être extrait, ensuite le sentiment en lien avec cet aspect doit être mis en valeur. ***Par***

exemple, la nourriture et le service sont des aspects d'un restaurant.

— **Niveau du mot :** La recherche de sentiment à ce niveau s'apparente à une analyse par mots clés en recherche d'information classique . On détermine si le mot implique une polarité négative, positive ou neutre. Habituellement, l'aspect de polarité est utilisé pour des tâches de sentiment à granularité plus forte. Cependant, à ce niveau, la recherche peut être vue comme une approche de construction de lexique de sentiments.

1.4 Les tâches liées à l'analyse d'opinion

La fouille d'opinion se compose de plusieurs tâches, qu'il est utile de mettre en œuvre selon les applications visées. On retrouve :

— Détection de la subjectivité : consiste à identifier dans une collection textuelle les textes porteurs d'opinion, ou encore à localiser les passages porteurs d'opinion dans un texte. Plus précisément, on parle ici de classer les textes ou les parties de texte en objectifs ou subjectifs .

— Classification de l'axiologie de l'opinion (positif, négatif, neutre) :a pour but d'attribuer une étiquette au texte selon l'opinion qu'il exprime. On considère généralement les classes positive et négative, ou encore positive, négative et neutre.

— Classification de l'intensité de l'opinion :a pour but d'attribuer une étiquette au texte selon l'opinion qu'il exprime avec un degré d'intensité par exemple(très négatif,négatif,neutre,,positif, très positif,)

— Identification de l'objet de l'opinion (ce sur quoi porte l'opinion) :consiste à rendre l'information rapidement et facilement accessible en mettant en avant les opinions exprimées et les cibles de ces opinions présentes dans un texte.

— Identification de la source de l'opinion :consiste à déterminer qui exprime l'opinion.

Toutes ces tâches peuvent se pratiquer à différents niveaux, selon l'application envisagée. Cela peut aller de l'analyse au niveau global du texte, au sein d'un aspect particulier du texte, ou alors à des niveaux intermédiaires tel que la phrase, le paragraphe ou la thématique.[3]

1.5 Problématiques de l'analyse d'opinions

L'analyse d'opinion pourrait se comparer à une classification de texte classique « c-à-d. étant donné les mots présents dans le texte, j'en déduis une classe ». ***Par exemple,*** pour une classification thématique, le fait de trouver un certain nombre de mots en lien avec le sport, la politique ou le cinéma aide à prédire l'éventuel classe d'un document. Par contre, les opinions sont rarement exprimées à travers un seul terme.De plus, les discours contenant des appréciations sont par définition des discours subjectifs et l'interprétation de la subjectivité est souvent une tâche délicate. Parmi les problématiques communément rencontrées en analyse d'opinions, on note :

— **problèmes de pertinence** :Il arrive que lors d'une analyse, celle-ci n'attribue pas le sentiment à la bonne cible. Cela arrive, lorsque l'évaluation du texte porte sur différents aspects d'un même produit, ou bien, lorsque'on attribue deux opinions à un même locuteur alors que

deux locuteurs étaient en train de s'exprimer. ***A titre d'exemple,*** *la phrase : « Je trouve que le film est excellent mais ma sœur le trouve mauvais ». Dans ce cas, deux opinions sont exprimées par deux personnes différentes et cela induit souvent à une analyse erronée.*

- **problèmes sémantiques** : de nombreuses figures sémantique ne sont pas repérables de façon fiable : humour, sarcasme, métaphore. Cela nécessite une analyse du contexte plus approfondie, chose qui est hors de la portée des machines actuelles.
- **Problèmes syntaxiques ou terminologiques** :Cela suppose la prise en compte des divers cas particulier de la négation.
- **Problèmes du style d'écriture** :Habituellement sur les réseaux sociaux, la ponctuation et les marques de fin de phrase n'existent pas ce qui rend difficile la distinction linguistique. En effet un langage plus familier, des phrases grammaticalement incorrectes ou des expressions locales empêchent l'analyse correcte de ces opinions.
- **Problème de catégorisation** : trop d'opinions sont classées neutres par défaut, ou présentent deux opinions opposées dans une même phrase et sont classées au hasard (ou en fonction de la fréquence) dans l'une ou l'autre catégorie .
- **Problème de dépendance du domaine** : La dépendance du domaine est en partie une conséquence des changements de vocabulaire. En effet, le passage d'un domaine à un autre implique souvent l'adaptation des ressources utilisé pour l'analyse, Ceci entraîne souvent un énorme travail de reprise, d'adaptation et d'enrichissement des lexiques et dictionnaires utilisés à chaque nouveau domaine.
- **Difficulté de distinction entre les opinions implicite et explicite** :On retrouve généralement une même expression dont la polarité est différente en fonction du contexte, cela impacte fortement les résultats de l'analyse. Pang & Lee dans « Opinion Mining and Sentiment Analysis » illustrent ce cas de figure avec ***l'exemple :*** *« Go read the book »* qui est positif en tant que critique littéraire mais négatif en tant que critique cinématographique.

L'expertise humaine est plus souvent utilisée pour vérifier, corriger, affiner les résultats de la machine, qui travaille elle-même à partir d'entrées et d'algorithmes qui auront été validés par des experts.

1.6 Processus typique de l'analyse d'opinion

Les étapes nécessaires à l'analyse d'opinions ressemblent à celles utilisées pour la fouille de textes classique. La figure ci-dessous illustre un processus typique de classification d'opinions dont les grandes lignes sont l'acquisition du corpus (ensemble des textes), analyse du corpus obtenu et application de méthodes de classification.

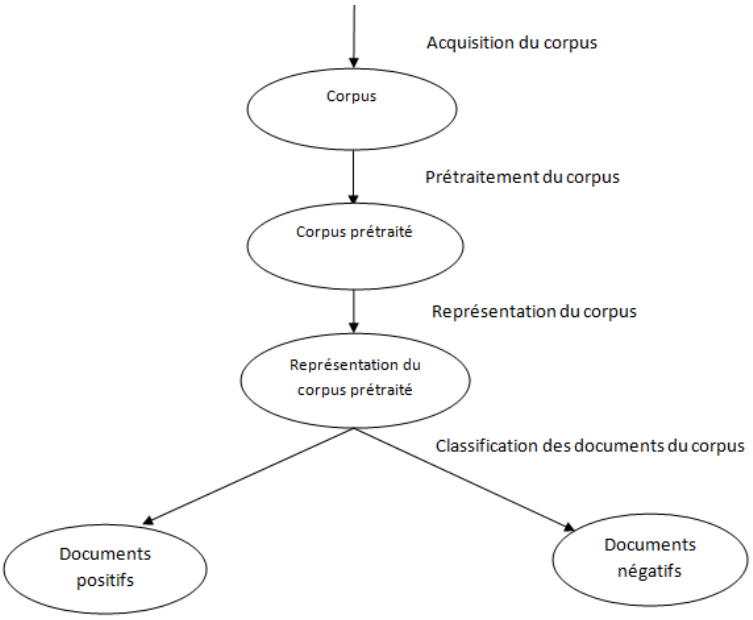


FIGURE 1.1 – Processus d'analyse d'opinion

1. **L'acquisition du corpus d'apprentissage :** L'objectif de cette phase est d'extraire de manière automatique à partir du web des documents d'opinions exprimant des avis positifs ou négatifs.
2. **Le prétraitement du corpus :** Les informations disponible sur le web ne sont pas toujours fiables et peuvent être écrites d'une manière incompréhensible. Pour améliorer les performances de l'analyse des textes opiniâtres, un prétraitement et nettoyage de ces textes est essentiel. Cette tâche consiste à l'élimination des doublons, correction orthographique,tokénisation, suppression des mots vides etc.
3. La **représentations du corpus :** La représentions des textes est une étape très importante dans le processus de fouille d'opinion, pour cela il est nécessaire d'utiliser une technique de représentation efficace permettant de représenter les textes sous une forme exploitable par la machine. La représentation la plus couramment utilisée est celle du modèle vectoriel dans laquelle chaque texte est représenté par un vecteur de n termes pondérés. Les différentes méthodes existantes pour la représentation des textes sont :

   (a) En fonction des approches statistiques :
   - **Représentation en sac de mots (bag of words) :** Les textes sont transformés simplement en vecteurs dont chaque composante représente un mot. Utiliser les mots comme termes a comme avantage d'exclure toute analyse grammaticale et toute notion de distance entre les mots.
   - **Représentation en n-gramme :** Cette méthode consiste à représenter le document par des n-grammes. Le ngramme est une séquence de n caractères consécutifs. Elle consiste a découper le texte en plusieurs séquences de n caractère en se déplaçant avec une fenêtre d'un caractère.

   (b) En fonction des approches lexicales :

- **Représentation en racines lexicales (racinisation) :** Cette méthode consiste à remplacer les mots du document par leurs racines lexicales, et à regrouper les mots de la même racine dans une seule composante.
- **Représentation en lemmes :** La lemmatisation consiste à utiliser l'analyse grammaticale afin de remplacer les verbes par leurs forme infinitive et les noms par leurs forme au singulier. En effet, Un mot donné peut avoir différentes formes dans un texte, mais leurs sens reste le même.

4. **La classification :** Cette dernière phase a pour but de valider l'utilité des termes appris ou choisi lors des phases précédentes puis on classifie de manière automatique des documents. Dans les sections suivantes, nous présentons en détail ces différentes phases.

1.7 Domaine d'application

Plusieurs domaines sont étroitement liés à l'exploitation des opinions ; la figure 1.2 montre les différents secteurs où peut s'appliquer l'analyse d'opinion. Ces multiples applications ont motivé plusieurs travaux répartis sur divers thématiques et que l'on peut classer, en deux grandes catégories :

1.7.1 Usage personnel

Chaque jour de plus en plus d'internautes postent leurs commentaires et partagent leurs opinions sur des sujets variés. Ces opinions constituent une source d'information importante pouvant influencer les internautes dans leur choix. Plusieurs études se sont intéressées aux comportements des internautes autour de cette masse d'informations. Une étude a été réalisée par L'OFT (Office Of Fair Trading) sur le comportement des internautes lors de leurs achat en ligne, elle décrit les différentes raisons pour lesquelles les consommateurs achètent en ligne. Cette étude a été effectuée sur des internautes Britanniques durant les mois de novembre 2006 et 2009.Il a été constaté que l'achat des internautes sur le Net n'arrête pas d'accroitre et que cela est dû essentiellement à la facilité et à l'accessibilité de certaines informations liées aux avis et sentiments qui influencent et mettent en confiance l'internaute pour l'achat de son produit.

Des études américaines confirment ce fait. Elles analysent les achats effectués par les internautes ainsi que l'impact des évaluations postés par d'autres internautes sur leurs habitudes de consommations. L'étude se base sur 2000 internautes américains durant le mois d'octobre 2007 dans des contextes variés tels que les : restaurants, hôtels, voyages, services médicaux, automobiles,..etc. Cette étude révèle que les internautes sont prêts à payer 20% de plus pour les services ayant obtenu la meilleure évaluation ou un étiquetage de «5 étoiles». Actuellement l'information opiniâtres est omniprésente dans plusieurs applications. ***A titre d'exemple*** *pour l'achat d'un ordinateur portable, la consultation des avis d'internautes influe sur la décision finale du consommateur.Il en est de même au niveau des films, la consultation des avis opiniâtres joue aussi sur la décision d'aller voir le film.* Comme nous le montrent ***les exemples suivants :*** *deux films sont extraits du site IMDb, le premier a obtenu une note de 8.1 , tandis que le deuxième obtient une note de 5.0. On aura plus tendance à aller voir le premier film que le second.*

Les services (e.g. avis sur un produit) ne sont pas l'unique motivation des personnes mais l'information politique est un autre facteur important. Une étude a été faite par [3] et ont constaté que plus de 31% des Américains durant l'élection présidentielle de 2006 ont échangé des avis et des informations sur la campagne électorale. Un autre fait plus récent a été l'utilisation des différentes plateformes (Youtube, Facebook), qui a permis de réunir une masse importante d'internautes portant un avis commun sur une révolte gouvernementale (le printemps arabe). Les publications croissantes à teneur politique se font de plus en plus en ligne. Certains chercheurs essayent de déterminer l'accord ou le désaccord des internautes sur un projet de loi.Dans [4] espèrent faciliter la reconnaissance de la position d'un internaute dans un débat politique grâce à l'analyse de ses sentiments.

Les avis des internautes n'intéressent pas uniquement les individus mais aussi les marques, les politiciens et les entreprises qui veulent avoir accès à ces avis ordinaires pour pouvoir les exploiter à leurs avantages.

1.7.2 Usage professionnel

Le marketing a rapidement compris l'intérêt de l'analyse des sentiments. Ainsi, des agences vendent aux entreprises la traque des moindres mots sur leur image ou produit afin que ces derniers s'améliorent. Certains sites repèrent les meilleures critiques émises par les internautes et essayent de les mettre en premier pour donner bonne impression (site Ebay2)[9]. D'autres luttent contre les spams en contribuant à détecter des faux avis postés par des internautes (ou des agences uniquement) pour nuire ou dévaloriser ces entreprises ou ces produits. Une autre utilisation plus ou moins récente concerne les systèmes de recommandations. Cela consiste à extraire les avis des internautes puis à partir de leurs avis positifs leur prédire des produits identiques à leurs attentes.

Plusieurs autres applications utilisent l'analyse des sentiments. Ces applications multiples et variées sont principalement liées à l'évolution des plateformes de réseaux sociaux (blogs, tweets, facebook, etc...) qui motivent d'avantage les recherches sur le sujet.[2]

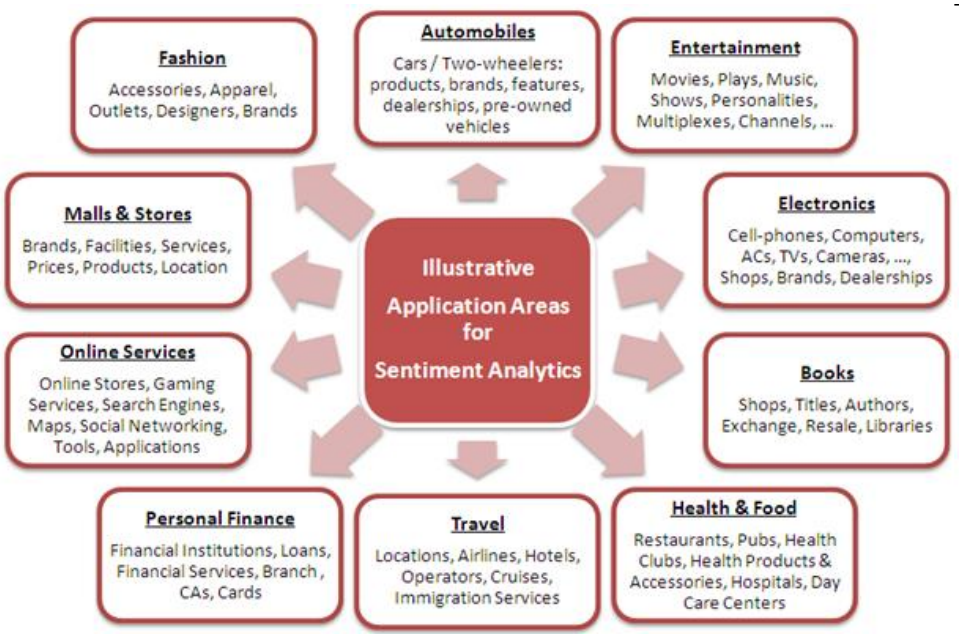


FIGURE 1.2 – Domaines d'application du sentiment analysis selon le cabinet Beacon [4]

1.7.3 Exemple d'applications de la fouille d'opinions

- **Flipkart :** Est une application de shopping en ligne créé par une entreprise indienne de commerce en ligne basée à Bangalore. Cette application utilise le système de classification de produits par étoile afin de les évaluer par les utilisateurs , plus le nombre d'étoile est élevé plus le produit est bon et vice versa.[5]

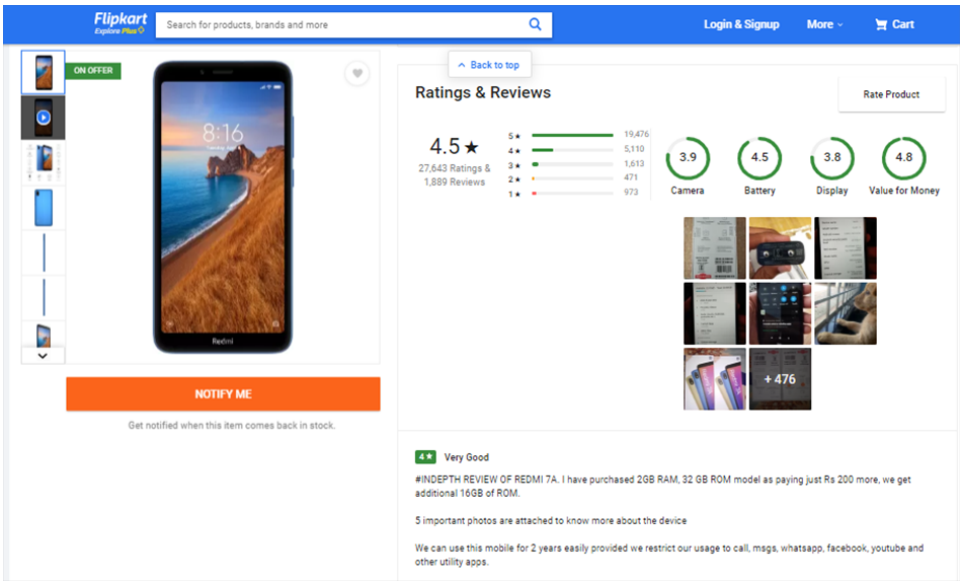


FIGURE 1.3 – Application d'achat en ligne Flipkart

- **Google Play :** Est un magasin d'applications (boutique en ligne) créé par Google le 6 mars

2012 par fusion des services Android Market, Google Movies, Google ebookstore et Google Music. Elle regroupe une boutique d'applications pour le système d'exploitation Android, une boutique de location de films et de séries télévisées, une boutique d'achat de musique, de livres, de magazines et de la gamme de smartphones et tablettes. Il permet à ses utilisateurs de faire un choix parmi les divers applications, sur la base d'un système d'évaluation par étoile.[6]

FIGURE 1.4 – Google play

- **Quot'&Vous :**Derrière le site internet "Quot'&Vous" se trouve l'entreprise TNS Sofres. Il s'agit du deuxième groupe mondial d'étude de marché et d'opinion après sa fusion en 2008 avec l'un de ses concurrents, le groupe Kantar [2].

  Le site Quot'&Vous permet à cette entreprise d'étendre leur immense panel de consommateurs en proposant de répondre à des questionnaires pour donner leurs opinion au sujet d'un produit, d'un service ou de leurs propres habitudes de consommation.

  A chaque sondage auquel l'utilisateur répond des points lui seront attribués et lui permettront de participer à un tirage au sort mensuel. Le principe est simple, plus on répond à des questionnaires, plus on obtient de points et donc plus on augmente les chances de gagner un lot lors du tirage au sort.[7]

2. Kantar TNS, anciennement TNS Sofres et Sofres (Société française d'enquêtes par sondages), est une entreprise de sondages français, créée en 1963 par Pierre Weill. Kantar TNS est aujourd'hui l'une des premières sociétés d'études marketing et d'opinion en France. Elle fait partie du groupe international d'études marketing et de sondages TNS acquis par le leader de la communication et la publicité WPP et intégré à Kantar en 2008.

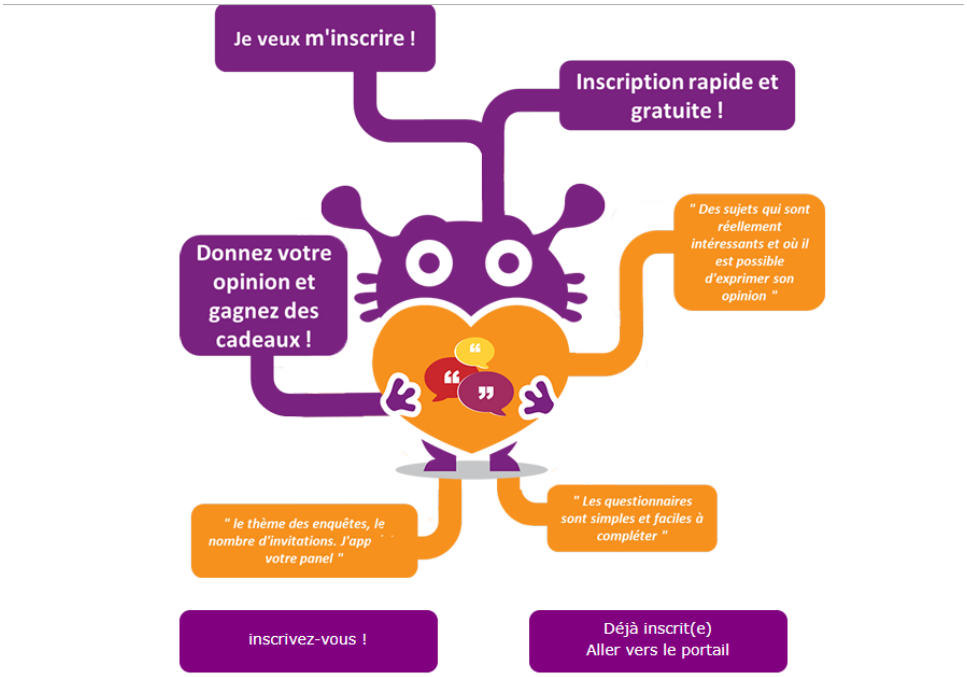


FIGURE 1.5 – Page d'accueil du site Quot'&Vous

- **Analyse de la tonalité sur Twitter :** Le célébre réseau social et de micro-blogging Twitter permet à un utilisateur d'envoyer gratuitement de brefs messages, appelés tweets, sur internet, par messagerie instantanée ou par SMS.

  Avec un nombre moyen de 500 millions de messages envoyés par jour, Twitter est l'un des lieux privilégiés pour recueillir des opinions spontanées sur des sujets très variés, il existe de nombreux services proposant d'analyser la tonalité des messages partagés sur Twitter,et parmis eux on trouve Twitter Sentiment.

  Ce dernier est un outil en ligne gratuit créé par trois étudiants en informatique issue de Stanford. Il s'agit donc d'un projet académique où une timeline est disponible et affiche les courbes de sentiments positifs et négatifs.Un système de retour de pertinence manuel est associé à chacun des résultats et permet d'améliorer le service au fur et à mesure en utilisant l'expertise humaine agrégée des utilisateurs.[8]

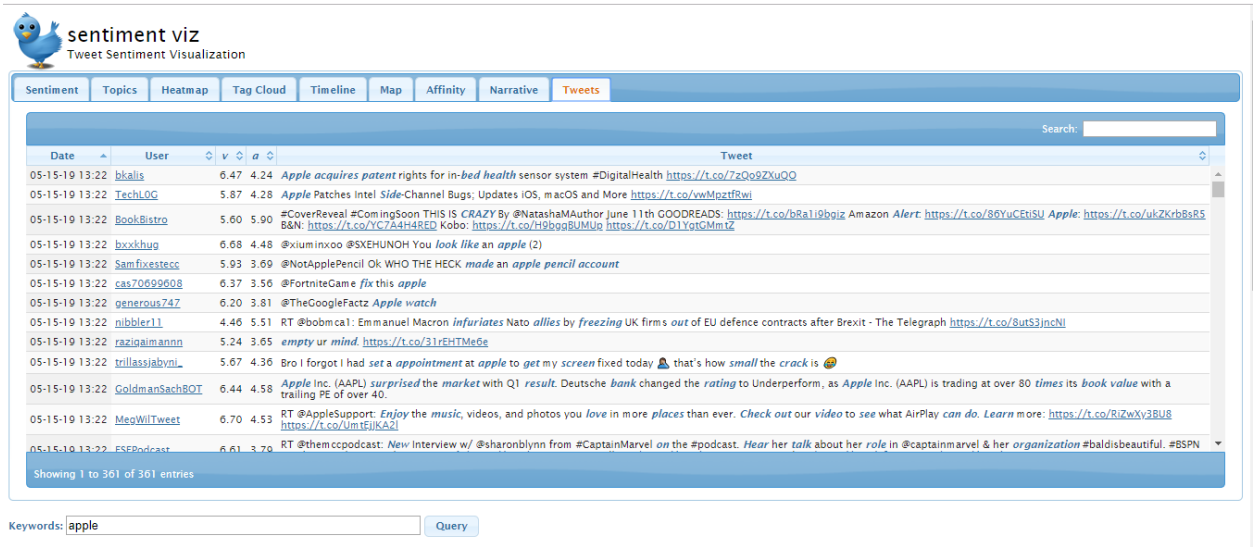


FIGURE 1.6 – Page d'accueil du site sentiments viz

Conclusion

Dans ce chapitre, nous avons introduit quelques notions appartenant au domaine de la fouille d'opinions et ses composants à travers la représentation de Bing & Liu nous avons présenté les principales notions et différents concepts propres à la fouille d'opinion. Nous avons illustré avec un exemple, le processus de fouille d'opinion et ces différentes étapes. Nous avons également discuté des problématiques ainsi que les domaine d'application de l'analyse d'opinion.

Le prochain chapitre présentera les différentes approches et méthodes utilisées en détection de polarité. On énuméra quelques travaux de recherches liées à la fouille d'opinions

Chapitre 2

Etat de l'art sur L'analyse d'opinion

Introduction

L'année 2001 marque le début d'une demande croissante en systèmes automatiques d'analyse d'opinions. Cette demande a émané d'une combinaison entre l'essor de méthodes d'apprentissage automatique et de la disponibilité d'ensembles de données riche grâce à l'expansion des informations sur le Web.

L'analyse d'opinions est justement un produit de toutes les innovations dans le domaine de la catégorisation de textes. En effet, les linguistes s'intéressent de plus en plus au domaine. Cet intérêt accru est principalement dû aux évolutions dans le domaine des méthodes d'apprentissage automatique et des rapprochement qui existent entre les deux domaines, On y retrouve ainsi, deux grandes catégories d'approches connue sous l'anglicisme (lexicon-based approach & machine learning based approach).

Dans ce chapitre, on présentera ces deux catégories d'approches, en mettant l'accent sur les techniques d'apprentissage automatique. On énumèrera les principaux classifieurs et mesures d'évaluations employées en analyse d'opinion.

2.1 Les approches de détection d'opinions

Tels que décrit ci-dessus, les approches utilisées en détection d'opinions se divisent en deux catégories. La première est basées sur des lexiques (dictionnaires, thésaurus etc.) de mots subjectifs et pondérés (mots exprimant une opinion). Si un document comporte ces mots, une somme du score de polarité des mots constituants le texte sera alors calculé pour déduire la polarité globale du document. Les approches basées sur l'apprentissage supervisé, quand à elle, utilisent différents types de classifieurs tels que SVM (Machine à Vecteur de Support), "Naıve Bayes" etc. afin de générer des modèles de classification, soit de manière supérvisée ou non supérvisée. Ainsi, on note :

2.1.1 Approches basées sur les lexiques

La principale tâche dans cette approche est la conception de lexiques ou dictionnaires d'opinion. L'objectif de ces lexiques ou dictionnaires est de répertorier le plus de mots porteurs d'opinion possible.Ces mots permettent ensuite de classer les textes en deux(positif et négatif)ou trois catégories(positif,négatif,neutre). Dans[14] Liu et al décrivent un système, Opinion Observer, qui permet de comparer des produits concurrents en utilisant les commentaires écrits par les internautes. ils ont une liste prédéfinie de termes désignant des caractéristiques de produits. Lorsqu'une de ces caractéristique est présente dans un texte, le système extrait les adjectifs proches dans la phrase. Ces adjectifs sont ensuite comparés aux adjectifs présents dans leur dictionnaire d'opinion et ainsi, une polarité est attribuée à la caractéristique du produit.

Cette méthode nécessite donc la construction d'un dictionnaire d'opinion. Pour construire un tel dictionnaire, trois genres de techniques sont possible :

— ***la méthode manuelle*** : cette méthode demande un effort important en terme de temps mais il faut savoir que toute les autres méthodes nécessitent également de créer initialement,de façon manuelle,un ensemble de mots et expressions porteurs d'opinions.Cet en-

semble de mots est appelé *graine* Il est ensuite utilisé afin de trouver d'autres mots et expressions porteurs d'opinions.

— ***la méthode basée sur les corpus :*** afin d'agrémenter cet ensemble de mots l'utilisation de corpus de textes,[5] propose la méthode suivante : afin de déterminer la polarité de mots ou expressions non classés, il compte le nombre de fois où ces mots ou expressions apparaissent dans le corpus à côté de mots ou expressions déjà classés. Un mot apparaissant plus souvent à côté des mots positifs sera donc classé dans la catégorie positive et inversement. Yu propose une méthode similaire, mise à part qu'ils utilisent la probabilité qu'un mot non classé soit proche d'un mot classé afin de mesurer la force de l'orientation du premier nommé.

— ***la méthode basée sur les dictionnaires :*** cette méthode consiste à utiliser des dictionnaires de synonymes et antonymes existants tels que SentiWordNet.Afin de déterminer l'orientation sémantique de nouveaux mots,[6] utilisent ces dictionnaires afin de prédire l'orientation sémantique des adjectifs.Dans SentiWordNet les mots sont organisés sous forme d'arbre (voir la figure ci-dessous) . Afin de déterminer la polarité d'un mot, ils traversent les arbres de synonymes et d'antonymes de mots et s'ils trouvent déjà un mot classé parmi ces derniers , ils affectent la même polarité au mot étudié ou bien la polarité opposée s'ils trouvent un mot déjà classé parmi les antonymes. S'ils ne croisent aucun mot déjà classé, ils réitèrent l'expérience en partant de tous les synonymes et antonymes et ce jusqu'à rencontrer un mot d'orientation sémantique connue.

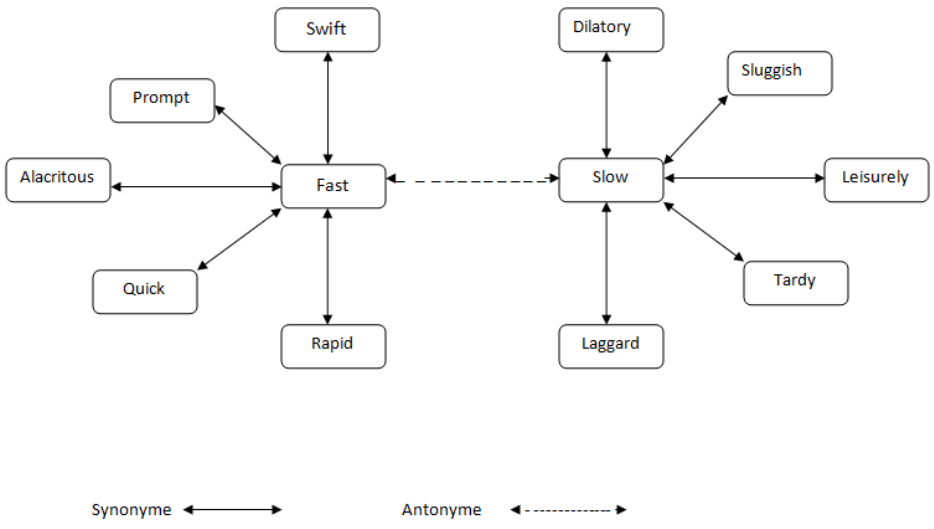


FIGURE 2.1 – Exemple d'arbre de synonymes et d'antonymes présents dans WordNet. [15]

2.1.2 Approches basées sur l'apprentissage automatique

Certaines des approches les plus populaires en détection d'opinions se basent sur les méthodes d'apprentissage automatique. Ces approches considèrent la tâche de classification de sentiments comme une catégorisation de textes, où ces derniers sont classés dans une des catégories prédéfinies en utilisant des informations (nommées caractéristiques en français et features en anglais)

pour entrainer les algorithmes de classification. En catégorisation de textes classique, divers algorithmes d'apprentissage automatique ont été appliqués et ont prouvé leurs efficacité. Ces mêmes algorithmes ont également été appliquées avec succée à la classification de sentiment, parmi lesquelles on retrouve :

1. **Naive Bayes :** La classification naïve bayésienne est un type de classification probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses. Elle met en oeuvre un classifieur bayésien naïf, ou classifieur naïf de Bayes, appartenant à la famille des classifieurs linéaires.En termes simples, un classifieur bayésien naïf suppose que l'existence d'une caractéristique pour une classe, est indépendante de l'existence d'autres caractéristiques. Un fruit peut être considéré comme une pomme s'il est rouge, arrondi, et fait une dizaine de centimètres. Même si ces caractéristiques sont liées dans la réalité, un classifieur bayésien naïf déterminera que le fruit est une pomme en considérant indépendamment ces caractéristiques de couleur, de forme et de taille. [1]
2. **Machine à vecteurs de support (SVM) :** Cet algorithme de classification repose sur la notion d'hyperplan séparateur et de marge maximale. Un hyperplan séparateur entre deux ensembles de points représente la frontière entre deux catégories d'ensembles (ensemble de documents subjectifs/objectifs ou Positif/Négatif). La marge représente la distance entre les points (vecteurs de support) et l'hyperplan. *Considérons l'exemple suivant :* On se place dans le plan, et l'on dispose de deux catégories : les ronds rouges et les carrés bleus, chacune occupant une région différente du plan. Cependant, la frontière entre ces deux régions n'est pas connue. Ce que l'on veut, c'est que quand on lui présentera un nouveau point dont on ne connaît que la position dans le plan, l'algorithme de classification sera capable de prédire si ce nouveau point est un rond rouge ou un carré bleu.

   Voici notre problème de classification : pour chaque nouvelle entrée, être capable de déterminer à quelle catégorie cette entrée appartient.
3. **Régression Logistique :** C'est une méthode statique permettant de produire un modèle pour décrire des relations entre une variable catégorielle et un ensemble de variables de prédiction. Cette méthode de classification a été mise en œuvre dans le cadre de nombreuses applications : en médecine, en économie, en sciences sociales et en législation. Elle donne de bons résultats, cependant, elle a été très peu utilisée dans le cadre de la détection d'opinions en particulier [2].

2.2 Méthodes d'apprentissage automatique

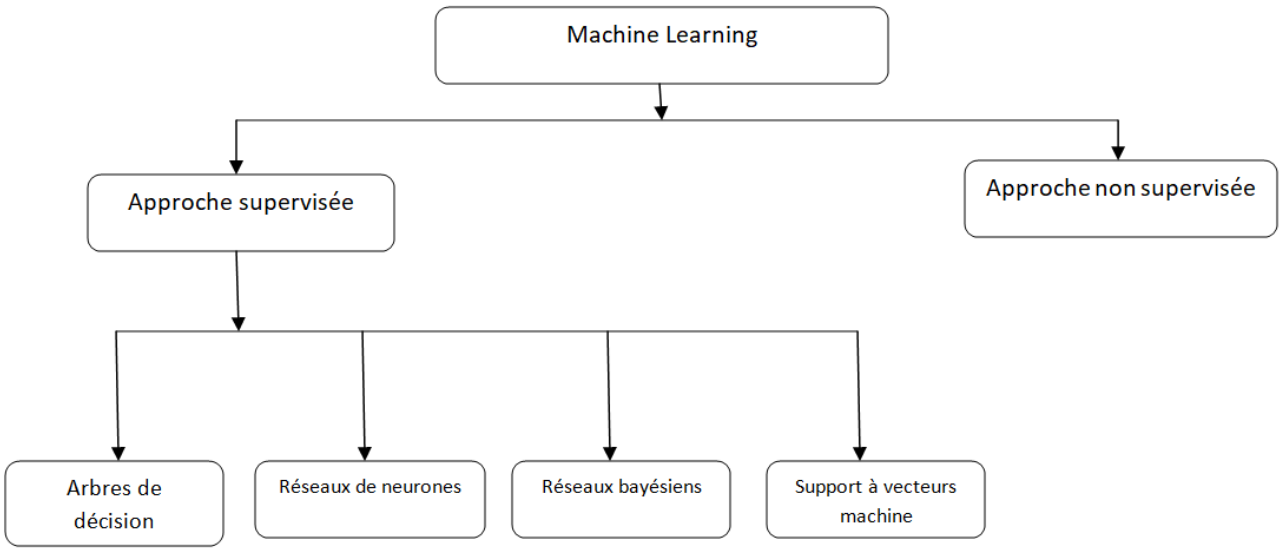


FIGURE 2.2 – les méthodes d'apprentissage supervisée

Deux des méthodes d'apprentissage automatique communément adoptées sont :

- **L'apprentissage supervisé :** qui permet de générer des modèles sur la base de données d'entrée étiquetées manuellement.
- **L'apprentissage non supervisé :** auquel on ne fournit aucune données étiquetées à l'algorithme afin de lui permettre de trouver une structure et de découvrir une logique dans les données en entrées.

Ainsi, on note :

2.3 L'apprentissage non supervisé ou clustering

L'apprentissage non supervisé vise à concevoir un modèle structurant l'information. La spécificité dans ce type de méthodes est que les catégories(les classes) des données d'apprentissage ne sont pas connus à l'avance, c'est ce que l'on cherche à trouver.

Un système d'analyse en clusters prend en entrée un ensemble de données et une mesure de similarité entre ces données, puis l'algorithme produit en sortie un ensemble de clusters. Les données sont généralement des enregistrements (ou objets) composés de champs ou d'attributs. Formellement, un système de clustering prend un tuplet (D ; S) où D représente l'ensemble de données et S la mesure de similarité, et retourne un ensemble $(C_1, C_2, \ldots, C_m)$ tel que :
C_i (i = 1..m) sont des sous ensembles de D qui vérifient :

$C_1 \cup C_2 \cup C_3 \cup ..... \cup C_n = D \ \& \ C_i \cap C_j = \emptyset$

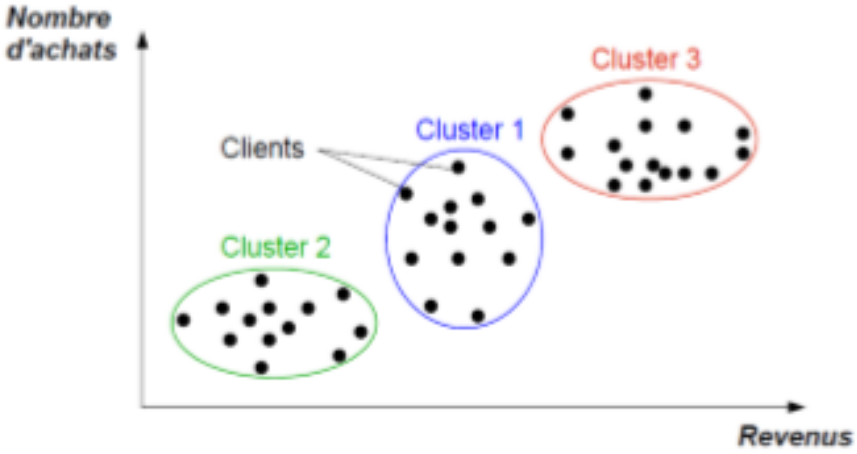


FIGURE 2.3 – un exemple de clustering des clients selon leurs revenus et leurs achats. [4]

Chaque C_i est considéré comme un cluster qui représente une ou plusieurs caractéristiques de l'ensemble D.
Chaque cluster issu de ce processus doit vérifier les deux propriétés suivantes :

- **La cohésion interne :** les objets appartenant à ce cluster sont les plus similaires possibles.
- **L'isolation externe :** les objets appartenant aux autres clusters sont les plus distinct possibles.

Cependant La problèmatique qui se pose est de savoir précisément le nombre de clusters à rechercher

— Dans certains cas c'est un expert du domaine d'application qui fournit le nombre de clusters.

— Mais Dans la majorité des cas, on définit une mesure de stabilité du processus d'analyse sur la base à laquelle on peut atteindre le meilleur nombre de clusters décrivant au mieux les données.

Souvent, les clusters peuvent se chevaucher (un objet peut appartenir à plusieurs clusters), dans ce cas on parle de clustering recouvrant,et dans le cas contraire on parle de clustering exclusif.

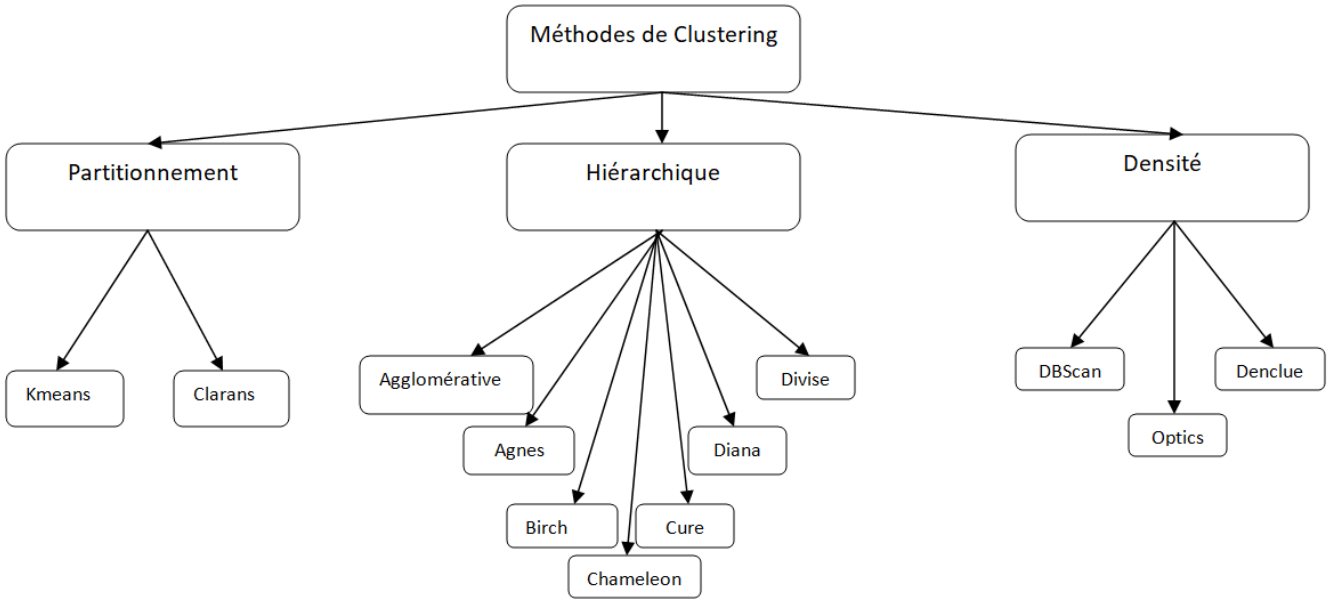


FIGURE 2.4 – les méthodes de clustering.[4]

Parmi les différents algorithmes non supervisé, le plus simple et le plus utilisé est l'algorithme K-means (Mac- Queen 1967) dit aussi méthode des centres mobiles ou méthode des K-moyennes :

2.3.1 L'algorithme de K-means

Cet algorithme est considéré comme un des algorithmes d'apprentissage de classification non supervisé le plus utilisé, du fait de sa simplicité de mise en œuvre. Son objectif consiste à construire les k meilleurs centres de classes (clusters ou groupes) de l'ensemble de données d'apprentissage.

L'algorithme démarre d'une partition arbitraire des enregistrements sur les k clusters, à chaque itération il calcule les centres de ces clusters puis il effectue une nouvelle affectation des enregistrements aux plus proches centres. Il s'arrête dés qu'un critère d'arrêt est satisfait,généralement on s'arrête si aucun enregistrement ne change pas de cluster[4] .

La description formelle du déroulement de cet algorithme est comme suit :

Entrée : un échantillon de m enregistrements $X_1, \ldots, X_m$

1. Choisir k objets formant ainsi k clusters C_i de centre M_i
2. (Ré) affecter chaque enregistrement xi au cluster Ci de centre Mi tel que distance (x_i, M_i) est minimal
3. Recalculer Mi de chaque cluster, pour tout i, M_i est la moyenne des éléments du cluster i
4. Aller à l'étape 2 jusqu'à ce qu'aucun enregistrement ne change de cluster alors arrêt et sortir les clusters.

L'avantage de cette méthode est la facilité de l'implémentation avec des grands volumes de données. Mais Le choix du paramètre k n'est pas découvert il est choisi par l'utilisateur, aussi la solution dépend des k centre de gravité choisie lors de l'initialisation.

2.4 L'apprentissage supervisé

En apprentissage supervisé, la machine s'appuie sur des classes prédéterminées et sur un certain nombre de paradigmes connus pour mettre en place un système de classification à partir de modèles déjà catalogués. Ce processus est réalisé en deux étapes. Dans un premier temps, une phase d'apprentissage permet d'engendrer une modélisation des données annotées. Ensuite, dans un second temps, on attributs des classes à de nouvelles données non annotées introduites dans le système, afin de les classer sur la base du modèle généré.

Il existe deux types de problèmes auxquels l'apprentissage supervisé est appliqué :

- **La classification :** consiste à identifier les classes d'appartenance de nouveaux objets à partir d'exemples antérieurs connus, la variable à prédire peut donc prendre des valeurs discrètes appelées classes (exemple : positif, neutre, négatif).
- **La régression :** est utilisée lorsqu'il s'agit de prédire une variable continue, qui peut donc prendre un nombre infini de valeurs .

Ainsi, à partir des éléments ci-dessus nous pouvons en déduire l'interprétation suivante d'un classifieur :

Un classifieur est une procédure qui, à l'aide d'un ensemble d'exemples, produit une prédiction de la classe auquel appartient des données en entrée. Beaucoup de méthodes de classification supervisée existent comme le montre la figure 2.2 :

Habituellement, le fonctionnement d'un classifieur suit le cheminement suivant :

1. **Phase d'apprentissage :** Dans cette phase l'algorithme (classifieur) reçoit en entrée des exemples d'apprentissage étiquetés (documents d'entrainement) et produit un modèle de prédiction en veillant à produire le moins d'erreurs de prédiction possible. Le concept de prédiction est définie comme étant une tâche qui vise à présager statistiquement une ou plusieurs caractéristiques inconnues à partir d'un ensemble de caractéristique préalablement connues.

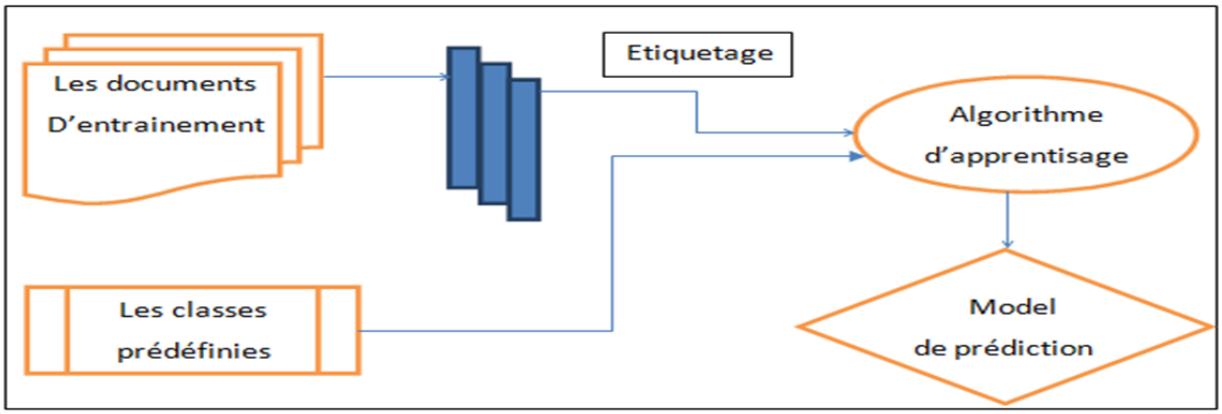


FIGURE 2.5 – La phase d'apprentissage

2. **Phase de test** : Dans cette phase, le modèle obtenu lors de la phase d'apprentissage doit être capable de prédire l'étiquette d'un nouvel exemple en fonction des valeurs d'entrées :

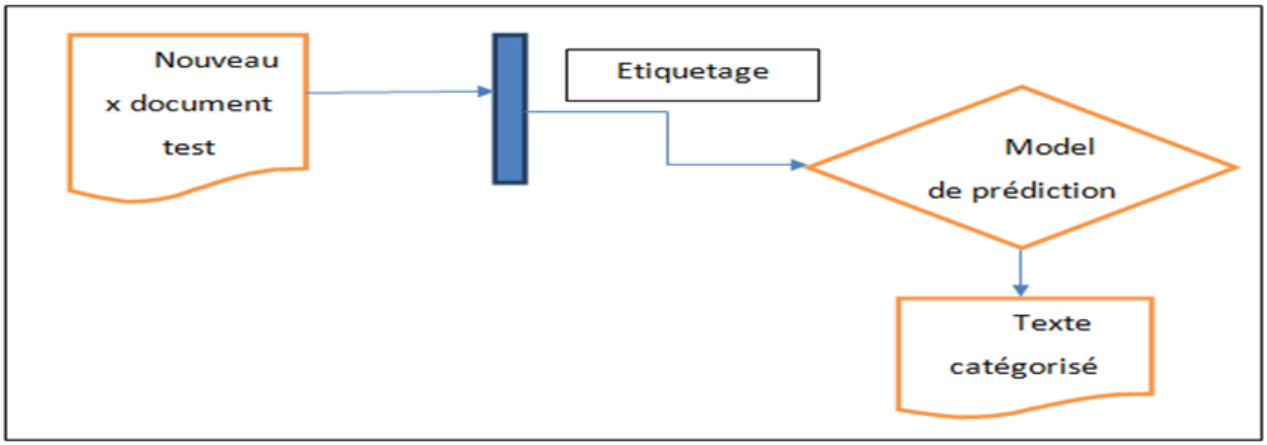


FIGURE 2.6 – La phase de test

Ci dessous, nous présentons brièvement certains des classificateurs les plus fréquemment utilisés en analyse d'opinions :

2.4.1 Les arbres de décision

Les arbres de décision sont des outils d'aide à la décision représentant un ensemble de choix sous la forme graphique d'un arbre[3]. Le formalisme de ce classifieur permet de classifier de nouveaux documents en testant ses caractéristiques les unes à la suite des autres. La classification se fait à travers une séquence d'évaluations dans laquelle chaque test dépend du résultat de l'évaluation précédente.Cette séquence est représentée par un arbre de décision dont les noeuds (feuilles) de l'extrémité représentent les classes. Pour la phase de construction du classifieur, les exemples de l'ensemble d'apprentissage sont décomposés hiérarchiquement par des tests définis sur les caractéristiques pour obtenir des sous-ensembles d'exemples qui refléteront les différentes classes de prédilections.

La figure ci-dessous montre un arbre de décision pour le jeu de Tennis. L'objectif est de déterminer si les conditions climatiques du jour sont convenable pour le jeu ou non :

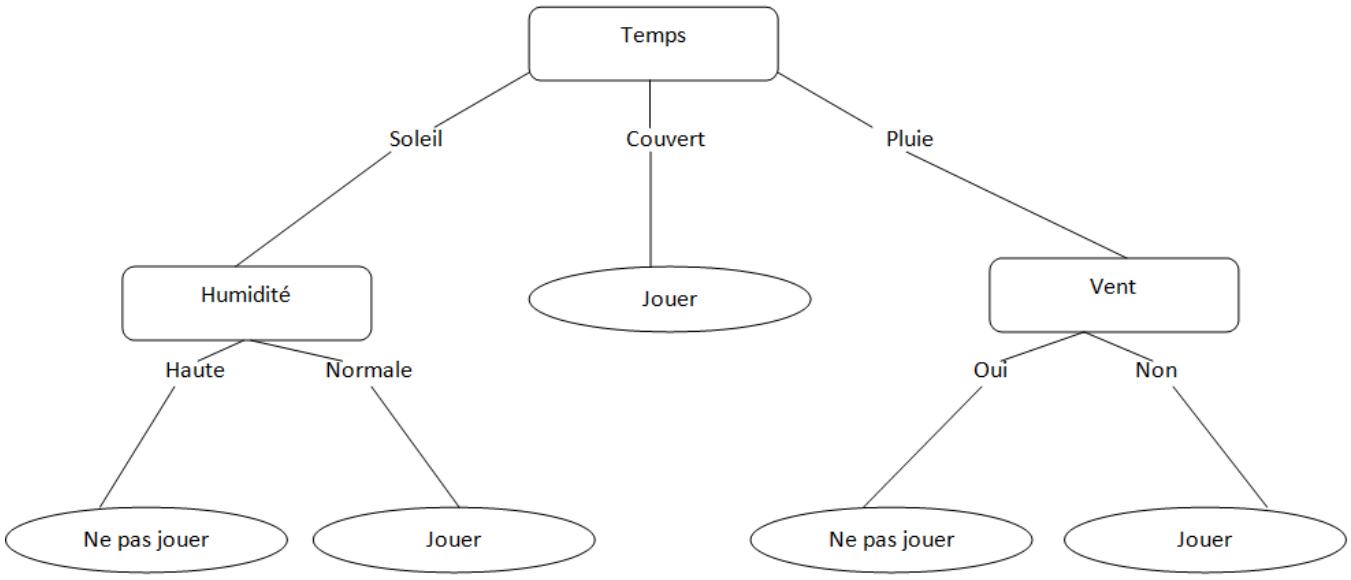


FIGURE 2.7 – Exemple d'application de l'algorithme d'arbre de décision

Exemple d'algorithme de construction des arbres de décision

- Algorithme de CHAID (CHi-squared Automatic Interaction Detector) :Est un algorithme de construction d'arbre développé vers 1980 par Gordon V. Kass , il a été l'un des algorithmes à être implémenté dans des logiciels commerciaux, il peut être utilisé pour la prédiction ou pour la détection d'interaction entre variables. À la fin des années 70 et au début des années 80, J. Ross Quinlan , un chercheur dans l'apprentissage de la machine, a développé un algorithme d'arbre de décision connu sous le nom ID3 (Induction of Decision Tree), ce travail étendu sur des travaux antérieurs dans les systèmes d'apprentissage de concepts, Quinlan plus tard a présenté C4.5 c'est une version améliorée de l'algorithme ID3, qui est devenu une référence à laquelle les nouveaux algorithmes d'apprentissage supervisé sont souvent comparés. Contrairement à ID3 qui ne manipule que les attributs nominaux, C4.5 peut manipuler les attributs numériques et les attributs discrets. En 1984, un groupe de statisticiens a publié la méthode de la classification et des arbres de régression (CART) décrite en détail dans une monographie qui fait encore référence aujourd'hui, elle permet de décrire la génération d'arbres de décision binaires.
- Algorithme ID3 :L'ID3 est un algorithme développé par Ross Quinlan en 1986.Il construit un arbre de décision de façon récursive en choisissant l'attribut qui maximise le gain d'information selon l'entropie de Shannon. Cet algorithme fonctionne exclusivement avec des attributs catégoriques et un nœud est créé pour chaque valeur des attributs sélectionnés.
- Algorithme C4.5 : Le C4.5 est un algorithme d'apprentissage automatique, il a été créé par Ross Quinlan en 1993 et il est considèré comme une amélioration de l'algorithme ID3, afin de pallier les limites de son prédécesseur, il offre la possibilité de :
  - Meilleur adaptation de la fonction de gain.
  - gérer des attributs avec des valeurs manquantes, nulles.

— élaguer l'arbre pour éviter un "overfitting".

— manipuler des valeurs continues.

- Algorithme de CART (Classification And Regression Tree) : Est un algorithme développé par Breiman, Friedman, Olshen et Stone. C'est une méthode de discrimination basée sur la construction d'un arbre de décision binaire qui a pour but de construire, à partir d'une population, des sous-groupes qui soient le plus homogène possible pour une caractéristique donnée (variable à expliquer). Correspond à deux situations bien distinctes selon que la variable à modéliser ou prévoir est :
  — Qualitatives : classification
  — Quantitative : régression

2.4.2 Avantages et inconvénients des arbres de décision

Avantages	**Inconvénients**
• Simple à comprendre et à interpréter. • Capable de traiter à la fois des données quantitatives et qualitatives. • Performant sur de grands jeux de données. • Nécessite un minimum de préparation de données.alors qu'en revanche d'autres techniques nécessitent souvent la normalisation des données, des variables indicatrices doivent être créées et des valeurs vides doivent être enlevées.	• Les arbres de décision peuvent être instables en raison des petites variations dans les données qui risque de générer un arbre complètement différent. • Dans certain cas, l'apprentissage par arbre de décision peut nous conduire à la génération d'arbres de décision très complexes, qui généralisent mal l'ensemble d'apprentissage ce problème est connue sous le nom de surapprentissage.

2.4.3 Les réseaux de neurones

Les réseaux de neurones artificiels sont des méthodes d'apprentissage automatique inspirées du modèle de neurone biologique, proposés par les biophysiciens Mac Cullogh et Pitts en 1943. Un réseau de neurones artificiels est un ensemble de neurones formels connectés les uns avec les autres selon plusieurs modalités et architectures afin de réaliser certaines tâches tels que la reconnaissance facial, la traduction automatique, la reconnaissance vocal...etc. A ce jour les réseaux de neurones constituent une méthode de traitement de données bien comprise et bien maîtrisée. De façon formelle, un RN est une fonction mathématique associant à des entrées, des grandeurs de sortie à l'aide de paramètres ajustables appelés des poids.[4]

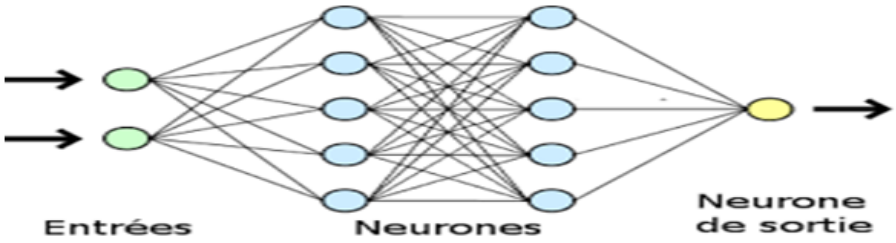


FIGURE 2.8 – schéma d'un réseau de neurones artificielles.
[4]

D'autre part, un neurone formel est une représentation mathématique, sous forme d'un ensemble d'entrées auxquelles on associe des poids synaptique et une fonction d'activation qui diligente la sortie du neurone lorsque la somme des entrées multiplié par les poids synaptiques atteint un seuil défini.

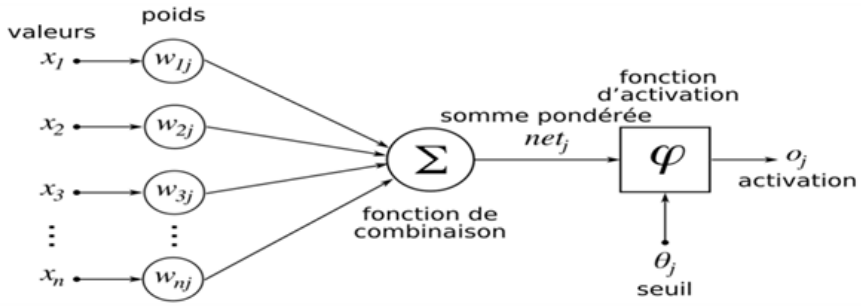


FIGURE 2.9 – schéma d'un neurone formel.[4]

La Figure ci-dessus illustre la structure générale d'un neurone formel, caractérisée par :

- **Les entrées** (X_1,X_2,...,X_i,...,X_n) collectées depuis les données que l'on souhaite traiter , ou bien à partir des sorties d'autres neurones.
- **La somme pondérée :** Correspond à la somme définissant le prétraitement (combinaison linéaire) effectuée sur les entrées comme le montre l'équation :

$$\text{Sum } W_i.X_i+\text{b}$$

  Où W_i est le poids synaptique attaché à l'entrée **i** et **b** désigne le seuil d'activation (biais).
- **La fonction d'activation (ou d'état** f_0**)** définissant l'état interne du neurone en fonction de son entrée totale.
- **Et enfin une sortie** calculant le résultat du neurone en fonction de son état d'activation.

2.4.4 Avantages et inconvénients des réseaux de neurones :

Avantages	Inconvénients
• La capacité d'adaptabilité et d'auto-organisation. • La possibilité de résoudre des problèmes non-linéaires avec une bonne approximation. • La rapidité d'exécution.	• La difficulté d'interpréter le comportement d'un réseau de neurones est un inconvénient pour la mise au point d'une application. • Il est souvent impossible d'utiliser les résultats obtenus pour améliorer le comportement d'un réseau de neurones.

2.4.5 Les réseaux bayésiens

Les réseaux bayésiens sont des modèles graphiques probabilistes qui permettent de représenter les connaissances probabilistes d'un système de support à la décision. Ils sont essentiellement basés sur le théorème de BAYES et se modélisent sous forme d'un graphe orienté, dont les nœuds sont des variables aléatoires et des tables de probabilités. La représentation graphique permet de proposer un langage facilitant la compréhension des relations entre les variables, il s'agit de l'aspect qualitatif du modèle. A chaque nœud est associée une table de probabilités qui définit sa relation avec les nœuds parents, il s'agit de l'aspect quantitatif du modèle.[5]

Un cas particulier de ces réseaux est appelé « réseaux naïf de BAYES » dans lequel on suppose l'indépendance des variables entre elles.

La classification naïve bayésienne

L'algorithme naïf de BAYES est un algorithme d'apprentissage supervisé dédié aux problèmes de classification. C'est une approche probabiliste basée sur les probabilités conditionnelles et le théorème de BAYES. Il permet de prévoir des comportements futurs pour l'aide à la prise de décision. Cet algorithme est connu pour sa simplicité et son efficacité.

Cette méthode suppose que les variables sont indépendantes les unes des autres et ne prend en compte aucune corrélation ou relation entre elles. Même si ces variables sont liées le classificateur naïf de BAYES déterminera qu'un objet donné appartient à une classe en considérant indépendamment ses caractéristiques, d'où l'appellation « classification naïve ».

Pour mettre en place un classificateur naïf de BAYES, il faudra :

- Déterminer un ensemble d'apprentissage.
- Déterminer les probabilités à priori de chaque classe.
- Calculer les probabilités conditionnelles d'appartenance aux classes pour toutes les valeurs de chaque variable.
- Appliquer la règle de BAYES pour obtenir les probabilités à postériori des classes au point X.
- Choisir la classe la plus probable.

De manière plus formelle, le calcul de la probabilité qu'un ensemble de caractéristiques appartiennent à une classe particulière est calculée à l'aide du théorème de Bayes.

$$P(classe|carcatristique) = \frac{P(caractristique|classe) * P(classe)}{P(caractristique)}$$

P (classe) est la probabilité à priori de la classe, elle est aussi appelée « probabilité marginale de la classe ». P (caractéristiques |classe) est la fonction de vraisemblance de la classe sachant la probabilité de la caractéristique. P (caractéristiques) est la probabilité à priori qu'un ensemble de caractéristiques donné se soit produit. Compte tenu de cette hypothèse naïve qui sous-entend que toutes les caractéristiques sont indépendantes, la première équation pourrait donc être réécrite comme suit :

$$P(classe|carcatristique) = \frac{P(classe) * P(classe|f_i) * ..... * P(classe|f_n)}{P(caractristique)}$$

2.4.6 Avantages et inconvénients des réseaux bayèsiens

Avantages	Inconvénients
• L'algorithme offre de bonne performance. • Un réseau bayésien est polyvalent, nous pouvons nous servir du même modèle pour évaluer, prévoir, diagnostiquer, ou optimiser des décisions. • La représentation graphique d'un réseau bayésien est explicite, intuitive et compréhensible par un non spécialiste, ce qui facilite à la fois la validation du modèle, ses évolutions éventuelles et surtout son utilisation. • Les réseaux bayésiens donnent la possibilité de rassembler et de fusionner des connaissances de diverses natures dans un même modèle.	• La généralité du formalisme des réseaux bayésiens aussi bien en termes de représentation que d'utilisation les rend difficiles à manipuler à partir d'une certaine taille. • La prédiction devient erronée si l'hypothèse d'indépendance conditionnelle est invalide.

2.4.7 Les Supports Vecteur Machines

Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais Support Vector Machine, SVM) est une méthode d'apprentissage supervisé permettant de résoudre des problèmes de classification et de régression et pouvant gérer des variables quantitatives et qualitatives. Son but est de catégoriser des exemples représentés dans un hyper-espaces de telle sorte que la distance entre les différentes classes de données et la frontière qui les sépare soit maximale.

Cet algorithme vise donc à créer une frontière de décision (ou de séparation) entre deux classes pour permettre la prédiction d'étiquettes à partir d'un ou plusieurs vecteurs de caractéristiques (les axes de cet hyperespace). Cette frontière de séparation, appelée hyperplan, est orientée de manière à être aussi éloignée que possible des points de données les plus proches caractérisant chacune des classes. Ces points les plus proches sont appelés vecteurs de support. Dans les SVM, la frontière de séparation est choisie comme étant celle qui maximise la marge. La marge représente la distance entre l'hyperplan optimal et les vecteurs de support. Les Machines à Vecteur de Support à vaste marge, sont des classifieurs binaires très populaire en classification de polarité.

Pour illustrer son fonctionnement, considérons le cas où on dispose de données linéairement séparables. Nommons positif et négatif les deux classes appartenant à Y. Si le problème est

linéairement séparable, les documents positifs sont séparables des documents négatifs par un hyperplan H. Notons H+ l'hyperplan parallèle à H qui contient le document positif le plus proche respectivement H- pour le document négatif. Une machine à vecteur de support linéaire recherche alors l'hyperplan qui sépare les données de manière à ce que la distance entre H+ et H- soit la plus grande possible. Cet écart entre les deux hyperplans H+et H- est appelés la marge.

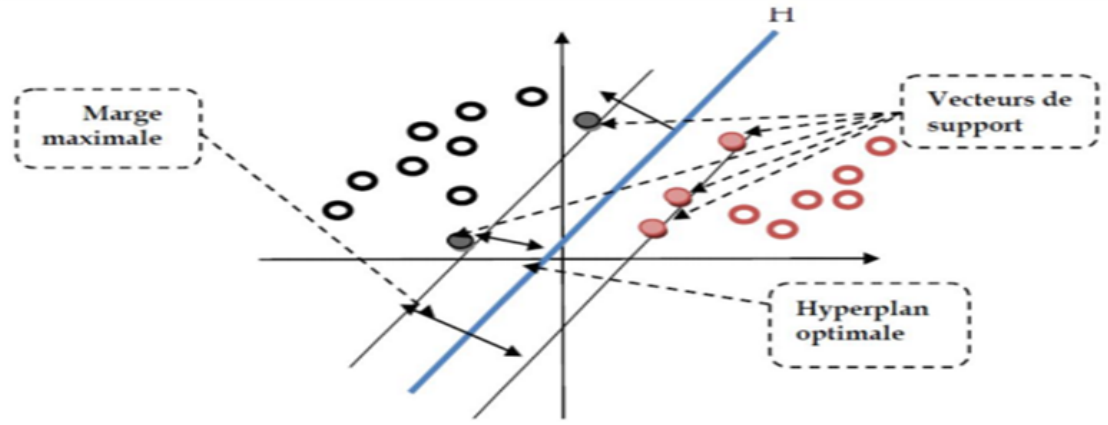


FIGURE 2.10 – exemples d'un séparateur à vaste marge
[4]

2.4.8 Avantages et inconvénients des SVM

Avantages	Inconvénients
• Très efficace dans le cas d'une forte dimentionnalité. • Peu de paramètres à régler. • Une grande vitesse d'apprentissage. • Moins gourmante en terme d'espace mémoire.	• Si le nombre d'attributs est beaucoup plus grand que le nombre d'échantillons, les performances seront moins bonnes. • Comme il s'agit de méthodes de discrimination entre les classes, elles ne fournissent pas des estimations de probabilités.

2.5 Étapes d'un processus de catégorisation de texte

L'analyse d'opinion utilisant l'apprentissage automatique peut être assimilée d'une manière générale à un processus classique de classification supervisée de texte. Ces méthodes reposent, généralement, sur l'existence préalable d'un corpus de données constitués généralement par des

experts du domaine. Une fois que le corpus est disponible, l'idée est d'apprendre automatiquement des unités linguistiques ou termes, pour modéliser des catégories particulières, des thématiques ou encore des opinions.

Toutefois, avant d'appliquer un quelconque algorithmes d'apprentissage, de nombreux problèmes de langue doivent être préalablement résolus en procédant au nettoyage,transformation et représentation de ces données textuelles mal structurées. Ci-dessous, nous présenterons les différentes étapes de pré-traitement linguistique communément utilisées en classification de textes. Nous aborderons ensuite, la représentation de textes utilisée.

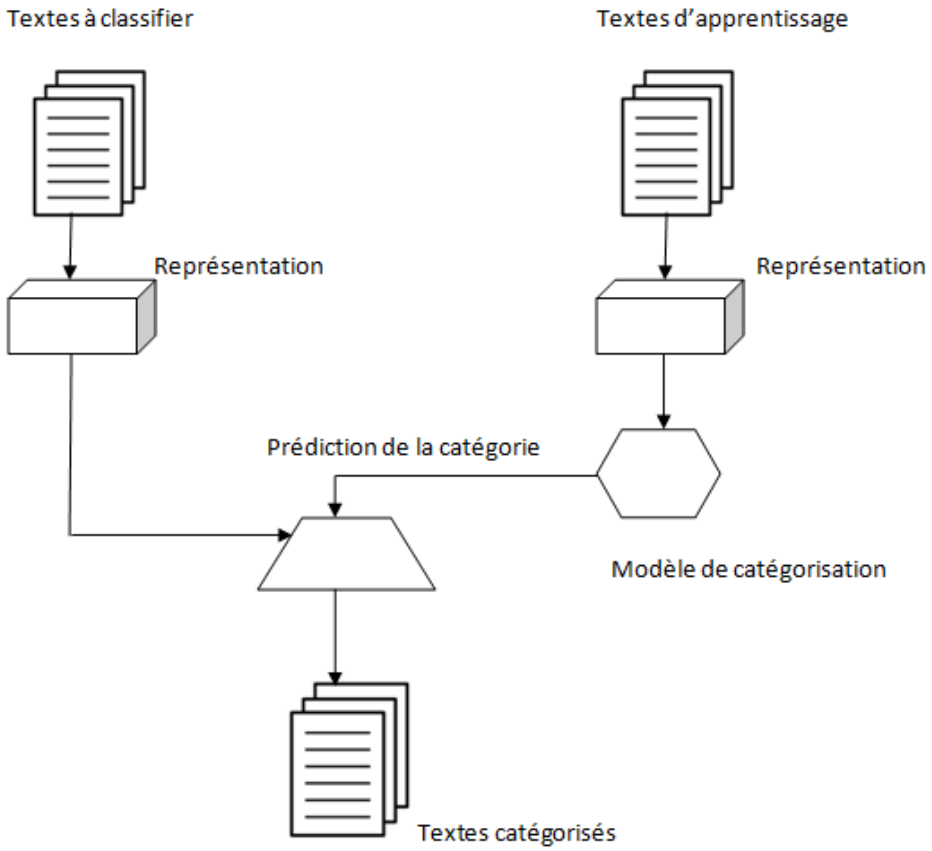


FIGURE 2.11 – Processus de catégorisation de textes

2.5.1 Acquisition et constitution du corpus

C'est une étape cruciale, portant sur la récupération de données textuelles en vue de constituer des corpus de données générale ou spécialisées. Ces corpus ont pour caractéristiques :

- **Un format :** un format de document correspond aux types des données extraites. il peut s'agir d'un format de type Word, Pdf, Xml, Text Brute etc. Des heuristiques spécifiques à chaque format sont souvent définies pour extraire ces textes.

- **Une langue :** identifier la langue des documents extraits.

- **Un encodage :** Un texte est une suite de caractères qui n'a de sens que si l'on connaît son encodage (ASCII, UTF-8, . . .), des erreurs dans la gestion de l'encodage peuvent conduire à des résultats erronés lors de l'utilisation du corpus.

2.5.2 La segmentation et la Tokenisation

la segmentation consiste à diviser un document textuel en unités significatives, telles que des phrases ou expressions. La plupart des outils de segmentation se basent sur les signes de ponctuation (ex. le point final) en tenant compte aussi d'heuristiques pour éviter certains cas d'exception (ex. c.-à-d.).

La Tokenisation consiste à fractionner le texte en des portions (Tokens) encore plus petites que lors de la phase de segmentation précédemment citée. Ces Tokens représentent soit des mots, des chiffres ou des acronymes. Une liste de délimiteurs (", '.', Etc.) est généralement utilisée pour séparer les Tokens entres eux, mais celle-ci peut être personnalisée selon l'application. Cette tâche est essentielle pour transformer les documents textuels non structurés en une forme communément appelé « Sac de mots » plus facile à exploiter par les techniques d'apprentissage automatique. De nombreux outils de tokenisation sont disponibles, tels que Stanford Tokenizer, OpenNLP Tokenizer.Dans [7] Sutton et McCallum ont dressé une introduction complète à ce sujet, notamment à propos des langues qui ne disposent pas de marqueurs explicites de fin de phrases tel que le chinois ou le japonais.

2.5.3 Les Pré-traitements linguistiques.

Des pré-traitements basés sur les techniques du TALN sont souvent nécessaires pour traiter des problèmes liées à la nature non structurée et complexe des données textuelles. Ces pré-traitement permettent d'ignorer certains détails contenus au niveau des textes(ponctuation, majuscules, fautes d'orthographes etc.). On note deux grandes étapes :

- **Élimination des mots vides** Étape préliminaire dont l'objectif est d'identifier les mots insignifiants qui n'apportent pas de sens au texte, puis de les exclure. Il existe plusieurs listes de mots vides pour diverses langues voir même plusieurs listes pour une même langue. On y retrouve des déterminants, des adverbes ainsi que les mots les plus fréquents. En règle générale, les mots qui apparaissent dans plus de 80% des documents d'une collection sont considérés comme inutiles. Ainsi, le fait de les supprimer permet d'économiser beaucoup d'espace lors de la phase d'indexation .
- **Normalisation des mots** : L'objectif de cette deuxième étape est de ramener les mots de la même famille à leur forme normale par :
  - **Lemmatisation** :Le processus de « lemmatisation » consiste à représenter les mots ou « lemmes » sous leur forme canonique. Par exemple pour un verbe, ce sera son infinitif. Pour un nom, son masculin singulier. L'idée étant encore une fois de ne conserver que le sens des mots utilisés dans le corpus.
  - **Racinisation** :(ou stemming en anglais). Cela consiste à ne conserver que la racine des mots étudiés. L'idée étant de supprimer les flexions (suffixes, préfixes) des mots afin de ne conserver que leur origine. C'est un procédé plus simple que la

lemmatisation et plus rapide à effectuer puisqu'on réduit essentiellement les mots contrairement à la lemmatisation qui nécessite d'utiliser un dictionnaire.

2.5.4 La représentation des textes

Un traitement efficace des données textuelles à l'aide du traitement automatique passe inévitablement par une bonne représentation de ces données. Ces dernières doivent être représentées en se basant sur un type d'unités élémentaires (descripteurs) pour être facilement manipulées par les algorithmes d'apprentissage. Cette étape consiste généralement en la représentation de chaque document par un vecteur. Ainsi, le problème peut se définir comme suit :

Définition : Soit un document d_j, sa transformation en un vecteur $v_j = (w_1, w_2, ..., w_T)$, où T est l'ensemble des termes (ou descripteurs) qui apparaissent au moins une fois dans la collection de documents d'apprentissage (corpus). Le poids de pondération w_k correspond à la contribution du terme tk à la sémantique du texte d_j. (Jalam,2003) [8].

On constate à travers cette définition que le choix des descripteurs (caractéristiques) est crucial pour une bonne représentation. Nous distinguons six grandes catégories de caractéristiques utilisées en analyse d'opinion :

Les mots simples : Les termes simples contenus dans les documents après élimination des mots vides.

- Les n-grams : un n-grams est une suite d'un nombre donné de caractères (bigrammes, trigrammes, 4 grams, etc.).L'utilisation de n-grams permet de capter des informations relatives à la syntaxe des cooccurrences.

- Les lemmes/Stemmes : Dans ce cas, on utilise les termes réduits à leur forme canonique (resp. radical) obtenus après l'étape de lemmatisation (resp. racinisation).
- Descripteurs basés sur l'étiquetage grammatical : l'utilisation des étiquettes POS met en valeur certaines caractéristiques liées aux opinions au sein d'un texte. Plusieurs descripteurs basés sur l'étiquetage grammatical ont été utilisées dans des travaux connexes, par exemple, des attributs tenant compte du nombre de noms, de verbes et d'adjectifs[9], on retrouve aussi des attributs comptant le rapport du nombre de noms par rapport aux adjectifs, le nombre de verbes par rapport aux adverbes [10] et le nombre de verbes négatifs obtenus à partir des étiquettes POS.

- Descripteurs lexical : des ressources lexicales supplémentaires telles que les lexiques d'opinions ou SentiWordNet[11] peuvent aussi être utilisés comme attributs. Ces ressources utilisent des connaissances externes pour améliorer les résultats de l'analyse d'opinions.

- L'émoticon comme descripteur : Dans ce cas, la liste des émoticônes positifs et négatifs est répertoriée, puis on recense le nombre d'occurrences de chaque classe d'émoticônes dans le texte [12].

Enfin, une fois que le type de caractéristiques a été choisi, la représentation du texte peut être réalisée sur la base du schéma adopté. L'un des premiers modèles de schéma mais aussi l'un des plus utilisés, est la représentation en « Sac de mots », également appelé Vector Space Model « VSM » [13]. Ce schéma consiste à représenter chaque document par un vecteur «documents X descripteur» de taille fixe dont chaque composante représente un mot (caractéristique) contenu dans les textes. Ainsi, les lignes correspondent aux documents à classer et les colonnes correspondent aux attributs qui apparaissent au moins une fois dans le document. Ce schéma a été largement utilisé dans les travaux d'analyse d'opinion.

Malheureusement, celui-ci présente certaines limites, dus à la dimensionnalité trop importante de la représentation :le nombre de dimensions obtenues à l'aide du VSM sont souvent exorbitants sur l'ensemble du corpus, ce qui conduit à une représentation vectorielle avec des matrices creuses. Pour y remédier, une étape supplémentaire mais élémentaire consiste à faire de la sélection de caractéristique afin d'élaguer l'espace original, sans pour autant compromettre la performance et la précision de classification.

2.6 La sélection de caractéristique

Chercher à réduire la taille d'un ensemble de données devient de plus en plus indispensable en raison de la multiplication des données. Dans de nombreux domaines, le système de résolution d'un problème est fondé sur un ensemble des variables (caractéristiques). L'augmentation du nombre de ces variables (caractéristiques) qui modélisent le problème introduit des difficultés à plusieurs niveaux comme la complexité, le temps de calcul ainsi que la détérioration du système de résolution en présence de données bruitées.C'est la raison pour laquelle une phase de réduction de la dimensionnalité s'avère nécessaire.[3]
Les méthodes de réduction de la dimensionnalité sont généralement classées en deux catégories :

- **L'extraction de caractéristiques** qui permet de créer de nouveaux ensembles de caractéristiques, en utilisant une combinaison des caractéristiques de l'espace de départ ou plus généralement une transformation effectuant une réduction du nombre de dimensions.
- **La sélection de caractéristique** qui est un processus important en analyse de sentiments et qui consiste à identifier et éliminer des caractéristiques redondantes et non pertinentes d'une liste initiale, en utilisant divers critères et différentes méthodes.

Dans ce qui suit nous présenterons certaines méthodes de sélection de caractéristiques.

2.6.1 Les approches et méthodes de sélection de caractéristique

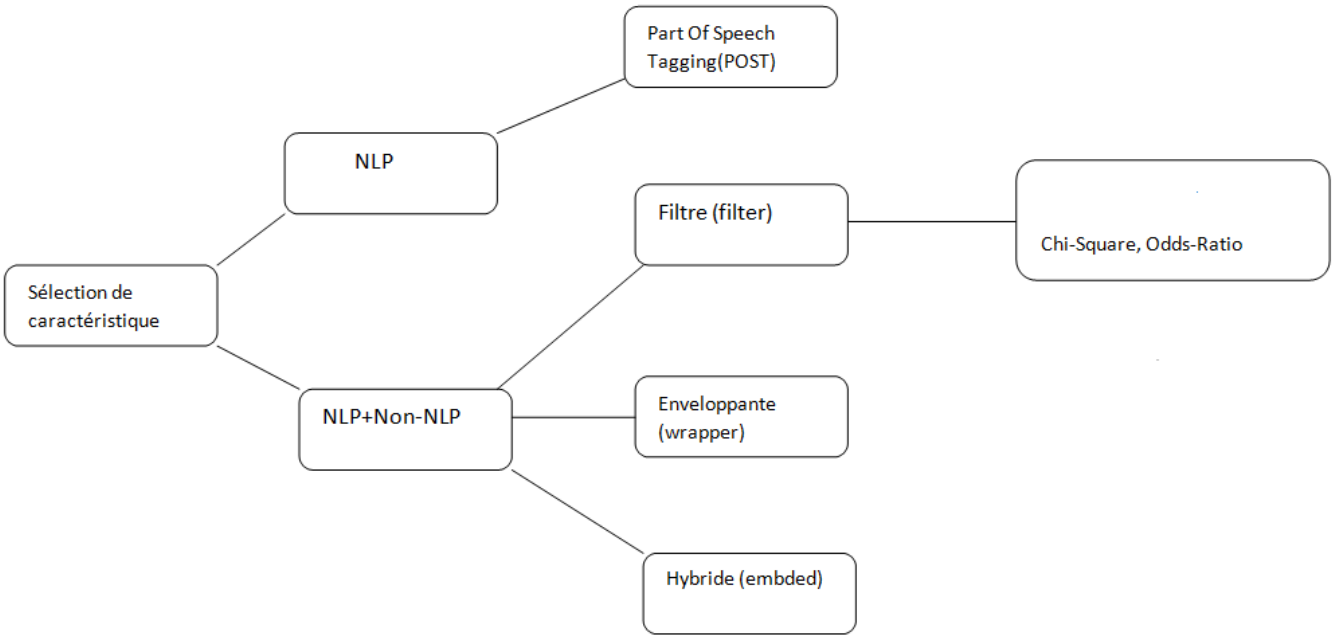


FIGURE 2.12 – Les catégories de sélection de caractéristiques en analyse de sentiments [3]

On retrouve principalement, trois grandes approches de sélection de caractéristiques :

— **Approche filtre (filter) :** Ce type d'approche a été la première à avoir été utilisée en sélection de caractéristiques et elle est indépendante de tout algorithme d'apprentissage automatique. En effet, au cours du processus de filtrage un score correspondant à chaque entité est calculé et les entités ayant un score faible sont ainsi supprimées. Le sous ensemble de caractéristiques résultant devient l'entrée de l'algorithme de classification. Les méthodes issue de ce modèle utilisent souvent une approche heuristique comme stratégie de recherche. La procédure du modèle "filter" est illustrée dans la figure ci-dessous : :

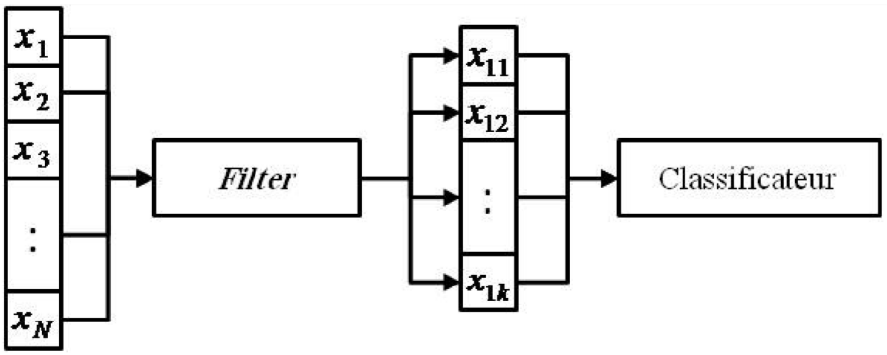


FIGURE 2.13 – La procédure du modèle "filtre" [3]

— **Approche enveloppante (wrapper) :** Le principal inconvénient des approches "filter" est le fait qu'elles ignorent l'influence des caractéristiques sélectionnées sur la

performance du classificateur à utiliser par la suite. Pour résoudre ce problème, [16] ont introduit le concept "wrapper" pour la sélection de caractéristiques. Cette approche utilise l'algorithme de classification comme une fonction d'évaluation, pour définir la pertinence d'un sous ensemble de caractéristique. L'appel de l'algorithme de classification est fait plusieurs fois à chaque évaluation, puis un score est attribué à chaque sous ensemble de caractéristiques retenus. Ce score est généralement un compromis entre le nombre de caractéristiques éliminées et le taux de bonne classification. des attributs par l'intermédiaire d'une prédiction de la performance du système final.

L'approche filtre est généralement plus rapide que l'approche Wrapper en terme de génération de résultats.Cependant, cette dernière à l'avantage de fournir des résultats mieux adaptés à l'algorithme de classification utilisé lors des phases d'évaluation.

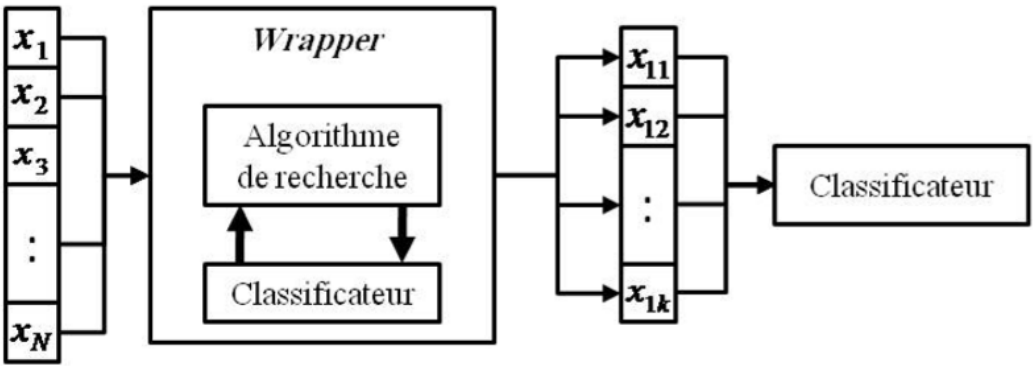


FIGURE 2.14 – La procédure du modèle "wrapper"
[3]

— **Approche intégrée (embedded) :** Ces méthodes incorporent la sélection directement au sein du processus d'apprentissage sans étape de validation, pour maximiser la qualité de la sélection et minimiser le nombre de caractéristiques. En effet, contrairement aux méthodes précédentes qui scindent leur ensemble d'apprentissage en deux afin d'y opérer l'apprentissage et la validation séparément, les méthodes embarquées ont l'avantage de se servir de l'ensemble entier afin de sélectionner les caractéristiques pertinentes, ce qui les rend moins coûteuses en complexité de calcul.

Ci-dessous, nous présentons quelques méthodes largement utilisées en classification de texte

Méthode Chi-Squre x^2 :

Le test de Chi-Square est un test statistique d'indépendance permettant de déterminer la dépendance de deux variables. De la définition de Chi-Square, on peut facilement déduire l'application de la technique du Chi-Square dans la sélection des caractéristiques. Dans [17] Chi mesure le manque d'indépendance entre l'entité et la catégorie, plus la valeur de X^2 est élevée, plus la relation entre la caractéristique et la classe est étroite.

$$x^2(f,c) = N\frac{(df_c * d\bar{f}_{(}\bar{c}) - (df_{(}\bar{c}) * d\bar{f}_c)}{df * d\bar{f} * N_c * N\bar{c}}$$

Méthode Odds-Ratio orr :

Odds-Ratio donne un score positif aux caractéristiques qui apparaissent plus souvent dans une catégorie que dans l'autre, et un score négatif s'il survient plus souvent dans l'autre. Un score de zéro signifie que la probabilité qu'une caractéristique apparaisse dans une catégorie est exactement la même que celle d'une autre catégorie.

$$orr(f,c) = log\frac{p(f|c) * (1 - p(f|\bar{c}))}{p(f|\bar{c}) * (1 - p(f|c))}$$

* Ces notations seront définie dans la section "Phase de sélection de caractéristiques".

Le tableau ci-dessous résume les inconvénients et les avantages des approches de sélection de caractéristiques [5] :

	Avantages	**Inconvénients**
Approche filtre	— Rapidité — généricité	— Non relié aux caractéristiques de la méthode, rien de dit que les variables ainsi sélectionnées seront les meilleures
Approche enveloppante	— relié à un critère de performance	— Lourdeur des calculs — danger de sur-apprentissage — non connecté avec les caractéristiques intrinsèques de la méthode (ex. max de la marge pour les SVM).

2.7 Mesures d'évaluations de l'analyse d'opinion

Plusieurs méthodes permettent de valider (ou d'infirmer) la valeur d'un processus d'apprentissage. Une des approches consiste à n'utiliser qu'une partie des données pour apprendre et à se servir des autres données pour tester le résultat telles que la précision, le rappel, la F-mesure. Ces mesures sont essentiellement utilisées en apprentissage supervisé et non supervisé.

— **La précision :** Est définie comme étant la probabilité conditionnelle qu'un exemple choisi aléatoirement soit bien classé par le système. Il s'agit du rapport entre le nombre de bonnes prédictions positives (solutions pertinentes) et le nombre de prédictions

positives (vraies et fausses) . Elle mesure donc la capacité du système à refuser les solutions non-pertinentes. La précision exprime le ratio entre le nombre d'exemples d'entraînements correctement lassés dans la classe (C) sur le nombre total d'exemples d'entraînements auxquels la classe (C) est assignée, selon l'équation suivante :

$$\textbf{Précision}=\frac{Vp}{Vp+Fp}$$

Où : Vp : le nombre de prédiction vraie positive.
Vn : le nombre de prédiction vraie négative.
Fp : le nombre de prédiction fausse positive.
Fn : le nombre de prédiction fausse négative.

— **Rappel** : Le rappel mesure la largeur de l'apprentissage. Il correspond au rapport entre le nombre de bonnes prédictions positives et le nombre total d'exemples d'entraînement. Il mesure la capacité du système à donner toutes les solutions pertinentes.

Le Rappel exprime le ration entre le nombre d'exemples d'entraînements correctement classés dans la classe (C) sur le nombre total d'exemples d'entraînements appartenant àla classe (C), selon l'équation suivante :

$$\textbf{Rappel}= \frac{Vp}{Vp+Fn}$$

— **F-score** : Est définie comme la moyenne harmonique de la précision et du rappel

$$\textbf{F-score}=2*\frac{(precision*Rappel)}{precision+Rappel}$$

Le F-score rend compte de la qualité d'une classification en fonction des classes, mais ne tient pas compte de l'éventuel déséquilibre entre les classes .

2.7.1 La validation croisée :

Certains classifieurs comme les arbres de décision, les SVMs et les régressions à noyau sont souvent sujettes au phénomène de sur-apprentissage. Ainsi, en évaluant les indicateurs de performances sur les données d'entraînement, on trouve une estimation largement optimiste des performances du classifieur.
il existe une solution systématique pour se rendre compte qu'un modèle sur-apprend **: *la validation croisée*** est une technique d'évaluation des modèles d'apprentissage-machine via la formation de plusieurs modèles d'apprentissage-machine sur des sous-ensembles des données d'entrée disponibles et via leur évaluation sur le sous-ensemble complémentaire des données ,elle est utilisé sous différentes variantes.

- **Division entraînement test :** La version la plus simple de la validation croisée consiste à diviser le jeu de données en deux sous-ensembles : l'ensemble d'entraînement et l'ensemble de test.On entraîne notre algorithme sur le premier sous-ensemble de données. Ensuite, on compare les indicateurs de performances 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, notre algorithme sur-apprend et il peut valoir le coup de retoucher le modèle pour améliorer les performances sur l'ensemble de test.Sinon, il faut penser à

augmenter la complexité du modèle afin d'obtenir de meilleures performances.

- **Validation croisée à k plis :**L'idée est d'aller plus loin en divisant l'historique en k sous-ensembles aussi appelés plis. On procède à l'apprentissage sur un ensemble et au test sur les k-1 ensembles restants, et ce k fois. On compare ensuite les moyennes des indicateurs sur les ensembles d'entraînement et de test pour savoir si le modèle sur-apprend.
- **LOO (Leave One Out)** :C'est un cas particulier de la variante 2 où le nombre de sous-ensembles k est égal au nombre d'échantillons n.

2.8 Travaux connexes

Dans cette partie, nous présenterons une synthèse des différents travaux autours de l'analyse d'opinions. On s'intéressera principalement à ceux de la tâche de classification de polarité. On discutera également des travaux ayant traités de la sélection de caractéristiques.

2.8.1 Travaux basées sur l'apprentissage automatique

Pang et Lee

Pang et Lee [18] ont été les pionniers dans les approches d'apprentissage automatique pour l'analyse d'opinions de critiques cinématographiques. Pour les expériences, ils ont recueilli des critiques à partir de IMDb.com [1] . Plusieurs types de caractéristiques ont été expérimentés à l'aide de trois algorithmes de de classification à savoir. NB, Entropie maximum (MaxEnt) et SVM . Il s'avère que SVM a donné les meilleurs résultats avec une précision de 82,9% en utilisant les Unigrams. Ils affirment que l'utilisation des techniques d'analyse du discours et de co-références a nettement aidé à l'amélioration de la précision.

Dang et al

Dans [19], l'analyse d'opinions a été réalisée à l'aide des SVM et du Gain d'information (IG) comme méthode de sélection de caractéristiques. Les expériences ont été effectuées sur deux corpus, le premier comprend 305 critiques positifs et 307 critiques négatives à propos d'appareils photo numérique. Le deuxième corpus ensemble de données multi-domaine a été proposé par Blitzer [20]. Trois variantes de caractéristiques ont été utilisé pour l'apprentissage SVM (dépendantes au domaine, indépendant au domaine et caractéristiques opiniâtres). Les caractéristiques obtenues après sélection ont réalisé de meilleurs résultats de classification (précision de 84,15%) sur l'ensemble de données multi-domaine que sur celui concernant l'appareil photos.

Tan et al

Tan et al [21] ont proposé une approche automatique pour l'extraction de règles à base de concepts subjectifs afin de détecter la polarité des opinions au niveau des phrases. Des règles

1. http ://www.imdb.com/

séquentielles de classes (Class sequential rules) ont été utilisées pour apprendre automatiquement des modèles de dépendance typés, ces modèles associent en outre les dépendances aux relations grammaticales, telles que le sujet ou l'objet indirect. Les expérimentations ont été réalisées sur le corpus de Pang et Lee, le taux de f-mesure obtenu est estimé à 85,37% en utilisant la sélection de caractéristiques sur les modificateurs d'adjectifs. Dans [22], les auteurs comparent les performances de trois types de techniques (le Bagging, Boosting et le Random Subspace) basées sur les ensembles de classificateurs (EoC). Aussi, cinq algorithmes de classifications, à savoir, NB, MaxEntont, le k plus proche voisins, SVM et les arbres de décisions, ont été utilisés pour évaluer les performances de ces trois techniques sur l'analyse d'opinions. Dix ensembles de données ont été étudiés pour vérifier l'efficacité de l'apprentissage à l'aide des ensembles de classificateurs. Sur la base de 1200 expériences comparatives, les résultats révèlent que les approches EoC améliorent considérablement la performance de la phase d'apprentissage pour la classification d'opinions notamment en utilisant celle du Random Subspace. Cependant, on constate que ces techniques sont rarement utilisées en analyse d'opinions en raison, peut-être, de leur forte complexité de calcul.

2.8.2 Travaux liés à la classification de polarité

Considérée comme l'une des tâches les plus étudiées en analyse d'opinions, la classification de polarité consiste à déterminer le type d'opinions exprimé dans un texte ou dans une phrase supposée subjectif. La tâche est habituellement considérée comme binaire (positif contre négatif), toutefois, l'analyse peut être étendue sur un axe plus large : ternaire (positif vs. négatif vs. neutre) ou ordinal (ex. en utilisant un classement en étoile). Mais théoriquement, les travaux autours de la polarité peuvent aussi se scinder en fonction du niveau de granularité de l'analyse. Cette granularité dépend de l'unité de texte à analyser (document, phrases ou portions de phrases). La figure ci-dessous illustre le compromis entre ces différentes catégories.

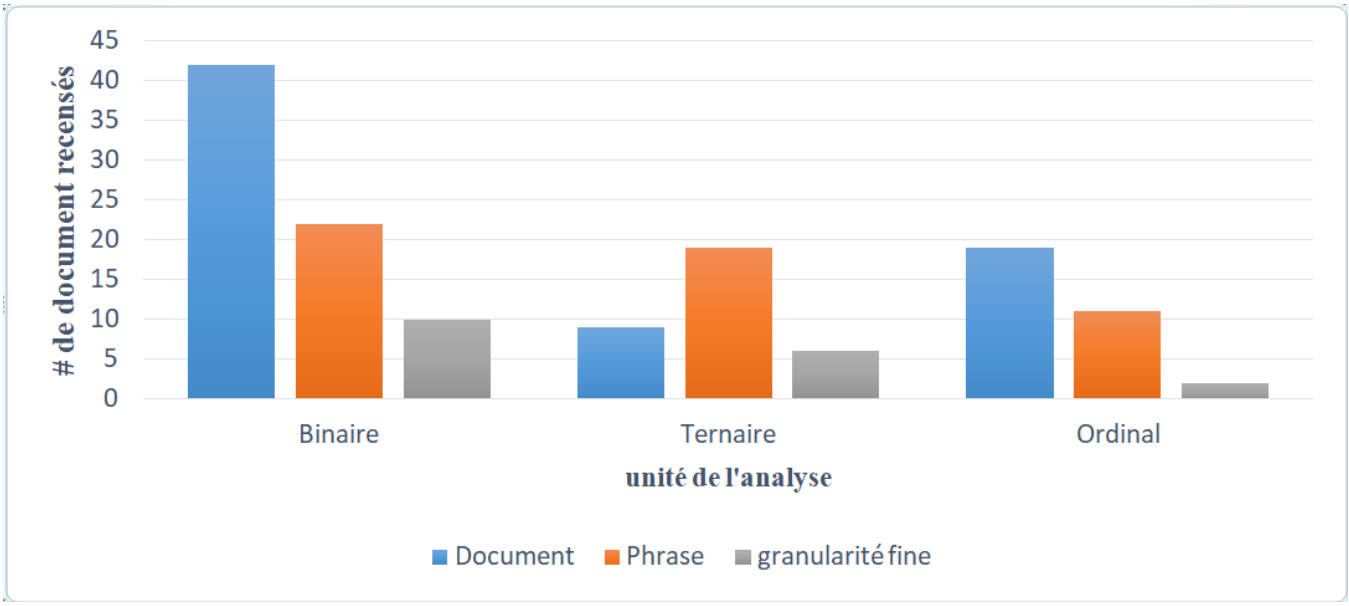


FIGURE 2.15 – Aperçu des travaux existants selon la granularité de l'analyse [42]

2.8.3 Travaux au niveau du document

Classification binaire de la polarité : A travers la figure 2.11, on voit clairement que la majorité de travaux se focalisent sur la classification binaire. L'objectif de cette tâche est de déterminer si le ton global d'un document est positif (recommandation) ou négatif (désapprobation). Les premiers et les plus influents travaux ayant abordés la tache sont ceux de Pang [18].

L'engouement pour ce type de classification s'explique certainement par le nombre de données d'apprentissage disponible et adapté pour la tâche. En effet, plusieurs ensembles de données (annotés en positifs et négatifs) ont été collectés pour de nombreux cas d'applications à l'aide des techniques du Crowdsourcing. Par exemple, dans le cas de l'extraction des avis de clients, les corpus réservés à la classification de polarité binaire, au niveau du document peuvent facilement s'obtenir via les évaluations fournies par Amazon. Pour n'en nommer que quelques-uns, le corpus de Pang [18] est à titre d'exemple obtenu en collectant parmi un ensemble de critiques cinématographique issues de la base de données IMDb. Depuis, des corpus similaires ont été crawler par [23] et comprennent respectivement 320000, 7000 et 40000 documents et rapports clientèles. Pour le français on retrouve DEFT'07 avec 28 000 débats parlementaires autours de l'environnement [24] ou encore celui de Thomas avec 3000 débats traduit du Congrès américain.

Plusieurs travaux ont émergés à partir de l'exploitation des corpus ci-dessus. On retrouve par exemple, Xia [54] où un ensemble de caractéristiques et algorithmes d'apprentissage machine ont été utilisés pour la classification de polarité. Deux types de caractéristiques, à savoir, celles basées sur l'étiquetage grammatical (POS) où les relations entre mots (word-relation based feature sets). Naïve Bayes, l'entropie maximale et les SVM ont été choisis comme algorithmes de classification en utilisant trois sortes de techniques à savoir, la combinaison fixe (fixed combination), la combinaison pondérée et combinaison de méta classifieurs (meta-classifier combination). La classification pondérée basée sur la relation de mots a obtenu les meilleurs résultats de classification de polarité avec une précision de 87,7% (resp. 85,15%) sur le corpus de Pang (resp. Blitzer). La méthode de sélection de caractéristiques basée sur les dépendances syntaxiques à fortement améliorer les taux de précision obtenus.

Khan et al

Khan dans [25], présente une approche hybride pour la classification de polarité sur Twitter. L'approche proposée comprend différentes étapes de prétraitement avant la phase de classification. Les résultats montrent que la technique proposée fournie de bons taux de précision par rapport à des techniques similaires, notamment en présence de données bruités (sparsity data) et expressions de sarcasme.

Classification ternaire de la polarité

La tâche de classification binaire de polarité, suppose qu'un document comporte des opinions majoritairement positives ou négatives. Ceci est généralement valable pour le domaine des évaluations de produits, mais ce n'est toujours pas le cas pour d'autres types de domaine

(ex. l'analyse d'articles de journaux, blogs etc.). En effet, ces documents peuvent ne pas être subjectifs ou peuvent contenir des opinions mitigées sans pour autant révéler clairement un point de vue positif ou négatif. Pour remédier à la problématique, deux stratégies de recherche s'imposent La première consiste à trier les documents, de l'ensemble d'apprentissage, qui ne correspondent pas entre eux en genre et par domaine. A titre d'exemple, nous pouvons procéder d'abord au tri des documents selon deux critères « documents semblables » et « documents sans critique », puis procéder à une classification binaire de la polarité sur la première catégorie (documents semblables). Cette démarche a été étudiée par Barbosa [26] et Frasincar [27]. La deuxième stratégie, plus générale, consiste à introduire une troisième catégorie «neutre» ou «objectif».

L'analyse s'effectue, en apprenant un modèle de classification multi-classe ou en utilisant une approche en cascade. L'approche consiste à former un classifieur de subjectivité en plus du classifieur binaire de polarité. Les documents ainsi identifiés comme subjectifs seront transmis au classifieur de polarité, puis les documents non classés seront considérés comme neutre. Les classifieurs pour la subjectivité et la polarité peuvent être formés sur des corpus différents, cela n'influe pas sur le résultat final. Cette approche a été présentée dans les travaux de Koppel et Schler ainsi que Das et Chen . La Figure tirée de [28] illustre parfaitement l'approche.

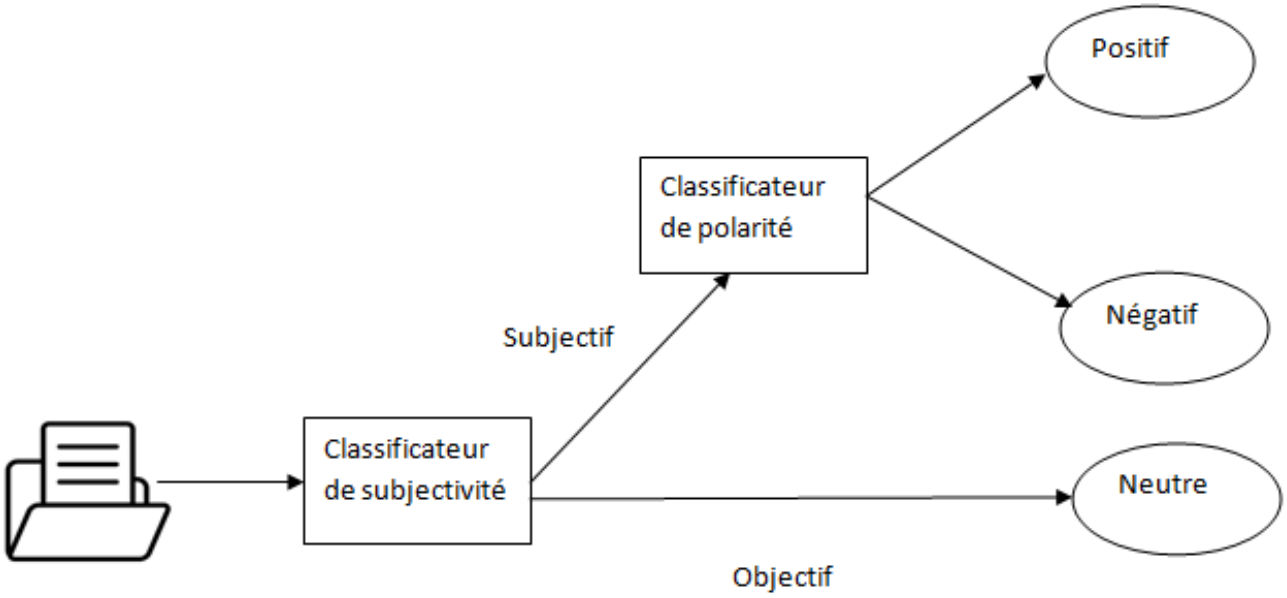


FIGURE 2.16 – Classification ternaire de la polarité

2.8.4 Travaux au niveau de la phrase

La classification de polarité au niveau de la phrase permet d'effectuer des analyses plus fines. Avec ce niveau de granularité, des tâches plus générales telles que l'analyse d'opinions des caractéristiques de produits (aspect-based review mining) ou encore les systèmes de questions réponses basés sur les opinions (sentiment aware question answering). Cependant, comparé à la catégorie précédente, beaucoup moins de travaux ont été publiés pour

ce niveau de granularité. Ceci est probablement dû au manque d'ensembles de données appropriées pour la tâche. Aussi, la tâche est éventuellement plus complexe vu qu'à ce niveau les informations sont peu nombreuses. En effet, les phrase contiennent moins de mots ce qui rend la représentation vectorielle des caractéristiques résultantes assez clairsemés (en anglais, Sparce). Par conséquent, la phase d'extraction de caractéristiques joue un rôle très important à ce niveau de granularité. Le travail de Klakow montre clairement l'impact qu'a un changement ou une diversification de caractéristiques sur la tâche[29].

Comme pour la classification au niveau des documents, nous pouvons scinder les différents travaux en fonction du type de la tâche à opérer. Par exemple pour la classification binaire les travaux de[30] et pour une classification ternaire nous retrouvons [31]. Une approche en cascade combinée à la détection de subjectivité a été proposée par IrynaGurevych (IrynaGurevych) et Hatzivassiloglou [32]. Par contre, beaucoup d'autres travaux modélisent la classe « neutre » directement en apprenant un classifieur multi-classes. Toutefois, nous n'avons répertorié aucun travail portant sur une échelle de polarité ordinal.

2.8.5 Travaux concernant la Sélection de Caractéristiques

Plusieurs méthodes de sélection de caractéristiques ont été proposées en analyse d'opinions, à savoir, Le Gain d'information, l'information mutuelle, le Chi square ou encore la sélection basée sur la fréquence [33]. Un travail assez simpliste a été proposé dans [4]. Celui-ci s'est appuyé sur l'aspect fréquentiel des termes (Document Frequency DF) où l'auteur a exploité les termes les plus fréquents dans le corpus pour en déduire les éléments les plus dominants au sein du vecteur de caractéristique. Tan et Zhang [34] ont expérimenté quatre méthodes de sélection pour l'analyse d'opinions (IG, MI, X2 et DF) sur des documents en chinois à l'aide de cinq algorithmes d'apprentissage automatique, Les résultats montrent que les meilleurs taux de classification ont été obtenus grâce à l'utilisation du gain d'information comme méthode de sélection. Abbasi [35] a montré qu'après avoir utilisé conjointement l'IG et les algorithmes génétiques pour sélectionner ses caractéristiques, cela lui a permis d'obtenir de nettes améliorations pour sa tâche d'analyse d'opinion sur un corpus de critiques cinématographiques. Il a également proposé un Framework basé sur un algorithme génétique pondéré grâce un calcul d'entropie (EWGA). Nicholls and Song [36] ont proposés une nouvelle méthode de sélection, nommée « Document Frequency Difference ». La méthode a été évaluée et comparée à une série de méthodes de sélection de caractéristiques. Gamon et al[37] ont également proposé d'utiliser la méthode de l'estimateur du maximum de vraisemblance (log-likelihood ratio) pour sélectionner les caractéristiques pertinentes pour l'analyse d'opinions.

Agarwal et Mittal [33] proposent une nouvelle méthode de sélection (Categorical Probability Proportion Difference - CPPD). Cette dernière combine deux méthode de sélection, la première exclusivement basée sur les notions de probabilité catégorielle (Probability Proportion Difference - PPD), la second méthode (Categorical Proportional Different - CPD) est proposée pour la catégorisation de texte [38]. CPD mesure le degré avec lequel un terme contribue à discriminer les différentes classes du problème. Puis les principaux termes discriminant sont sélectionnés pour la classification. Tandis que PPD sert à mesurer le degré d'appartenance ou la probabilité qu'un terme appartienne à une classe prédéfinie. L'intérêt

du procédé CPD est qu'il mesure pour un terme particulier, son degré de distinction entre les classes. Cette notion est importante car elle permet d'éliminer les termes qui occurrent conjointement et de manière égale dans les deux classes d'étude. Ceci facilite l'élimination des termes peu discriminants et dont les fréquences d'apparition sont extrêmement élevées. Ce qui est généralement le cas avec les mots vide (stopword).

Wang [22] propose une méthode innovante basé sur le discriminant de Fisher. Cette mesure a fait nettement améliorée les résultats obtenus en comparaison avec d'autres méthodes de sélection. Duric et Song [39] ont proposé modèle de contenu et de syntaxe afin de pouvoir séparer les entités en relation avec les opinions (c'est-à-dire les modificateurs de polarité). Leurs résultats ont montré que l'utilisation de cette caractéristique avec un classificateur d'entropie maximal fournit des résultats compétitifs comparés à ceux présentés dans l'état de l'art.

O'keefe et Koprinska [40] ont introduit deux nouvelles méthodes de sélection, l'une se repose sur un score de subjectivité tiré à partir de SentiWordNet (SWNSS), capable de distinguer les termes objectifs et subjectifs, et la seconde SentiWordNet Proportional Difference (SWNPD) est capable d'opérer une discrimination de classe avec une prise en compte du contexte pour un meilleur résultat de sélection. Le travail présenté dans [41] consistait à d'abord épurer les termes sémantiquement moins importants et discriminants en se basant sur un score sémantique extrait de SentiWordNet [11], par la suite une étape de sélection à l'aide du Gain d'information permet de sélectionner les caractéristiques les plus importantes pour une meilleure précision de classification.

Conclusion

Dans ce chapitre, nous avons tout d'abord introduit quelques notions appartenant au domaine de la fouille de texte. Nous avons dressé un état de l'art de différentes approches d'apprentissage automatique(supervisée et non supervisée). Ensuite nous avons détaillé les principales étapes de catégorisation de texte ainsi que les métriques d'évaluation,commune à l'analyse d'opinions. Enfin, nous avons terminé avec une synthèse des divers travaux autours de l'analyse d'opinions,Dans le prochain chapitre , nous allons présenté la démarche suivi et l'approche proposé afin de répondre à la problématique posée.

Chapitre 3

Description et présentation de la solution

Introduction

La recherche accorde ces dernières années, beaucoup d'importance au traitement et à la classification des données textuelles pour plusieurs raisons : le nombre croissant de collections mises en réseau et distribuées au plan international, le développement de l'infrastructure de communication et de l'Internet. Les traitements manuels de ces données s'avèrent très coûteux en temps et en personnel, ils sont peu flexibles et leur généralisation à d'autres domaines est presque impossible ; c'est pour cela que l'on cherche à mettre au point des méthodes automatiques.
Le domaine de la fouille de textes (texte mining) s'est développé pour répondre à volonté à la gestion par contenu des sources volumineuses de textes. À l'heure actuelle, de nombreux logiciels de classification de textes sont disponibles, ils ont fait l'objet de publications et leurs champs d'applications s'élargissent de jour en jour.
Dans ce mémoire, nous abordons le problème de détection de la polarité des opinions comme un problème de classification de textes en deux catégories : la classe des textes qui expriment des opinions positives, et la classe des textes qui expriment des opinions négatives. Il est donc important de bien comprendre ce qu'est la classification de textes avant de traiter le problème de détection des polarités.
Dans ce chapitre, nous exposons quelques techniques de prétraitements et de représentations pour la classification des polarités. Ensuite, nous proposons un modèle à base du naïf Bayes et des Machines à Vecteurs de support . Enfin, nous terminons par une conclusion.

3.1 Motivations et objectifs

La classification de textes a pour objectif de regrouper les textes similaires, c'est-à-dire thématiquement proches, au sein d'un même ensemble. L'intérêt d'une telle démarche est d'organiser les connaissances de façon à pouvoir effectuer par la suite une recherche ou une extraction d'information efficace.
Pour détecter la polarité exprimée dans un texte, il faut pouvoir le classifier en tant que texte qui exprime une opinion positive ou négative. Pour ce faire, nous devons d'abord réaliser le prétraitement de textes bruts, puis leurs donner une bonne représentation pour qu'ils deviennent exploitables par les modèles. Enfin, nous comptons construire un modèle à base du naïf Bayes et des Machines à Vecteurs de support pour faire la classification[B1].
Comme nous l'avons vu dans le chapitre précédent, le processus de catégorisation de texte nécessite plusieurs étapes commençant par l'acquisition du corpus, le prétraitement qui inclut plusieurs tâches et qui permettent d'extraire un vocabulaire à partir des textes brutes du corpus. Ce vocabulaire est constitué d'un ensemble varié de termes qui ne sont pas tous pertinent pour une bonne catégorisation de texte. Dans ce contexte, il nous a été demandé d'améliorer les résultats obtenus durant le processus précédent de façon à extraire les termes les plus discriminants pour une meilleure classification.

3.2 Outils et environnement de développement

Pour les besoins de notre approche, nous avons opté pour certaines API largement utilisées en traitement de données. Ainsi, en ce qui concerne le traitement du langage naturel , nous avons

opté pour la librairie StanfordCoreNLP. Pour ce qui est de la phase d'apprentissage automatique nous avons choisi de travailler avec l'API Weka ainsi que la librairie LibSVM pour la prise en charge de l'algorithme SVM au sein de WEKA. Une brève présentation de ces outils s'impose :

API Weka 3.8.0

Weka est un logiciel d'apprentissage automatique développé à l'université Waikato, en Nouvelle-Zélande , l'outil est développé en Java standard et permet de réaliser des expériences d'apprentissage machine et d'intégrer des modèles formés dans des applications Java.Il peut être utilisé pour un apprentissage supervisé et non supervisé. La bibliothèque de Weka fournit une vaste collection d'algorithmes d'apprentissage machine, implémentés en Java.Les algorithmes peuvent être appliqués directement à un jeu de données ou appelés à partir de notre propre code Java. Weka contient des outils pour le prétraitement, la classification, la régression, le regroupement, les règles d'association et la visualisation des données. Il se compose principalement :

- de classes Java permettant de charger et de manipuler les données
- de classe pour les principaux algorithmes de classification supervisée ou non supervisée.
- D'outils de sélection d'attributs, de statistiques sur ces attributs.
- de classe permettant de visualiser les résultats.

On peut l'utiliser à trois niveaux :

- Via l'interface graphique, pour charger un fichier de données, lui appliquer un algorithme, vérifier son efficacité
- Invoquer un algorithme sur la ligne de commande.
- utilisé les classes définies dans ses propres programmes pour créer d'autres méthodes, implémenter d'autres algorithmes, comparer ou combiner plusieurs méthodes via L'API Weka.

Dans le cadre de notre travail, Nous avons utilisé L'API Weka principalement pour certaines parties de la phase préliminaire de traitement et préparations de données textuelles. La phase de classification (apprentissage et test) a été intégralement réalisée à l'aide de l'api Weka.
Cependant, afin d'opérer les séries de prés-traitements et de préparation des données (corpus), nous avons défini notre propre filtre indépendamment de celui inclus dans Weka (StringToWordVector). Ceci nous permet d'affiner au préalable notre phase de sélection et convertir à notre convenance les attributs de chaîne, en un ensemble d'attributs représentant des informations d'occurrence de mots à partir du texte contenu dans les chaînes. Nous intéressons aux termes qui se produisent et au nombre de fois où chacun se produit. En outre, nous associons et transformons tous les déclinaisons N. Grams ,Bigrams de façon à éviter les erreurs.

StanfordCoreNLP 3.9.2

Stanford Core Nlp : un analyseur en langage naturel qui fournit un ensemble d'outils technologiques en langage humain. Il peut donner les formes de base des mots, leurs parties du discours, normaliser les dates, les heures et les quantités numériques etc.Stanford CoreNLP fournit :

- une boîte à outils intégrée de la PNL avec un large éventail d'outils d'analyse grammaticale.

— un annotateur rapide et robuste pour les textes arbitraires, largement utilisé en production.
— un package moderne, régulièrement mis à jour, avec une analyse de texte de la plus haute qualité.
— Prise en charge d'un certain nombre de langues (humaines) majeures.

L'objectif de Stanford CoreNLP est de faciliter l'application de nombreux outils d'analyse linguistique à un texte. Un pipeline d'outils peut être exécuté sur un morceau de texte brut avec seulement deux lignes de code. CoreNLP est conçu pour être extrêmement flexible et extensible. Avec une seule option, on peut modifier les outils à activer ou à désactiver. Stanford CoreNLP intègre de nombreux outils de PNL de Stanford, notamment l'indicateur de partie de parole (POS), l'identificateur d'entité nommée (NER), l'analyseur. De plus, un pipeline d'annotateurs peut inclure des annotateurs personnalisés ou tiers supplémentaires. Les analyses de CoreNLP fournissent les éléments de base des applications de compréhension de texte de niveau supérieur et spécifique à un domaine.[3]

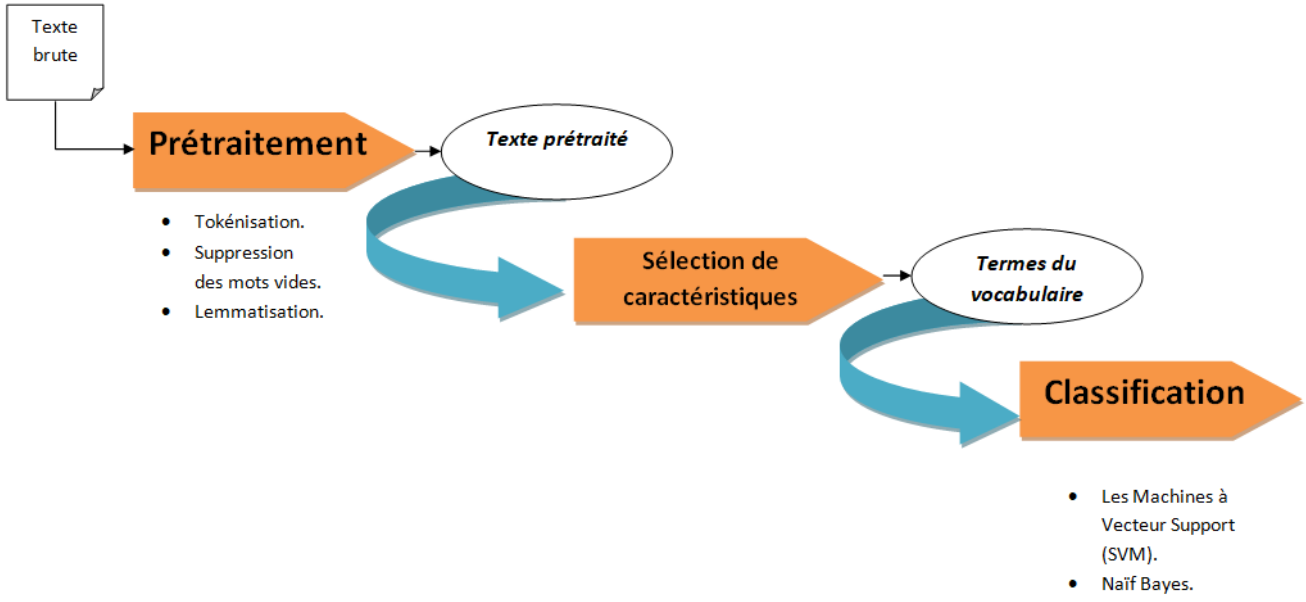
Librairie LibSVM de l'API Weka

La librairie LibSVM fournit une classe très spécifique d'algorithmes de classification Supports de Machines à Vecteurs(SVM) ;

```
LibSVM classifier = new LibSVM() ;
classifier.setGamma(0.1) ;
classifier.setCoef0(0.12) ;
classifier.buildClassifier(data) ;
```

3.3 Description de la chaîne de traitement

Notre travail a été scindé en trois étapes essentielles, le schéma ci-dessous illustre l'enchaînement de ces dernières.



- **Le prétraitement :** Le prétraitement du texte est une phase critique dans le processus d'analyse des sentiments. L'objectif était d'extraire des variables (sous forme de mots ou de séquences de mots) pour les utiliser dans la classification. La qualité du traitement a un impact majeur sur les performances des modèles de classification et les résultats obtenus à la fin du processus. Dans notre cas ce processus commencé par un nettoyage et une normalisation du texte : suppression des signes, des symboles, des lettres répétées, des mots vides et tous les mots qui ne fournissent aucune information sur le sujet étudié. La tâche suivante est l'opération de tokénisation par laquelle le texte est divisé en unités lexicales(tokens). Dans un texte, ces unités sont plus complexes puisqu'elles sont composées souvent de plus d'un mot, d'où l'importance de la tâche de lemmatisation ou racinisation. Pour appliquer une lemmatisation sur les mots collectés, nous avons utilisé la librairie Stanford NLP qui offre plusieurs méthodes de traitement de texte, Parmi ces méthodes nous avons utilisé "LemmaAnnotation" qui produit la forme canonique des mots en éliminant des préfixes et suffixes les plus courants d'un token.
- **La sélection de caractéristiques :** Durant cette étape, on s'attachera à structurer les caractéristiques qui serviront pour la phase d'apprentissage. l'instance utilisée pour les besoins de cette tache est définie comme ceci (texte, vocabulaire, label de classe) ou
  - texte : représente le texte de notre corpus à l'état brute.
  - vocabulaire : représente l'ensemble des tokens extrait après l'étape de prétraitement.
  - label de classe : définit l'étiquette du texte brute(positif ou négatif).

  A ce niveau, nous faisant correspondre aux instances de l'API weka, des pondérations que nous détaillerons ci-dessous, à savoir delta tf-idf,Odds Ratio ,Chi-Square.L'ensemble de ces instances sera transformé en instance globale qui permettra la génération d'un fichier Arff utilisé comme support de travail pour la phase d'apprentissage.
- **La classification :** Le processus de catégorisation de la polarité des textes intègre la construction d'un modèle de prédiction qui reçoit en entrée un texte et qui lui associe en sortie, une ou plusieurs étiquettes. Le déroulement d'une classification de textes suit quatre étapes principales résumées dans la figure ci-dessous :

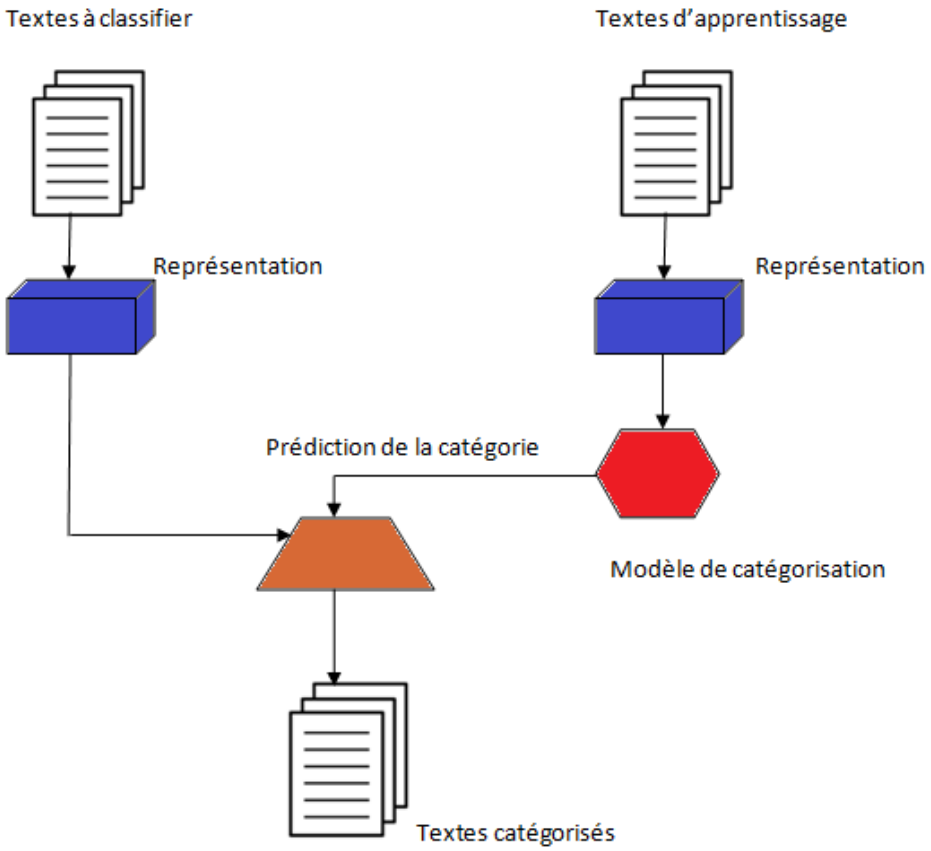


FIGURE 3.1 – La classification de textes

Deux classifieurs ont été utilisés pour la classification supervisée : un classifieur fondé sur les Machines à vecteur support (SVM) et un classifieur probabiliste Naïf Bayes. Le choix a porté sur ces deux types de classifieurs vus les bons résultats obtenus dans de nombreux travaux en relation avec l'analyse d'opinions.

*

3.3.1 Description du corpus d'étude

Pour évaluer notre approche, nous utilisons un corpus de critiques de films « Polarity dataset ». C'est une archive du groupe rec .arts. Films reviews, qui ont été publiés par « Pang et Lillian » et disponible en ligne à http ://www.cs.cornell.edu/people/pabo/-movie-review-data/. Les auteurs se sont basés uniquement sur des exemples de critique fiables (critiques avec des étoiles ou une valeur numérique) extraient à partir de la base de données internet movie Database (IM Db). Ils ont converti et classé les notations en trois catégories : positive, négative et neutre. Pour ce corpus, ils se sont concentrés sur les déclarations positives et négatives supposées discriminantes. La version la plus récente de l'ensemble de données est « Polarity dataset V2.0 » (PLII). Cette version comprend 1000 critiques positives et 1000 critiques négatives. Celui-ci est l'un des corpus les plus utilisés et le plus typique à la tâche d'analyse de polarité au niveau des documents. Ci-dessous, un exemple de critiques extraites à partir du corpus :
"Numerous comparisons can be made with this movie To past sci-fi, suspense thrillers. Soldier is a multicrossbreed between the likes of Terminator, Aliens and offspring. The problème with such

mixed genes is that the final product is a réal mongrel not well made and should have been put down before production got off the ground. Besides this, the action is mediocre when compares To the standard action flicks of this day and age. This movie has not made me change my opinion about director Paul Anderson, whose last epic Évent Horizon has left an unusually bitter taste in my mouth. Although this movie does not comme anywhere close To the strangeness of former, it is stilla long way from anything considered desirable."

Dans l'exemple, l'auteur de la critique commence par situer le contexte du film « Soldier », ensuite, il commence à y exprimer son avis (évaluation) à propos des scènes de combat, des personnages.L'évaluation mentionne également "Paul Anderson", le réalisateur, avec une brève évaluation de sa carrière. Enfin, l'auteur conclu que le film est "loin d'être souhaitable" ce qui correspond donc à un avis négatif. On comprend à travers cet exemple, que le texte est constitué d'un ensemble de phrases subjectives et objectives. Ainsi la difficulté de la tâche de classifications est de pouvoir trouver un compromis entre ses différentes phrases pour en déduire l'opinion globale.

Le tableau ci-dessous décrit les statistiques du corpus :

	Positif	Négatif
# documents	1000	1000
# minimum de mots dans le document	16	134
# maximum de mots dans le document	2253	2755
Moyenne de mots dans un document	721	803
Total de mots dans l'ensemble des documents	721 257	803 117

Table 3.1 – Statistique du corpus

3.4 Description de l'approche

Pour appliquer un quelconque algorithme de classification de texte, chaque document doit être représenté à l'aide d'un ensemble de termes prédéfinie. Un des modèles les plus répandus en

recherche d'informations est celui du bag of Words, ou l'ordre des termes dans le document est littéralement ignoré et chaque document est représenté par le nombre d'occurrences d'un terme dans le vocabulaire. Dans ce qui suit, nous utiliserons souvent l'appellation « Caractéristiques » pour faire référence au terme du vocabulaire. Cependant, plusieurs paramètres peuvent influer sur les performances d'un système de classification. Les caractéristiques extraites à partir des textes afin de représenter ces derniers peuvent être considérées parmi les paramètres les plus importants. La réduction de la dimensionnalité, via l'extraction et la sélection de caractéristiques, est l'une des étapes les plus importantes notamment en apprentissage automatique.
Le but de notre travail est de, justement, pouvoir sélectionner des caractéristiques pertinentes avant de procéder à la phase d'apprentissage automatique. Ceci permet de réduire le temps et les ressources nécessaires pour le calcul. Aussi, élaguer les caractéristiques non pertinentes rend l'apprentissage plus performant et offre une meilleure compréhension des spécificités de l'opinion. Notre travail est composé en deux parties, la première porte sur une sélection de caractéristique par échantillonnage à base de mesures de pondérations . La seconde partie s'appuie sur une approche d'optimisation multicritère adaptée afin d'essayer de parer au problème de la redondance. Pour ce faire, on recherche un ensemble de solutions qui constituent un compromis entre différents critères (mesures de pondération). Nous avons opté pour l'utilisation du paradigme des requêtes skyline afin de sélectionner l'ensemble des termes qui ne sont dominés (au sens de Pareto) par aucun autre.
Ci-dessous, nous détaillerons pas à pas, les étapes suivies pour la réalisation de notre travail.

3.4.1 Acquisition et prétraitement des textes (Preprocessing)

L'acquisition des données :Tel qu'expliqué dans la section 3.3.2, nous avons fait le choix d'utiliser un corpus de critiques cinématographiques préalablement annotés (Positif et Négatif). Ainsi, aucun processus d'acquisition de données n'a été nécessaire dans le cadre de ce travail.
Cependant, un certain nombre de prétraitements de textes ont été élaborés afin d'épurer l'ensemble de données de toutes les informations inutiles et ainsi opérer une première réduction de l'espace des caractéristiques.
Dans cette première partie, nous nous contentons juste de présenter les prétraitements classiques que nous avons utilisés et qui sont suivis dans la majorité des articles scientifiques. Cependant les traitements spécifiques, propre à notre cas, seront détaillés ultérieurement. Ainsi, on retrouve :

la tokénisation :cette opération consiste à décomposer une séquence de chaînes en morceaux appelées tokens(jetons). Les tokens peuvent être sous forme des mots individuels (uni-gramme), des mots deux à deux (bi-gramme), suite de trois mots (tri-gramme), jusqu'à n-gramme, dans notre cas nous avons opté pour une tokénisation unie-gramme et bi-gramme. Pour ce faire, nous avons implémenté notre propre classe de tokénisation, indépendamment de celle proposée dans l'api Weka. Cela nous a permis d'avoir un libre accès aux délimiteurs, qui marquent les frontières de mots et de phrases. En effet, certains de ces délimiteurs sont non ambigus (comme le point d'exclamation, les doubles points), d'autres sont plutôt ambigus (comme l'apostrophe, l'espace, le tiret ou le point), ce qui a nécessité un traitement plus fin notamment en ce qui concerne l'Anglais.

Suppression des mots vides :dans ce cas, nous effectuons une suppression de mots non influents. Ils représentent 30% des mots dans un texte. Aussi, la présence de ces mots n'apporte

absolument aucune différence tant sur le plan sémantique que sur le plan lexical. Cela veut dire que leur présence dans tous les textes du corpus les rend non discriminants et du coup leur utilisation pour une tâche de classification s'avère inutile.Ajoutée à cela, leur présence augmente considérablement et inutilement la dimensionnalité de l'espace de représentation. Ces mots sont :
- Les conjonctions de coordination (for, and, nord, butor, yet , so).
-Les déterminants (à/an, the, this, that, thèse, those).
-Les prépositions (at, in, To).
Pour cibler et éliminer les mots vides, nous avons utilisé une liste des mots vides en anglais si l'un de nos tokens obtenu dans la phase de tokénisation est apparu dans cette liste, il sera supprimé.

Lemmatisation : La lemmatisation représente un procédé plus avancé que la racinisation [1], elle réduit un mot à sa forme canonique appelée lemme qui est toujours un mot correct contrairement à la racine, ce qui se formalise par une absence de pluriels, des verbes conjugués etc. La lemmatisation recherche généralement dans un dictionnaire pour trouver le lemme d'un mot, car il est difficile dans certains cas d'avoir le lemme d'un mot en se basant uniquement sur des règles morphologiques et syntaxiques, c'est pour cette raison que nous avons procédé à l'utilisation de la bibliothèque StanfordCoreNLP décrite précédemment.
La lemmatisation est étroitement liée à la racinisation, la différence est qu'un stemmer fonctionne sur un seul mot sans connaissance du contexte, et ne peut donc pas faire la distinction entre des mots qui ont des significations différentes selon la partie du discours.Cependant, les stemmers sont généralement plus faciles à implémenter et à exécuter, aussi ce manque de précision peut ne pas avoir d'importance pour certaines applications comme c'est le cas pour nous lors de cette première phase de traitement.

L'étiquetage grammatical l'étiquetage grammatical ou étiquetage morpho-syntaxique (parts-of-speech tagging ou POS tagging) est un processus dans lequel chaque entité primaire extraite du texte est caractérisée par une catégorie lexicale par exemple : le nom, verbe, adverbe, COD, informations concernant le genre, le nombre. C'est EC qui nous a permis d'obtenir les tokens du types adjectifs et les collocations bigramme (adjectif+adverbe).

Transformation de données en instance weka Après avoir prétraité nos données, nous procédons à la transformation des textes au format d'entrée par défaut au sein de Weka. Ces données seront représentées par le standard ARFF (Attribute relation file Format), qui constitue le format le plus courant pour les données utilisées dans Weka. Chaque fichier ARFF doit avoir un en-tête décrivant à quoi devrait ressembler chaque instance de données. Après l'en-tête, chaque instance doit être répertoriée avec le nombre correct d'instances.L'instance utilisée dans le cadre de notre travail est définie comme suit :
(texte, vocabulaire, label de classe) où :

— texte : représente chaque texte du corpus à l'état brute.
— vocabulaire : représente l'ensemble des tokens extrait après l'étape de prétraitement.
— label de classe : définit l'étiquette du texte brute(positif ou négatif).

1. la racinisation est le procédé de transformation d'un mot en gardant que sa racine, donc la partie du mot restante une fois que l'on a supprimé son préfixe et son suffixe

3.4.2 Phase de sélection de caractéristiques

L'apport majeur de notre travail à principalement porté sur la phase de sélection. L'approche proposée se base donc sur un processus à deux étapes.
La première étape, consiste à effectuer un filtrage à l'aide de plusieurs métriques de pondération supervisée, permettant ainsi de mesurer les capacités de discriminations des termes, de sorte que, tous les termes peu informatifs ou peu spécifiques aux deux classes de polarité (Positive et Négative) soient élagués de l'espace de recherche. Deux classifieurs (SVM et NB) ont été utilisés à ce propos. Pour chaque méthode de pondération, une phase d'apprentissage est réalisée à l'aide de la validation croisée sur la base de 25 % puis 50% et 75% du vocabulaire initial. Au final, les résultats obtenus à chaque itération seront comparés à l'aide d'un écart-type de la F-mesure.
La seconde étape, a pour but de raffiner le sous-ensemble précédemment obtenu. Ainsi, l'objectif principal est de sélectionner parmi diverses pondérations les termes discriminants pour la classification d'opinions. L'idée de cette proposition s'appuie sur le concept des préférences et établit une optimisation par domination de Pareto.

1.Première Phase de sélection

Les méthodes de sélection filtre se basent principalement sur des techniques de pondérations et de classements (Weighting and Ranking methods) comme principal critère de sélection. Ainsi, nous considérons ces dernières comme méthodes de sélection à part entière, vu qu'elles s'appliquent antérieurement à la phase de classification et permettent de filtrer les caractéristiques les plus pertinentes.

On retrouve une multitude de mesures utilisées en fouille de texte, à l'image du DeltaIDF (Martineau and Finin, 2009 ; Paltoglou and Thelwall, 2010) ou encore le Gain d'Information et l'Information Mutuelles (Deng et al., 2014). Ces dernières sont pour la plupart issue de la recherche d'informations et prennent généralement en compte l'aspect fréquentiel des termes.

Dans le cadre de ce travail, nous proposons d'utiliser un certain nombre de métriques de pondération non pas uniquement comme étape préliminaire à la phase de représentation de donnée (utilisation lors de la représentation des matrice Termes-Documents), mais plutôt comme une procédure de sélection de caractéristiques à part entière (Ranking method). Ainsi, après avoir prétraité l'ensemble des documents du corpus, on génère notre premier espace de caractéristiques (Vocabulaire généré). Puis, on opère divers pondérations sur ce dernier afin de procéder, par la suite, à la phase de classification.

Habituellement, les méthodes de ranking se basent sur un seuil de réduction préalablement choisie, afin d'opérer la sélection des termes candidats à la phase classification. Ce seuil prête souvent à confusion, car aucun standard ne permet de fixer au préalable le taux de réduction à opérer.

Ainsi, pour évaluer l'impact de ce seuil sur nos métriques de pondérations, nous proposons d'effectuer un échantillonnage pour lors de la réduction de notre espace de caractéristiques.

Notre sélection a été effectués sur trois sous ensembles définit à des pourcentages différents de notre vocabulaires puis nous avons associé un poids à chaque mot du sous ensemble en appliquant les mesures de pondération(Delta TfIdf,ChiSquare,OddsRatio) .Au cours du processus un score

correspondant à chaque entité est calculé, le sous ensemble de caractéristiques résultant servira d'ensemble d'entrée aux algorithmes de classification.

les paramètres utilisés par les métriques de pondérations sont illustrés ci-dessous :

	signification
P(f/c)	la probabilité que la caractéristique f appartient à la classe C
P(f/ $\bar{c}$**)**	La probabilité que la caractéristique f appartient à la classe complémentaire de C
N	Le nombre de documents total du corpus
$\mathbf{d}f_c$	Nombre de documents de la classe C qui contiennent la caractéristique f
$\mathbf{d}\bar{f}\bar{c}$	Nombre de documents qui n'appartiennent pas à C et qui ne contiennent pas la caractéristique f
$\mathbf{df}\bar{c}$	Nombre de documents qui n'appartiennent pas à C et qui contiennent la caractéristique f
$\mathbf{d}\bar{f}\mathbf{c}$	Nombre de documents qui ne contiennent pas la caractéristique f dans la classe C
df	Nombre de documents qui contiennent la caractéristique f dans le corpus global
$\mathbf{d}\bar{f}$	Nombre de documents qui ne contiennent pas la caractéristique f dans le corpus global
Nc	Nombre de documents qui appartiennent à la classe C
$\mathbf{N}\bar{c}$	Nombre de documents qui n'appartiennent pas à la classe C

TABLE 3.2 – Signification des paramètres des métriques de pondération utilisées

-oddsRatio : cette mesure permet de donner un score positif aux caractéristiques qui apparaissent plus souvent dans une catégorie que dans l'autre, et un score négatif si cela se produit plus dans l'autre. Un score de zéro signifie que les chances pour une caractéristique d'apparaître dans une catégorie sont exactement la même que la probabilité que cela se produise dans l'autre catégorie.

$$orr(f,c) = log(\frac{p(f|c).(1-p(f|\bar{c}))}{p(f|\bar{c}).(1-p(f|c))})$$

-Chi-Square : Cette mesure permet de calculer le manque d'indépendance entre variables nous avons deux variables cibles c'est-à-dire l'étiquette de classe positive/négative et des caractéristiques décrivant chaque échantillon de données. Nous calculons maintenant des statistiques du chi-square entre chaque variable caractéristique et la variable cible et observons l'existence d'une relation entre les variables et la cible. Si la variable cible est indépendante de la variable d'objet, nous pouvons ignorer cette variable d'objet. S'ils sont dépendants, la variable de caractéristique est très importante. Donc, une valeur élevée de chi square indique que l'hypothèse d'indépendance est incorrecte. En d'autres termes, plus la valeur du chi square est élevée, plus la caractéristique est susceptible d'être corrélée à la classe. Elle doit donc être sélectionnée pour la formation sur modèle.

$$x^2(f,c) = \frac{N.((df_c.d\bar{f}\bar{c})-(df\bar{c}.d\bar{f}c))^2}{df.d\bar{f}.N_c.N_{\bar{c}}}$$

-DeltaTfIdf : TfIdf est une mesure statistique qui permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus.
Le *Delta TFIDF* calcule la différence des TfIdf positif et TfIdf négatifs, et permet de donner un score positif aux caractéristiques qui apparaissent plus souvent dans une catégorie que dans l'autre,

$$C_{t,d} * \log_2\left(\frac{N_t}{P_t}\right)$$

Ainsi, la première étape de notre série d'expérimentations a été menées sur la base de ces trois méthodes de sélection (Chi-square, Odds Ratio et Detlta TF-IDF). Ensuite, pour chaque liste de termes obtenue à l'aide de ces trois mesures respectives, nous avons fait le choix de ne garder qu'un nombre restreint de caractéristiques réparties comme suit : 25%, 50% et 75%. L'intérêt de cette répartition est de pouvoir évaluer le comportement des algorithmes de classification en fonction de l'espace de caractéristiques utilisé.

La phase de classification a été réalisée à l'aide des algorithmes Naive Bayes et SVM. Les résultats de la classification sont présentés en termes de F-mesure. Cette mesure est calculée grâce à

la validation croisée à 10 échantillons (cross validation 10-fold). Puis, afin d'estimer la dispersion des résultats autour de la moyenne de classification, nous avons fait le choix de calculer la valeur de l'écart type de notre série de tests.

	25%	**50%**	**75%**	*Ecart-type*
Chi-Square	0.631	0.657	**0.727**	0.0496
Odds Ratio	0.651	0.670	0.686	**0.0175**
Delta TF-IDF	0.605	0.628	0.689	0.043

TABLE 3.3 – Résultat de la classification bayesienne des unigrammes de mots

	25%	**50%**	**75%**	*Ecart-type*
Chi-Square	0.518	0.662	0.524	0.0814
Odds Ratio	0.697	0.671	0.752	**0.0502**
Delta	0.660	0.753	**0.809**	0.0610

TABLE 3.4 – Résultat de la classification SVM des unigrammes de mots

Les tableaux 3.1 et 3.2 représentent les performances de classification obtenues pour chacun

des classifieurs SVM et NB, à l'aide des trois combinaisons citées ci-dessus. Les résultats montrent une divergence de la F-mesure. En effet, les meilleurs taux de classification avec NB et SVM ont été obtenus respectivement, grâce aux métriques Chi-squre et Delta TF-IDF avec un taux de réduction de 75%. On note également que les plus faibles taux de classification ont été réalisés avec Chi-squre et Delta TF-IDF en utilisant, respectivement, NB et SVM avec un taux de réduction de 25%. On note également une certaine dispersion des résultats obtenus avec Chi-squre et Delta TFIDF, tout au long de l'expérimentation et ce, quel que soit l'algorithme de classification employé. Ceci nous amène à penser que, les performances obtenu avec Chi-squre et Delta TF-IDF augmentent à mesure que le nombre de caractéristiques augmente, ce qui est contraire au but recherché.

Pour valider notre constatation, nous proposons d'utiliser une mesure largement répandu en recherche d'information et qui permet d'estimer la dispersion d'un échantillon statistique. Cette mesure qu'est l'Ecart-type (noté σ) sert principalement à mesurer la dispersion autour de la moyenne d'un ensemble de données. Une valeur de σ proche de 0 signifie que les valeurs sont très peu dispersées autour de la moyenne. En contrepartie, plus ces valeurs sont éloignées de la moyenne, plus l'écart-type est élevé.

On constate à travers les tableaux 3.1 et 3.2, que la plus faible valeur de l'écart-type a toujours été obtenue grâce à la métrique Odds ratio. On remarque d'ailleurs que les taux de classification réalisés avec cette dernière présente une certaine stabilité, quel que soit le taux de réduction appliqué. Aussi, les taux de classification obtenus respectent parfaitement la condition de l'écart-type, selon laquelle une série normalement constituée doit avoir environ 68 % de ces valeurs qui appartiennent à l'intervalle [Moyenne – σ ; Moyenne + σ]. Ainsi, de manière naïve, nous avons pris la décision de continuer la suite de nos expérimentations à l'aide de la métrique Odds ratio.

Certe, les meilleurs taux de classification ont été réalisés à l'aide de Chi-squre et Delta TF-IDF. Néanmoins, nous estimons qu'Odds ratio a été la métrique la moins impactante pour les algorithmes de classification, losque le nombre de caractéristiques est disproportionné.

Ainsi, pour la suite de notre travail, nous choissisons d'utiliser Odds ratio avec l'algorithme SVM, sur un taux de réduction correspondant à 75% du vocabulaire initial.

2.Deuxième phase de séléction :

Tel que décrit dans la section précédente, les résultats obtenus lors de la phase de validation croisée sont, malheureusement, peu attrayant. Certes que, l'utilisation des différents échantillonnages a permis de mettre en valeur une mesure de pondération. Cependant, celle-ci ne permet pas, à elle seule, d'opérer une sélection efficace des caractéristiques discriminantes. Aussi, l'approche proposée reste dépendante du type de classifieur utilisé lors de la phase d'apprentissage et les caractéristiques sélectionnées risquent, quelque peu, d'influencer les résultats de prédiction dans le cas d'un changement de classifieur.
Par ailleurs, les caractéristiques obtenues par le biais d'une unique mesure de pondération peuvent théoriquement être infructueuses, du fait que certains critères n'ont pas été pris en considération lors de la sélection. Or que, leur prise en compte aurait pu être bénéfique du point de vue sé-

mantique. Ainsi, dans cette deuxième partie nous essayons d'obtenir des termes caractérisant au mieux, les deux polarités opposées (positive et négative) et ce, en tenant compte d'un grand nombre de critères possibles.
L'idéal dans le cas d'une sélection de caractéristique pour une classification de polarité, serait de prendre en considération des critères de nature contradictoire. En effet, cela permet d'évaluer l'impact qu'a une mesure de pondération (facteur) sur la tâche de sélection, et ce, en la confrontant à une autre mesure dont l'impact est contradictoire. C'est la raison pour laquelle nous avons opté pour l'utilisation des requêtes skyline.
Ainsi, lors de cette 2ème phase de sélection, on choisit un ensemble de mesures permettant d'évaluer chaque caractéristique en fonction d'un certain nombre d'objectifs ; chaque mesure correspondra à une dimension du skyline. Pour mieux discriminer les termes qui caractérisent chaque classe de polarité, nous procédons à la séparation du corpus d'apprentissage en deux sous-ensembles (SCP et SCN). Les sous corpus contiennent des documents positifs (resp. négatif) exprimant des opinions en faveur (resp défaveur) du sujet de prédilection du corpus. Après quelques notions préliminaires concernant les requêtes skyline, nous détaillerons ensuite les mesures utilisées dans notre approche

3.4.3 Requête Skyline

Skyline au sens de Pareto Les requêtes skylines sont un exemple spécifique et bien pertinent des requêtes à préférences.Elles permettent de sélectionner l'ensemble des points considérés les plus intéressants, lorsque différents critères et souvent conflictuels sont pris en compte. Elles s'appuient sur le principe de dominance au sens de Pareto qui peut être défini comme suit :

Définition 1 (Principe de dominance au sens de Pareto) : Soit D un ensemble de points d-dimensionnels et, p_i et p_j deux points de D. On dit que p_i domine (au sens de Pareto) p_j si et seulement si p_i est meilleur ou égal à p_j sur toutes les dimensions et strictement meilleur que p_j sur au moins une dimension.

On dit alors que p_i domine (est préféré à) p_j

$$\text{on note } (p_i) \geq (p_j)$$

Définition 2 (Skyline). Le skyline S de D est l'ensemble des points, dits points skyline, qui ne sont dominés par aucun autre point de D :

$$S = \{\ p \in D \mid !\exists\ p' \in D,\ p' \geq p\ \}$$

Les requêtes skyline calculent donc l'ensemble des tuples (points, éléments, objets) optimaux au sens de Pareto dans une relation, c à d, les tuples qui ne sont dominés par aucun autre tuple de la même relation.
Nous disposons maintenant, d'une brève définition formelle du concept de requêtes skyline. On s'intéressera dans ce qui suit, à la manière de déployer ces requêtes dans un cadre applicatif. Pour illustrer le concept, considérons l'exemple ci-dessous :
Supposons qu'on dispose d'une base de données contenant des informations sur des candidats comme montrés en table 1. Cette table comporte les informations suivantes : Code, Âge, Expérience en management (man-exp en année), Expérience technique (tec-exp en année) et la distance

séparant le travail au domicile (dist-td en Km).
La procédure de recrutement mise en place par la direction des ressources humaines visent à choisir les candidats ayant plus d'expérience technique (Max man-exp) et plus d'expérience en management (Max tec-exp), tout en ignorant les autres critères. L'application du skyline traditionnel sur la liste 1 renvoie les candidats suivants : M5, M8 voir figure 3.3 de [5].

code	age	man_exp	tec_exp	dist_td
M1	32	5	10	35
M2	41	7	5	19
M3	37	5	12	45
M4	36	4	11	39
M5	40	8	10	18
M6	30	4	6	27
M7	31	3	4	56
M8	36	6	13	12
M9	33	6	6	95
M10	40	7	9	20

FIGURE 3.2 – Liste des candidat

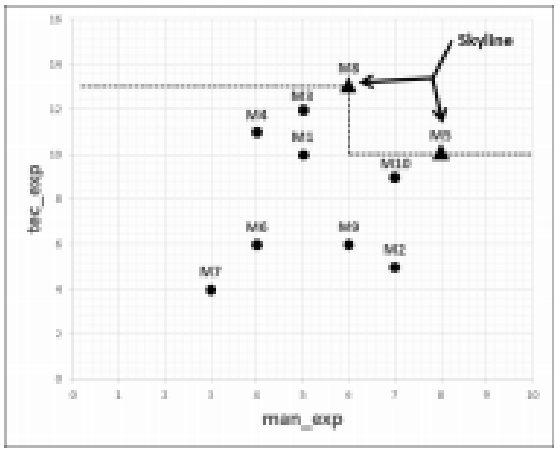


FIGURE 3.3 – Skyline des candidats

Ce problème de recherche des skylines a de nombreuses autres applications dans le monde réel. En effet, les décideurs de nombreux domaines d'activité (industrie, agriculture, finance, ...) doivent rechercher les meilleurs compromis afin d'atteindre de multiples objectifs parfois conflictuels.[2]. Cependant, très peu d'applications impliquant le Skyline, ont été recensées dans le domaine de la sélection de caractéristique, on retrouve [5],. Dans le cadre de notre travail, nous essayerons par le biais des requêtes skyline, d'obtenir un compromis entre différentes mesures de pondérations d'un même terme, afin de maximiser l'interaction entre aspect sémantique et statistique sur la base des travaux présenté dans [5].

	Rappel	**Précision**	**F-mesure**
Naîf Bayes	0.606	0.607	**0.683**
SVM	0.690	0.708	**0.768**

TABLE 3.5 – Résultats de la sélection de caractéristique en utilisant le skyline.

le tableau 3.3 représente les performances de classification obtenues pour chacun des classifieurs SVM et NB en utilisant la sélection des meilleurs points Skyline. On constate une légère amélioration des résultats de la F-mesure quelque soit l'algorithme utilisé, notamment ceux obtenus avec SVM et ayant enregistré le meilleur taux. Cependant, ces taux de classification restent en dessous des résultats escomptés notamment si l'on se base sur les récents travaux présentés en baseline dans [6]. En effet, les meilleurs taux enregistrés en utilisant les approches d'apprentissage automatique se situent aux alentours des 0.92 de la F-mesure, ce qui nous positionne à 16% en dessous de cette baseline.

Nous supposons que, la configuration des dimensions utilisées serait l'un des facteurs majeurs ayant conduit à la faiblesse de nos résultats. En effet, un choix peu adéquat des dimensions amène dans la majorités des cas soit à des réductions drastiques de l'espace d'attributs, ou au contraire, cela nous amène à des taux de réduction assignifiants. On peut facilement voir cela à travers le taux de réduction des caractéristiques initiales élevé obtenu lors de nos expérimentations et qui se situe environs à 30%. Cette problématique est très répandu au sein de la communauté scientifique étudiant l'optimisation à base de requêtes Skyline. Plusieurs solutions ont été proposées pour remédier à ce problème, on retrouve notamment les méthodes de raffinement et de relaxation de requêtes Skyline. Mais l'un des facteurs majeurs qui serai intéressant d'étudier dans un premier temps, est l'optimisation des mesures et critères servants à la représentation des dimensions de l'espace Skyline. Ce qui pourrai faire office de perspectives dans une prochaine étude.

3.4.4 Conclusion

Dans ce chapitre nous avons présenté l'approche qui nous a permis de construire notre modèle de classification en commençant par la description de la chaîne de traitement que nous avons suivi et en décrivant l'approche de sélection de caractéristique qui est répartie en deux phases : une première phase qui nous a permis de notre vocabulaire de termes et affiné la deuxième phase du processus de sélection qui est la méthode "Skyline" et qui consiste à extraire les termes les plus discriminants pour une meilleure classification.

Conclusion générale

Le nouveau web fournit une importante collection d'opinions sous plusieurs formes. Il est évident que cette nouvelle collection de données favorise de nouvelles directions dans le domaine de l'opinion mining. Au cours de ce mémoire, nous avons conclu l'étude et la présentation du domaine de l'opinion Mining et du sentiment analysis et ses problématiques, c'est un domaine émergeant et un nouvel axe de recherche permettant de faciliter et améliorer la vie quotidienne, nous avons présenté les différentes approches, méthodes et leurs résultats.

L'objectif principal de notre travail est de s'initier aux techniques de l'apprentissage automatique, à travers une approche supervisée de classification de polarité d'opinions.

D'abord, nous décrivons le problème de la fouille d'opinions dans son ensemble,nous avons introduit quelques notions appartenant au domaine de la fouille d'opinions et ses composants puis nous avons présenté les principales notions et différents concepts propres à la fouille d'opinion. Nous avons également discuté des problématiques ainsi que les domaine d'application de l'analyse d'opinion

Nous avons dressé un état de l'art de différentes approches d'apprentissage automatique(supervisée et non supervisée). Ensuite nous avons détaillé les principales étapes de catégorisation de texte ensuite nous avons détaillé les approches et les méthodes de la sélection de caractéristiques ainsi que les métriques d'évaluation commune à l'analyse d'opinions. par la suite, nous avons enchaîné avec une synthèse des quelques travaux autours de l'analyse d'opinions.

Enfin, nous avons décrit la chaîne de traitement qui nous a permis de construire notre modèle de classification en détaillant chaque phase du processus. En commençant par la phase de prétraitement puis la sélection de caractéristique en utilisant la méthode Skyline enfin nous nous avons présenté les résultats de la classification en utilisant les algorithmes d'apprentissage automatique.

En guise de perspectives, il serai intéressant d'étudier une approche de relaxation Skyline afin d'apporter plus de mieux appréhender l'espace de caractéristiques constituant l'ensemble Skyline.

Annexe

— **Corpus** : ensemble limité d'éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique, ensemble de textes réunis à des fins de comparaison, servant de base à une étude quantitative.

— **n-gramme** : Succession de N éléments du même type extraits d'un texte, d'une séquence ou d'un signal, les éléments pouvant notamment être des mots ou des lettres.Les N grammes sont beaucoup utilisés en traitement automatique du langage naturel mais aussi en traitement du signal.

— **Pondération** : relations entre des poids ou des puissances qui s'équilibrent mutuellement, Balancement des masses, équilibre des figures, Juste équilibre ; Caractère de ce qui est pondéré, bien équilibré.

— **SentiWordNet** : est une ressource lexicale pour l'extraction d'opinion basée sur WordNet. SentiWordNet attribue à chaque synset de WordNet trois scores de sentiment : la positivité, la négativité et l'objectivité.

— **Sac de mots (bag of words en Anglais)** : un document particulier est représenté par l'histogramme des occurrences des mots le composant : pour un document donné, chaque mot se voit affecté le nombre de fois qu'il apparait dans le document (voir la notion de multi-ensemble, bag en anglais). Un document est donc représenté par un vecteur de la même taille que le dictionnaire, dont la composante i indique le nombre d'occurrences du i-ème mot du dictionnaire dans le document.

— **Z-score** : est une mesure numérique de la relation entre une valeur à la moyenne dans un groupe de valeurs. Si un Z-score est de 0, le score est identique à la note moyenne. Le Z-scores peut également être positif ou négatif, avec une valeur positive indiquant le score est supérieur à la moyenne et un score négatif indiquant qu'il est en dessous de la moyenne.

— **Polarité (orientation sémantique)** : si le locuteur exprime une opinion positive ou négative, le score obtenu permet de mesurer la satisfaction globale en temps réelle, alors la classification peut être faite à peine si une phrase contient une opinion positive sur une caractéristique d'un objet ou il peut contenir un avis négatif sur elle .

— **Fouille d'opinion** : le terme fouille d'opinion (opinion mining)est utilisé pour évoquer le traitement automatique des opinions, des sentiments et de la subjectivité dans les textes. Ce domaine est connu sous le nom d'opinion mining.

Bibliographie

Chapitre 1 : *Contexte de l'analyse d'opinion.*

[1] Opinion mining et sentiment analysis méthodes et outils, Dominique Boullier et Audrey Lohard, OpenEdition Press, Sciences Po médialab, 2012.

[2] http ://www.theses.fr/2014TOU30076

[3]Rainie, H., Cornfield, M., & Horrigan, J. B. (2005). The Internet and campaign 2004. Pew Internet & American Life Project.

[4]Thomas, M., Pang, B., & Lee, L. (2006, July). Get out the vote : Determining support or opposition from Congressional floor-debate transcripts. In Proceedings of the 2006 conference on empirical methods in natural language processing (pp. 327-335). Association for Computational Linguistics.

[5] Pestian, J.P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Cohen, K.B., Hurdle, J., and Brew, C. (2012). Sentiment Analysis of Suicide Notes : A Shared Task. Biomed. Inform. Insights 5, 3–16.

Chapitre 2 : *Etat de l'art sur L'analyse d'opinion*

[1] Thèse de doctorat, Approches basées sur les modèles de langue pour la recherche d'opinions, Faiza BELBACHIR, Université de Toulouse, 2014.

[2] Mémoire de MAGISTER, Commande d'un onduleur par des approches basées sur des réseaux de neurones artificiels, Madjid BOUDJEDAIMI, Université Mouloud Mammeri Tizi Ouzou, 2011.

[3] Thèse pour l'obtention du grade de Docteur, Sélection de caractéristiques : méthodes et applications, Hassan CHOUAIB, Université Paris Descartes, 2011.

[4] Cours Data Mining 2 Mme S.Fellag

[5] Turney, P.D. (2002). 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, (Stroudsburg, PA, USA : Association for Computational Linguistics), pp. 417–424.

[6] Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., and Jurafsky, D. (2004). Automatic extraction of opinion propositions and their holders. In 2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text, p.

[7] Sutton, C., and McCallum, A. (2012). An Introduction to Conditional Random Fields. Found. Trends® Mach. Learn. 4, 267–373.

[8] Jalam, R. (2003). Apprentissage automatique et catégorisation de textes multilingues. PhD Tesis Univ. Lumiere Lyon 2.

[9] Ahkter, J.K., and Soria, S. (2010). Sentiment analysis : Facebook status messages. Unpubl. Masters Thesis Stanf. CA.

[10] Kouloumpis, E., Wilson, T., and Moore, J.D. (2011). Twitter sentiment analysis : The good the bad and the omg ! Icwsm 11, 164.

[11] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC, pp. 2200–2204.

[12] Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M.T., and Ureña-López, L.A. (2012). Random Walk Weighting over Sentiwordnet for Sentiment Polarity Detection on Twitter. In Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, (Stroudsburg, PA, USA : Association for Computational Linguistics), pp. 3–10.

[13] Salton, G., Wong, A., and Yang, C.S. (1975). A Vector Space Model for Automatic Indexing. Commun ACM 18, 613–620.

[14] Hu, M., and Liu, B. (2005). Mining and Summarizing Customer Reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (New York, NY, USA : ACM), pp. 168–177.

[15] Fouille d'opinions Méthodes et outils.Etude des méthodes existante de classification et teste d'opinions.Université Tébessa.

[16] Kohavi, R., and John, G.H. (1997). Wrappers for feature subset selection. Artif. Intell. 97, 273–324.

[17] Yang, Y., and Pedersen, J.O. (1997). A comparative study on feature selection in text categorization. In Icml, pp. 412–420.

[18]Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs Up ? : Sentiment Classification Using Machine Learning Techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, (Stroudsburg, PA, USA : Association for Computational Linguistics), pp. 79–86.

[19] Dang, Y., Zhang, Y., and Chen, H. (2010). A Lexicon-Enhanced Method for Sentiment Classification : An Experiment on Online Product Reviews. IEEE Intell. Syst. 25, 46–53.

[20] Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders : Domain adaptation for sentiment classification. In ACL, pp. 440–447.

[21]Tang, H., Tan, S., and Cheng, X. (2009). A survey on sentiment detection of reviews. Ex-

pert Syst. Appl. 36, 10760–10773.

[22] Wang, G., Sun, J., Ma, J., Xu, K., and Gu, J. (2014). Sentiment classification : The contribution of ensemble learning. Decis. Support Syst. 57, 77–93.

[23] Gamon, M. (2004). Sentiment Classification on Customer Feedback Data : Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis. In Proceedings of the 20th International Conference on Computational Linguistics, (Stroudsburg, PA, USA : Association for Computational Linguistics), p.

[24] Grouin, C., Berthelin, J.-B., El Ayari, S., Heitz, T., Hurault-Plantet, M., Jardino, M., Khalis, Z., and Lastes, M. (2007). Présentation de deft'07 (défi fouille de textes). Actes Trois. DÉfi Fouille Textes 3.

[25] Khan, F.H., Bashir, S., and Qamar, U. (2014). TOM : Twitter opinion mining framework using hybrid classification scheme. Decis. Support Syst. 57, 245–257.

[26] Barbosa, L., Kumar, R., Pang, B., and Tomkins, A. (2009). For a Few Dollars Less : Identifying Review Pages Sans Human Labels. In Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, (Stroudsburg, PA, USA : Association for Computational Linguistics), pp. 494–502.

[27] Vandersmissen, B. (2012). Automated detection of offensive language behavior on social networking sites. Master's thesis, Universiteit Gent, 2012. URL http ://lib. ugent. be/fulltxt/RUG01/001/8
001887239_2012_0001_AC. pdf.(Cited on page 13.).

[28] Koppel, M., and Schler, J. (2006). The Importance of Neutral Examples for Learning Sentiment. Comput. Intell. 22, 100–109.

[29] Wiegand, M., and Klakow, D. (2009a). The Role of Knowledge-based Features in Polarity Classification at Sentence Level. In FLAIRS Conference, p.

[30] (Blair-Goldensohn et al., 2008 ; Eguchi and Lavrenko, 2006 ; Ikeda et al., 2008 ; Meena and Prabhakar, 2007 ; Nakagawa et al., 2010 ; Wiegand and Klakow, 2009a).

[31] (Fink et al., 2011 ; Gamon et al., 2005 , and McDonald, 2011 .,Wiegand and Klakow, 2009b.

[32] Yu, H., and Hatzivassiloglou, V. (2003a). Towards Answering Opinion Questions : Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, (Stroudsburg, PA, USA : Association for Computational Linguistics), pp. 129–136.

[33] Agarwal and Mittal, 2013, 2014 ; Pang and Lee, 2008b ; Tan and Zhang, 2008.

[34] Tan, S., and Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. Expert Syst. Appl. 34, 2622–2629.

[35] Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment Analysis in Multiple Languages : Feature Selection for Opinion Classification in Web Forums. ACM Trans Inf Syst 26, 12 :1–12 :34., B., and Mittal, N. (2012). Categorical probability proportion difference (CPPD) : a feature selection method for sentiment classification. In Proceedings of the 2nd Workshop on Sentiment Analysis Where AI Meets Psychology, COLING, pp. 17–26.

[36] Nicholls, C., and Song, F. (2010). Comparison of feature selection methods for sentiment analysis. Adv. Artif. Intell. 286–289.

[37] Gamon, M., Aue, A., Corston-Oliver, S., and Ringger, E. (2005). Pulse : Mining Customer Opinions from Free Text. In Advances in Intelligent Data Analysis VI, (Springer, Berlin, Heidelberg), pp. 121–132.

[38] Simeon, M., and Hilderman, R. (2008). Categorical Proportional Difference : A Feature Selection Method for Text Categorization. In Proceedings of the 7th Australasian Data Mining Conference - Volume 87, (Darlinghurst, Australia, Australia : Australian Computer Society, Inc.), pp. 201–208.

[39] Duric, A., and Song, F. (2012). Feature selection for sentiment analysis based on content and syntax models. Decis. Support Syst. 53, 704–711.

[40] O'Keefe, T., and Koprinska, I. (2009). Feature selection and weighting methods in sentiment analysis. In Proceedings of the 14th Australasian Document Computing Symposium, Sydney, pp. 67–74.

[41] Verma, S., and Bhattacharyya, P. (2009). Incorporating semantic knowledge for sentiment analysis. Proc. ICON.

[42](Saidani, 2017), ''Contribution à la Sélection de Caractéristiques Discriminantes pour l'Analyse d'Opinion", Thèse de Doctorat 3ème cycle LMD. 2017

Chapitre 3 : *Description et présentation du modèle proposé*

[1] https ://projet.liris.cnrs.fr/inforsid/sites/default/files/a447c1bguEFeLWdKY.pdf

[2] Thèse de doctorat, Analyse multidimensionnelle interactive de résultats de simulation. Aide à la décision dans le domaine de l'agroécologie Tassadit BOUADI, Institut de recherche en informatique et systèmes aléatoires, 2013.

[3] Martineau, J., and Finin, T. (2009). Delta TFIDF : An Improved Feature Space for Sentiment Analysis. Icwsm 9, 106.]

[4] Paltoglou, G., and Thelwall, M. (2010). A Study of Information Retrieval Weighting Schemes for Sentiment Analysis. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, (Stroudsburg, PA, USA : Association for Computational Linguistics), pp. 1386–1395.

[5] (Saidani, 2017), ''Contribution à la Sélection de Caractéristiques Discriminantes pour l'Analyse d'Opinion", Thèse de Doctorat 3ème cycle LMD. 2017.

[6] Liu,C,Cheng,G, Chen,X,& Pang,Y(2018) ,Plantetary gears feature extraction and fault dignosis method based on VMD and CNNSensors, 18(5),1523.

Webographie

Chapitre 1 : *Contexte de l'analyse d'opinion.*

[1] https ://fr.wikipedia.org/wiki/Opinion_mining

[2] L. Bing, op. cit.

[3] https ://tel.archives-ouvertes.fr/tel-00777603/

[4] http ://www.thebeaconservices.com/sentiment_analysis.php ?fbclid=IwAR2qhGnvYKIQl24bHzJ6
egby806MSkWtvw7w

[5] https ://www.flipkart.com/

[6] https ://play.google.com/store/apps

[7] https ://quot.tns-sofres.com/Default.aspx ?Source=Toluna&AspxAutoDetectCookieSupport=1

[8] https ://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

[9] https ://www.prestashop.com/

Chapitre 2 : *Etat de l'art sur L'analyse d'opinion*

[1] https ://tadg5.files.wordpress.com/2015/03/classificateur-du-reseau-bayesien-naif.pdf

[3] https ://fr.wikipedia.org/wiki/Arbre_de_d%C3%A9cision

[4] https ://fr.wikipedia.org/wiki/R%C3%A9seau_bay%C3%A9sien

[5] https ://eric.univ-lyon2.fr/ ricco/cours/slides/svm.pdf

[6] http ://perso.ensta-paristech.fr

Chapitre 3 : *Description et présentation du modèle proposé*

[1] https ://waytolearnx.com/2018/11/difference-entre-jdk-jre-jvm.html

[2] https ://www.jmdoudoux.fr/java/dej/chap-maven.htm

[3] https ://stanfordnlp.github.io/CoreNLP/

[4] https ://blogs.msdn.microsoft.com/mlfrance/2014/08/05/evaluer-un-modle-en-apprentissage-automatique/