REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE MATHEMATIQUES ET DE L'INFORMATIQUE

DEPARTEMENT D'INFORMATIQUE

N° :……..

DOMAINE : MATHEMATIQUES ET DE L'INFORMATIQUE

FILIERE : INFORMATIQUE

OPTION : SIGL

Mémoire présenté pour l'obtention
Du diplôme de Master Académique

Par : SERAI Abderrahmane

Intitulé

Classification d'opinion en langue arabe

Soutenu devant le jury composé de :

……	Université de M'sila	Président
B. BELKACEM	Université de M'sila	Rapporteur
……	Université de M'sila	Examinateur

Année universitaire : 2018 /2019

Remerciements

Avant tout,

Je remercie ALLAH qui nous a donné la force,

le courage et l'espoir nécessaire pour accomplir ce travail.

Ce travail aussi modeste.

n'a été rendu possible que grâce aux orientations éclairées de mon encadreur : Mr BRAHIMI Belkacem que nous tenons à lui exprimer ma parfaite gratitude et mes sincères remercîments pour la qualité de son encadrement et pour ses conseils judicieux et avisés.

Je tenons à remercier également les membres de jury pour

avoir fait le plaisir d'accepter d'examiner ce travail. Je dédie ce mémoire à tous ceux qui ont contribué à ma formation et qui m' ont soutenus dans nos études

Dédicaces

A mes très chers parents qui ont toujours répondu présents dans les moments les plus difficiles et m'ont soutenu et encouragé tout le long de mes études, leurs confiances et leurs Sacrifices qui ont contribués à ma réussite.

A mes chers frères A ma famille

A tous mes amies

Et tous ceux qui par leurs conseils,

leur attention,

leurs encouragements et leur soutien m'ont aidé à réaliser cette œuvre.

Je ne manquerais pas de remercier tous les professeurs qui m'ont suivi pendant mon cursus universitaire

Table des matières :

CHAPITRE 1 : TEXT MINING

.

CHAPITRE 2 : LA Fouille D'opinion

Chapitre 03 : la méthodologie

CHPITRE 04 IMPLIMENTATION & RESULTAT

TABLE DES MATIERES

LISTE DES TABLEAUX ET FIGURES

Introduction générale :

De nos jours, les besoins de catégorisation automatique de documents en raison de l'augmentation constante du volume d'informations accessibles électroniquement, la conception et la mise en œuvre d'outils efficaces, permettant notamment à l'utilisateur de n'avoir accès qu'à l'information qu'il juge pertinente, devient une nécessité absolue. Comme la plupart de ces outils sont destinés à être utilisés dans un cadre professionnel, les exigences de fiabilité et de convivialité sont très importantes ; les problèmes à résoudre pour satisfaire ces exigences sont nombreux et difficiles.

La Fouille de données d'Opinion (Opinion Mining) vise l'étude de la subjectivité (subjective-objective) dans un texte : opinions, avis, sentiments, évaluations, croyances ou jugements personnel. Ensuite, il attribue une classe sémantique (positive, négative ou neutre) à ce texte.

Les documents porteurs d'opinion ont une importance stratégique et économique évidente pour les entreprises et les clients. En effet, l'étude de ces textes permet de connaître les tendencies des consommateurs, les avantages et les inconvénients des produits et des services. Pour les entreprises, cela permet d'améliorer la qualité de ses produits, alors que sur la base des critiques des consommateurs, les peuvent prendre la bonne décision concernant l'achat d'un produit.

Le but de notre travail de Master est d'étudier la faisabilité d'utiliser les algorithmes basés sur l'apprentissage automatique pour la catégorisation de textes d'opinions.

La structure proposée du mémoire peut être présentée comme suit :

Dans le premier chapitre nous introduisons des notions générales sur les domaines de :

Text Mining en donnant quelques définitions, les taches principales, les applications de chacun et surtoutla relation entre l'apprentissage automatique et le Text Mining. À présenter le processus de la catégorisation des textes et le prétraitement des textes, ainsi que les difficultés liées à cette catégorisation.

Le deuxième chapitre vise est dédié à la présentation des différents algorithmes d'apprentissage automatique supervisée ainsi que leurs avantages et leurs inconvénients. Nous avons également introduit les différents moyens d'évaluation d'un classificateur.

Le troisième chapitreen mettant l'accent sur l'algorithme utilisé dans notre travail : la naïve bayésienne.

Et le dernier chapitre permettra d'évaluer les performances des différentes approches implémentées en présentant les résultats obtenus avec interprétation. Et nous sommes terminé par une conclusion.

CHAPITRE 1 TEXT MINING

1.1 Introduction

Text Mining c'est une combinaison de recherche d'informations et de technique linguistique informatique traitant des opinions dans un document. Il sert à résoudre les problèmes liés aux text d'opinions consternants des produits. Avant le World Wide Web, les utilisateurs demandaient l'avis de famille et des amis avant acheter un produit, de même, quand une organisation devait prendre des décisions sur un produit, ils ont mené des enquêtes sur des groupes cibles ou des consultants externes embauchés.

L'extraction d'opinions facilite pour les clients a prennent des décisions en examinant les commentaires des utilisateurs publiés sur le Web, communautés, blogs, Social Media et les sites Web des produits.

L'analyse d'opinion a pour but d'extrait des textes d'opinion et les rassemblés dans des feuilles des données et les résumant sous une forme compréhensible pour les utilisateurs finaux. Il extrait des « opinions positives », « négatives » ou « neutres » à partir de données non structurées. Il implique la gestion informatique de la subjectivité des opinions et des textes.
Dans ce chapitre on va résumer en brève la définition de la fouille de données, leur processus en expliquant aussi les différentes taches de la fouille de données.

1.2 Définition de text mining

Le Text Mining, également appelé fouille de textes ou extraction de à partir de textes, est un ensemble de méthodes, de techniques et d'outils pour exploiter les documents non structurés que sont les textes écrits, commentaire sur le web de type txt, les commentairedans page web, les documents ...etc. Pour extraire du sens de documents non structurés, le Text Mining s'appuie sur des techniques d'analyse linguistique. Le Text Mining est utilisé pour classer des documents, réaliser des résumés de synthèse automatique ou encore pour assister la veille stratégique ou technologique selon des pistes de recherches prédéfinies [1].

1.3 Approches du Text Mining

Deux approches, peuvent être envisagées pour faire du Text Maning **[2]:**

1.3.1 Approche statistique

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

1.3.2 Approche sémantique

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

1.4 Les tâches de Text Mining :

Le Text Mining n'est pas un remplacement pour la recherche d'information ou le traitement du langage naturel. Les techniques qui permettent d'organiser un corpus de documents textuels selon leur contenu ont un spectre d'utilisation très large. Le Text Mining cherche des réponses aux questions difficiles ou impossibles à résoudre avec les seuls moteurs de recherche. Des exemples de tels services incluent : [2]

- Résumer des documents qui décrivent une consommation du produit dans certaines régions ;
- Etudier des réclamations des clients, raisons des changements de comportements de consommation, analyse de l'image de l'entreprise, …
- Faire la gestion de la relation client : orienter mes mails clients reçus sur le site vers les services adéquats et les aider à répondre le plus rapidement et correctement possible ;
- Connaître les réseaux relationnels des personnes ou entreprises. Chacune de ces tâches sera un cas particulier du schéma général de la figure ci-dessous [2]:

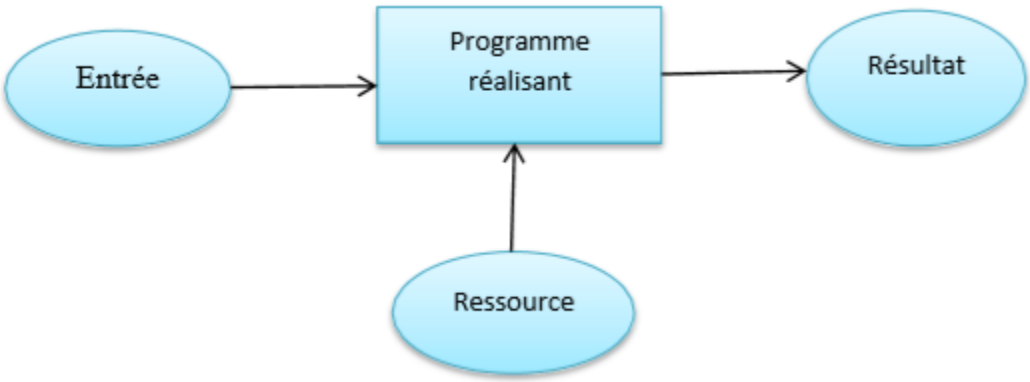


Figure 1.1 : Schéma général de la tache de fouille de textes.[2]

1.5 Processus de Text Mining :

Le Text Mining débute par la modélisation des textes en vue de leur préparation pour l'étape de Data mining et s'achève par l'interprétation de la fouille pour l'enrichissement des connaissances d'un domaine[4]. Son déroulement est tout à fait conforme à celui d'un processus. Le processus du Text Mining comprend la succession d'étapes suivantes

1.5.1 La définition du problème et identification des buts

Définition des buts attendus et des résultats souhaités [5]

1.5.2 La préparation des données

Les textes doivent être recueillis en utilisant, par exemple, des outils automatiques de récupération de l'information, ou de façon manuelle à partir de différentes sources**.** [5]

1.5.3 Nettoyage des données

Habituellement, le nettoyage consiste à éliminer les mots vides (stop-Word). Ces mots vides sont des mots ne jouant qu'un rôle syntaxique, contribuant peu au sens des documents. On les élimine pour deux raisons : (a) Minimiser la taille du fichier traité (contrainte d'espace). (b) Rendre le traitement plus rapide (contrainte de temps)**.** [5]

1.5.4 Lemmatisation :

La lemmatisation est l'opération qui consiste à ramener les variantes (flexionnelles) d'un même mot à une forme canonique, le lemme. Elle s'appuie sur une analyse grammaticale des textes afin de remplacer les verbes par leur forme infinitive et les noms par leur forme au singulier. Cette

opération permet de réduire le nombre de termes dans un index, ce qui est intéressant du point de vue du stockage des données. [5]

1.5.5 L'étude lexicométrique : La lexicométrie est l'étude quantitative du vocabulaire ; elle consiste à mesurer la fréquence d'apparition des mots dans un même texte, et il en résulte une représentation mathématique du texte**.** [5]

1.5.6 Le traitement des données (techniques de Data Mining)

On choisit l'une des techniques du Data Mining telles que les arbres de décisions, les algorithmes génétiques ou les réseaux de neurones, pour l'appliquer aux textes transformés (représentation mathématique), ce qui permettra de réaliser plusieurs tâches telles que: la classification, la traduction automatique, l'identification de la langue, etc. [5]

1.6 Applications du Text Mining

L'importance de la fouille de textes (Text Mining) ne cesse d'évoluer d'un jour à l'autre. Plusieurs domaines vitaux exploitent les techniques et les outils du Text Mining pour trouver l'information pertinente fouillée dans des quantités énormes de textes de différentes formes, parmi ces domaines on peut citer [6]:

- La recherche d'information.
- Les applications biomédicales.
- Le filtrage des communications.
- Les applications de sécurité.
- La gestion des connaissances.
- L'Analyse du sentiment**.**

1.7 TECHNIQUES LIEES A LA FOUILLE DE TEXTES

La fouille de textes s'apparente à d'autres domaines avec qui elle est très complémentaire traitement automatique des langues (TAL) et la recherche documentaire (RI) et l'extraction de l'information (EI) [7].

1.7.1 Le traitement automatique des langues « TAL»

Depuis une quinzaine d'années, avec la généralisation de l'outil informatique et d'Internet, les applications du TAL au sens large du terme se multiplient dans les disciplines

philologiques. Le TAL est une discipline à la frontière de la linguistique et de l'informatique, qui concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain [7].

1.7.2. La recherche d'information « RI»

La recherche d'information « RI » s'intéresse aux documents dans leur globalité et aux thèmes qu'ils abordent, pour comparer les documents et détecter des typologies. Elle cherche à détecter tous les thèmes présents [7].

1.7.3. L'extraction d'information « EI »

L'extraction d'information « EI » recherche des informations précises dans les documents, sans les comparer, en tenant compte de l'ordre et de la proximité des mots pour discriminer des énoncés différents ayant des mots clés identiques. L'extraction d'information consiste en l'alimentation d'une base de données structurée à partir de données exprimées en langage naturel. Il s'agit de détecter dans le texte en langage naturel les mots correspondant à chaque champ de la base de données. L'analyse est locale. L'extraction d'information est plus
complexe, car elle nécessite d'effectuer une analyse lexicale et morphosyntaxique pour reconnaître les constituants du texte (phrases, mots, verbes, adjectifs), leur nature pour détecter les phrases pertinentes et en extraire les informations voulues [7].

1.8 Langue arabe :

La langue arabe est la langue des populations arabes qui firent leur entrée dans l'histoire depuis 3 millénaires environ et qui occupaient les zones septentrionales de l'Arabie.

La langue arabe considéré la 5éme langue courantes utilisées dans le monde. Elle est parlée par plus de 422 millions de personnes en tant que première langue et de 250 millions en tant que langue secondaire, La langue arabe fait partie de la grande famille des langues sémitique.[8]

Le système archaïque d'écriture arabe était consonantique. Chaque lettre de l'alphabet arabe représente une consonne unique depuis les temps anciens. Cependant, la fin du VIIe siècle,

En arabe peut spécifier toute une phrase grammaticale par exemple «شكرا موبليس» (ce qui signifie : "et nous allons les aider»).

L'un des éléments les plus efficaces dans les phrases distinctives ou limites symboliques est des signes de ponctuation.

Ils ont émergé dans le système d'écriture arabe en 1912. En fait, l'utilisation de la ponctuation ne persiste pas dans la langue arabe.

L'alphabet arabe se compose de vingt-huit (28) lettres fondamentales

(ب أ ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه ي و), (vingt-neuf (29) lettres)

1.8.1Quelques problèmes rencontrés dans la catégorisation d'opinion arabes

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

1.8.1.1.Sur-apprentissage

Le sur-apprentissage s'explique par le fait que le modèle de prédiction n'arrive pas à bien classer les nouveaux opinion, pourtant il l'a bien fait dans la phase d'apprentissage en classant correctement les opinion de la base d'apprentissage. Pour limiter le sur-apprentissage, on doit sélectionner des termes pour réduire la dimensionnalité. D'après les expériences antérieures, le nombre de termes doit être limité par rapport au nombre d'opinion de la base d'apprentissage. Quelques auteurs recommandent d'utiliser au moins 50 à100 fois plus d'opinion que de termes. En général le nombre d'opinion d'apprentissage est limité, c'est pour cela on cherche au but de signes diacritiques pour apprendre à les reconnaitre et à les prononcer correctement en contexte, pour distinguer des lettres ambigües et pour faciliter la lecture. [43]

Agir sur le nombre des termes utilisés en les diminuant, pour éviter ce sur-apprentissage. Sans bien sûr pénaliser le système en supprimant des termes pertinents. [9]

1.8.1.2. L'homographie (signification de mot)

Deux mots sont dits homographes si 'ils s'écrivent de la même façon sans forcément avoir la même prononciation. L'homographie est une sorte d'ambigüité supplémentaire. (Ex : avocat en tant que fruit et avocat en tant que juriste c'est en français), et en arabe le mot () قلب à trois significations comme un nom [10]

Signification de mot	Phrase
Noyau	في قلب الحدث
Cœur	أجرى عملية في قلب مفتوح
Centre, moyen	الكرة في قلب الملعب

Table 1.1 La signification du mot (قلب)comme nom

1.8.1.3. Polysémie (Ambiguïté) :

Un mot possède, dans différents cas, plus d'un sens et plusieurs définitions lui sont associées. Par conséquent, à cause de la polysémie, les mots seuls sont parfois de mauvais descripteurs ; un mot arabe peut avoir plusieurs significations ; Prenons à titre d'exemple le mot arabe non voyelles ذهب qui a au moins deux significations : بَهَبْ« aller », بَهَبَ« or », l'absence des voyelles dans l'arabe couramment écrite génère une ambiguïté sur certains mots qui pénalise la performance des systèmes de traitement de opinion en langue arabe[43]

1.8.1.4Les mots composés :

La non prise en charge des mots composés comme : Arc-en-ciel, peut-être, sauve-qui-peut en langue française et إحدى عشر, واحد و عشرون en arabe etc.., Dont le nombre est très important dans toutes les langues, et traiter le mot Arc-en-ciel et le mot واحد و عشرون par exemple en étant 3 termes séparés réduit considérablement la performance d'un système de classification néanmoins l'utilisation de la technique des n-grammes pour le codage des textes atténue considérablement ce problème des mots composés. [43]

1.8.1.5. Redondance (Synonymie)

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

Cette difficulté est liée à la nature des documents traités exprimés en langage naturel contrairement aux données numériques. Exemples des synonymes en arabe sont (اعطى منح بدل وهب) qui signifie (donner), (عائلة اسرة) qui signifie (famille), et (صف فصل) qui signifie (classe). Lors d'une représentation vectorielle d'un document, ces termes sont représentés

séparément, et les occurrences du concept sont dispersées. Il est alors important de rassembler ces termes en un groupe sémantique commun. [43]

1.8.1.6. La forme de mot selon son cas

La forme de quelques mots arabes peut changer selon leurs modes de cas (nominatif, accusatif ou génitif). Par exemple le pluriel du mot (مسافر) qui signifie que (voyageur) peut être la forme

(مسافرون) dans le cas du nominatif (مرفوعة) et la forme (مسافرين)dans le cas d'accusatif/génitif (منصوبة/مجرورة) Refouler arabe de lu La présence d'un mot dans le texte indique un propos que l'auteur a voulu exprimer, on a donc une relation d'implication entre le mot et le concept associé, quoique on sait très bien qu'il y a plusieurs façons d'exprimer les mêmes choses, de lors l'absence d'un mot n'implique pas obligatoirement que le concept qui lui est associé est absent du document. Cette réflexion pointue nous amène à être attentifs quant à l'utilisation des techniques d'apprentissage se basant sur l'exclusion d'un mot particulier. Mière peut manipuler ces caisses. [43]

1.8.1.7. Présence-Absence de termes

La présence d'un mot dans le texte indique un propos que l'auteur a voulu exprimer, on a donc une relation d'implication entre le mot et le concept associé, quoique on sait très bien qu'il y a plusieurs façons d'exprimer les mêmes choses, de lors l'absence d'un mot n'implique pas obligatoirement que le concept qui lui est associé est absent du document. Cette réflexion pointue nous amène à être attentifs quant à l'utilisation des techniques d'apprentissage se basant sur l'exclusion d'un mot particulier. [43]

1.8.1.8. Subjectivité de la décision

Parmi les problèmes classiques usuels dans le domaine de l'apprentissage supervisé c'est la subjectivité de la décision prise par les experts qui décident de la classe à laquelle le texte va être attribué. Certainement après la lecture du texte à classer, l'expert va trancher à quelle(s) catégorie(s) ce texte appartient en se basant sur le contenu sémantique et le contexte du texte et même en consultant d'autres textes préalablement associés à certaines classes, pour valider la décision prise qui ne peut être que subjective. Les experts humains ne lisent pas de la même manière ! Ne réfléchissent pas de la même manière ! Donc ne classent pas de la même manière ! Ainsi un même document peut être classé différemment par deux

experts, ou encore un même document peut être classé différemment par le même expert, soumis à deux instants différents. [11]

1.8.2. Complexité de la langue arabe :

L'arabe est une langue difficile pour un certain nombre de raisons :

- Orthographe avec diacritiques est moins ambiguë et plus phonétique en arabe, certaines combinaisons de caractères peuvent être écrites de différentes manières. [12]
- Langue arabe a voyelles courtes qui donnent la prononciation différente. Grammaticalement ils sont nécessaires, mais omis dans les textes arabes écrits. [13]
- La langue arabe a une morphologie très complexe que comparer à la langue anglaise.
- Les synonymes sont très répandus. [14]
- La Classification automatique d'opinion dépend du contenu des documents, un grand nombre de fonctionnalités ou des mots-clés peut être trouvée dans le texte arabe tel que les morphèmes qui peuvent générées à partir d'une racine qui peut conduire à une mauvaise performance en termes de précision et de temps. [14]

1.9. Conclusion :

Ce chapitre présente la Text Mining (TM), les différents processus du TM, et leur tâche, aussi on a focalisé sur la langue arabe et Complexité de la langue arabe

Dans le chapitre suivant, nous exposons les déférentes méthodes de classifications d'opinions.

2.1 Introduction

La classification est une tâche très importante dans la fouille de donnée, elle consiste à créer un modèle qui peut être appliqué aux données, quand en a préparer notre base de données (nettoyage, remplissage, …), en peut appliquer une classification soit supervisé avec différentes méthodes ou non supervisée avec d'autres méthodes.

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

2.2 Définitions

2.2.1 Opinion : [3]

L'opinion est un jugement que l'on porte sur un individu, un être vivant, un phénomène, un fait, un objet ou une chose. Elle peut être considérée comme bonne ou mauvaise, tout dépend de la nature de l'individu en fonction de son caractère, ses émotions, son comportement. L'opinion peut influencer et peut donner de mauvaises informations sur un sujet étudié au sein d'un groupe, d'une personne, d'un objet. Une opinion (terme issu du verbe latin opinari) est un ensemble de jugements que l'on se fait à propos d'un objet. Selon les Définitions du pseudo-Platon, l'opinion est la « conception que la persuasion peut ébranler ; fluctuation de la pensée par le discours ; pensée que le discours peut mener aussi bien au faux qu'au vrai ». D'après Priscien de Lydie, Théophraste et Aristote définissent l'intelligence comme une faculté

Différente de la sensibilité, aussi bien que de l'opinion et de la raison. Selon Théophraste, une opinion est une déclaration concernant ce qu'il faut faire. Les opinions peuvent être paradoxales, consensuelles ou douteuses.

2.2.2 Fouille de données opinions :

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

- Quelle partie est exprimer d'opinion
- Qui a écrit l'opinion
- Ce qui est commenté

Le Fouille de donnée opinions, (en anglais appelé Fouille opinions) est une technique permettant d'automatiser le traitement de gros volumes de contenus texte pour en extraire les principales tendances et répertorier de manière statistique les différents sujets évoqués ainsi découvrir des connaissances et des relations à partir des documents disponibles.

L'outil de Text Mining va générer de l'information sur le contenu du document. Cette information n'était pas présente, ou implicite, dans le document sous sa forme initiale, elle va être rajoutée, et donc enrichir le document. [15]

2.2.3. Objectifs de la fouille de données opinions : [13]

La fouille de données textuelle peut être utilisée en particulier dans les cas suivants :

- Pour mieux comprendre le positionnement d'un discours, d'une thèse, d'un communiqué.
- Pour appréhender les thèmes récurrents qui sont associés à une activité, une entreprise ou des concurrents.
- Pour mesurer les points faibles et les points forts dans une revue de presse.
- Pour comparer des textes sur un même thème afin d'en déterminer les points communs ou au contraire de distinguer les différences stylistiques.
- Pour créer automatiquement des répertoires de sites Web ou emails associés à des thématiques. Pour quantifier un texte ou les parties d'un texte pour en extraire les structures signifiantes les plus fortes telles que le résumé automatique et la segmentation thématique.
- Pour établir des liens entre les termes et les documents utilisés dans l'indexation.
- Pour établir des règles de classification de documents (classification supervisée ou non supervisée).

2.3 La Classification : [3]

La classification est la tâche la plus commune du Data Mining ou fouille de donnée et qui semble être une obligation humaine. Afin de comprendre notre vie quotidienne, nous sommes constamment classifiés, catégorisés et évalués.

La classification est un outil puissant d'exploration des données. Elle est parmi les taches les plus important du data mining, elle consiste à étudier les caractéristiques d'un nouvel objet pour lui attribuer une classe prédéfinie. Les objets à classifiés sont généralement des enregistrements d'une base de données, la classification consiste à mettre à jour chaque enregistrement en déterminant un champ de classe. La tâche de classification est caractérisée par une définition de classes bien précise et un ensemble d'exemples classés auparavant.

Leur objectif est de créer un modèle qui peut être appliqué aux données non classifiées dans le but de les classifiées.

Voici quelques exemples de l'utilisation des tâches de classification dans les domaines de recherche et commerce sont les suivants :

- Déterminer si l'utilisation d'une carte de crédit est frauduleuse
- Diagnostiquant si une certaine maladie est présente.
- Déterminer quels numéros de téléphone correspondent aux fax.
- Déterminer quelles lignes téléphoniques sont utilisées pour l'accès à Internet.

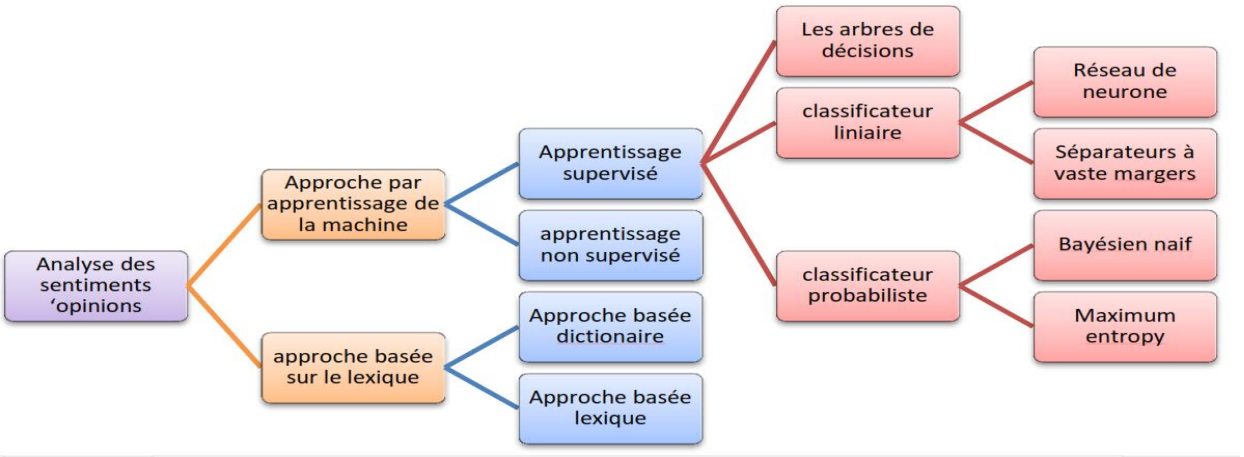


Figure 2.1 Les approche de classification[3]

2.3.1. Classification d'opinion (OC) :

Classification d'opinion (OC) également connu sous le nom de catégorisation de texte, est la tâche d'attribuer automatiquement un ensemble de documents en catégories ou classes ou sujets à partir d'un ensemble prédéfini. Cette tâche, qui tombe au carrefour de la recherche d'informations (IR) et l'apprentissage machine (ML), a été témoin d'un intérêt en plein essor au cours des dix dernières années, des chercheurs et les développeurs. [16]

2.3.2 Application de Classification d'opnion :

Peut fournir des vues conceptuelles de collections de documents et des applications importantes dans le monde réel par exemple [17] :

- Les reportages sont généralement organisés par catégories de sujets ou par des codes géographiques.
- Les Documents universitaires sont souvent classés par domaines techniques et sous- domaines.
- Les rapports des patients dans les organismes de santé sont souvent indexés à partir de plusieurs aspects : Le tri des fichiers dans les hiérarchies de dossiers, sujet des identifications, des intérêts dynamiques basés sur les tâches, organisation automatique des métadonnées, filtrage du texte organisation du document des bases de données et les pages Web.

- Analyse d'opinion (opinion mining) c'est classer automatiquement les revues (classification d'opinions) en positive, négative, neutre.
- Une autre application généralisée de la catégorisation de textes est le filtrage de spam, où les messages électroniques sont classés en deux catégories spam et non- spam
- Identification du l'auteur : dans ce cas, le système de classification doit identifier l'auteur d'opinion

2.3.3 Problème de Classification d'opinion :

Le problème de la classification d'opinion se compose de plusieurs sous-problèmes qui ont été étudiés de manière intensive dans la littérature tels que l'indexation de documents, l'attribution de la pondération, regroupement de documents, la réduction de dimensionnalité, de détermination de seuil et le type de classificateurs... [18]

Plusieurs méthodes ont été utilisées pour la classification d'opinion tel que :

- Support Vector Machines (SVM).
- K voisin le plus proche (KNN).
- Les réseaux de neurones (NN).
- Naïf Bayes (NB).
- Les arbres de décision (DT).
- Entropie maximum (ME).

2.4 Présentation de l'approche de classification :

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

2.4.1. Dictionnaire

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

contiennent que des mots vraiment spécifiques au corpus étudié. Nous pensons en effet que les lexiques d'opinion construits à l'aide des méthodes basées sur les dictionnaires (tels que WordNet), où l'on détermine la polarité des mots en fonction de leur synonymie, sont un peu trop aléatoires car beaucoup de mots peuvent avoir plusieurs sens selon le contexte. Nous avons aussi jugé que le corpus étudié n'est pas adapté aux constructions de lexiques d'opinion basées sur les corpus, les commentaires étant en règle générale très courts. [44] [45] Au final, 350 mots a priori porteurs d'opinion ont été classés dans deux catégories. Le lexique de mots positifs contient 150 éléments et le lexique de mots négatifs en contient 200. La construction de notre Dictionnaire (lexiques) a été réalisée à l'aide de [46].

2.4.2. Présentation de notre corpus de textes d'opinion

la classification supervisée nécessite des exemples (données étiquetées) afin de construire le « corpus d'apprentissage ».Ce corpus ayant un impact direct sur l'apprentissage des règles, et par conséquent sur la classification, il est nécessaire que les exemples soient représentatifs de l'apprentissage des règles , et par conséquent sur classification, il est nécessaire que les exemples soient représentatifs de l'ensemble des données. Cette hypothèse est généralement
difficile à vérifier. En classification d'opinion, ou plus généralement en classification de textes, les corpus étudiés sont assez restreints car l'étiquetage des exemples
est souvent effectué à la main. Ceci entraîne un coût élevé et ne permet donc pas l'obtention d'un gros corpus d'apprentissage. Pendant notre travail, nous avons utilisés un corpus de plusieurs commentaires sur des articles, recueillis à partir des journaux arabes disponibles sur le net
Une classification pour pouvoir déterminer la polarité des commentaires : Positive, Négative. C'est cela on a appliqué un ensemble des prétraitements manuels sur ce corpus. Exemples des commentaires : La table suivante présente un exemple de chaque type d'opinion que nous devons classifier [13]

2.5 Les approches de classification d'opinion

2.5.1 Approche linguistique

La principale tâche dans cette approche est la conception de lexiques ou dictionnaires d'opinion. L'objectif de ces lexiques ou dictionnaires est de répertorier le plus de mots porteurs d'opinion possible. Ces mots permettent ensuite de classer les textes en deux (positif et négatif) ou trois catégories (positif, négatif et neutre). [34]

5.1.1. Construction du dictionnaire :

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

– la méthode manuelle ;

– la méthode basée sur les corpus

– la méthode basée sur les dictionnaires.

La méthode basée sur les dictionnaires consiste à utiliser des dictionnaires de synonymes et antonymes existants tels que WordNet. Afin de déterminer l'orientation sémantique de nouveaux mots, on utilise ces dictionnaires afin de prédire l'orientation sémantique des adjectifs. Dans le WordNet, les mots sont organisés sous forme d'arbres (voir figure 3.3). Afin de déterminer la polarité d'un mot, ils traversent les arbres de synonymes et d'antonymes du mot et, s'ils trouvent un mot déjà classé parmi les synonymes, ils affectent la même polarité au mot étudié, ou bien la polarité opposée s'ils trouvent un mot déjà classé parmi les antonymes. S'ils ne croisent aucun mot déjà classé, ils réitèrent l'expérience en partant de tous les synonymes et antonymes, et ce jusqu'à rencontrer un mot d'orientation sémantique connue.[22]

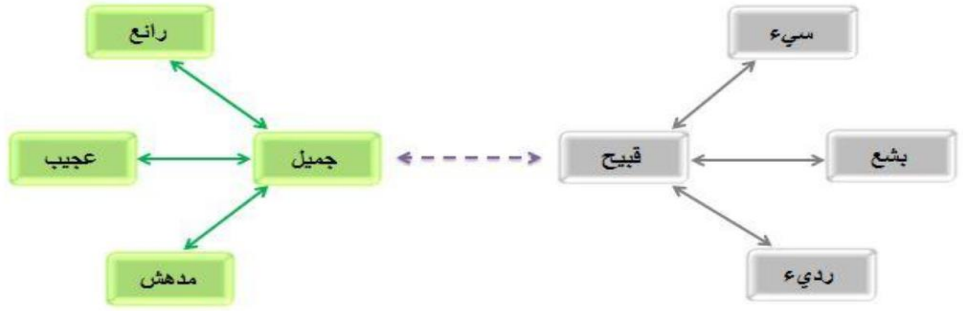


Figure 2.2 Exemple d'arbre de synonymes et antonymes présents dans WordNet (flèche pleine =synonymes, flèche hachurée = antonymes)

La construction du dictionnaire constitué le suivant : Marqueur : la table du marqueur contient tous les prédicats, les adjectifs et les adverbes construisent à partir du corpus avec leur polarité et l'intensité.

- Prédicat : **أحب**aimer, **اكره** détester, **ظن**penser.
- Adjectifs : **جميلة**bien fait, **رائعة**magnifique, **ركيكة**lâche.
- Adverbe : **غنية**riche, **مضجرة**fatigante, **مفيدة**intéressante.
- Intensité : **كثيرا** beaucoup, **جدا**très, **مئة بالمئة**cent pour cent
- Négation : (**لا** non, ni, pas), **لم لن ليس**pas.

2.6. Architecture du système de classification

Le système que nous réalisons a pour but essentiel de permettre la classification d'opinion en langue arabe dans un but de catégorisation et d'indexation. Pour ce faire nous nous proposons de définir un processus de traitement permettant d'avoir en entrée une opinion brut et de présenter en sortie la catégorisation de ce dernier. Cette catégorisation peut se faire par rapport à une référence existante ou par rapport à un autre opinion en entrée. Notre utilisation de la théorie de la distance intertextuelle pour la mise en place d'une métrique de classification nous contraint à l'intégration

d'un processus de lemmatisation des opinions (Prétraitement). Cette étape est nécessaire car elle prépare les opinions en les décomposant ce qui permet l'exploitation des structures grammaticales dans la détection des classes d'équivalence entre segments d'opinion. Nous avons exploité la richesse de la grammaire de la langue arabe pour intégrer la notion

2.7 Implémentation d'une classification :

On peut grouper les méthodes classificatoires en deux grandes familles, cette fois-ci, on prends en considération l'intervention ou non d'un « attribut classe » au fur et à mesure du processus de la classification, ces deux types sont : « supervisée (Classement)» et « non supervisée(Clustering) »,

1. supervisé (classement) : groupes fixés, exemples d'objets de chaque groupe.
2. non supervisé (Clustering) : on ne connaît pas de groupe.

Cependant, Il existe d'autres types de classification qui s'appuient sur d'autres types de méthodes d'apprentissages comme « l'apprentissage semi-supervisé » et « l'apprentissage par Renforcement ». En effet, l'apprentissage semi-supervisé est un bon compromis entre les deux

Types d'apprentissage « supervisé » et « non-supervisé », car il permet de traiter un grand nombre de données sans avoir besoin de toutes les étiqueter, et il profite des avantages des deux types mentionnés. Alors que L'apprentissage par renforcement est fort utilisé dans le cas d'apprentissage interactif,

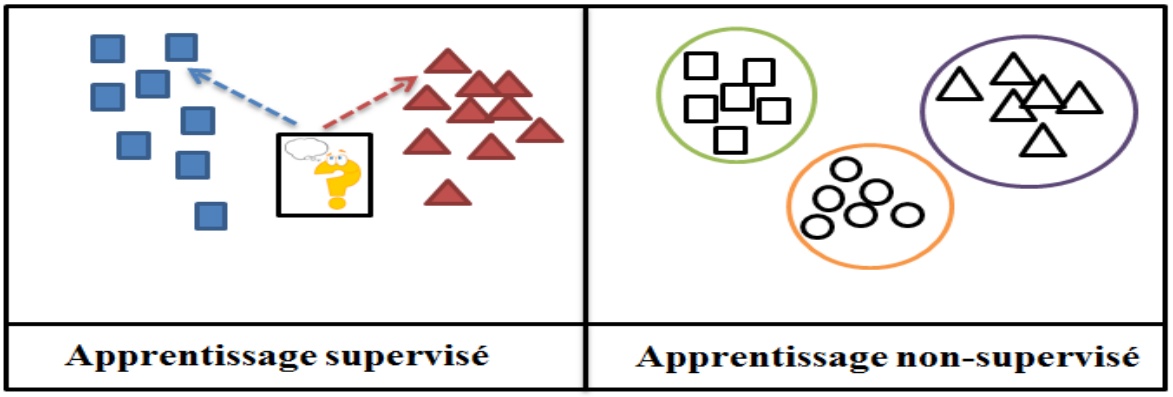


Figure 2.3 : d'apprentissage supervisé et d'apprentissage non supervisé

2.7.1 Classification supervisé :

Cette classification est aussi appelée " Classement"ou encore

" Discrimination".

Est une méthode supervisée qui consiste à définir une fonction qui attribue une ou plusieurs classes à chaque donnée. Dans cette approche on suppose qu'un expert fournit auparavant les étiquettes pour chaque donnée, les étiquettes sont des classes d'appartenance. Selon [Govaert, 2003] : « (la classification supervisée (appelée aussi classement ou a pour objectif « d'apprendre

» par l'exemple. Elle cherche à expliquer et a prédire l'appartenance de documents à des classes connues a priori. Ainsi c'est l'ensemble des techniques qui visent à deviner l'appartenance d'un individu à une classe en s'aidant uniquement des valeurs qu'il prend) ».

La conception supervisée d'un classifieur à C classe (ensemble fini de classe ci) est le fait de classifier N objets (xi) de même nature (des phonèmes, caractères manuscrits, . . .) sachant que ces N objets sont supposés avoir été préalablement « étiquetés » par un « superviseur » en C ensembles qui forme un ensemble d'apprentissage.

Le superviseur n'est qu'un Classifieur en lequel on a confiance (expert humain, caractère répétitif, le système visuel humain . . .), donc notre système de la classification supervisée va être conçu en basant sur les exemples du superviseur (l'ensemble d'apprentissage où pour tout exemple on connaît à priori sa classe.)

C'est-à-dire, on cherche à prédire si un objet (élément) « xi » de la base de données, décrit par un ensemble de descripteurs « d », appartient ou non à une classe « c j » parmi N Classes, pour le faire, on a un ensemble d'apprentissage décrit par :

A = (x1, c2), (x2, c4), (x3, c2) . . . (xi, c j)/xi ∈ R d, c j ∈ C (1.1)

Donc pour chaque objet xi de l'ensemble de données, on peut connaît sa classe a priori c j. La classification supervisée tente de chercher, à partir des données de A, une fonction de décision Γ qui va associer à tout nouvel élément xi de test une classe c j, puis on compare ce que nous a donné cette fonction avec la classe connai a priori de cet élément, de sorte à minimiser les mauvais classements (Γ(xi)6=Cj).

Donc l'objectif est de chercher à prédire la classe de toute nouvelle donnée.

Figure 2.5 schéma classification supervisé

2.7.2 Classification Non supervisé :

Cette classification est aussi appelée "classification automatique", "clustering" ou encore "regroupement".

Dans ce type de classification on est amené à identifier les populations d'un ensemble de données. On suppose qu'on dispose d'un ensemble d'objets que l'on note par :

X = {x1, x2, ..., xN} caractérisé par un ensemble de descripteurs D.

L'objectif du clustering est de trouver les groupes auxquels appartiennent chaque objet x qu'on note par C = {C1, C2, ..., Cn}. Ce qui revient à déterminer une fonction notée Ys^{-} qui associe à chaque élément de X un ou plusieurs éléments de C. Il faut pouvoir affecter une nouvelle observation à une classe. Les observations disponibles ne sont pas initialement identifiées comme appartenant à telle ou telle population. [23]

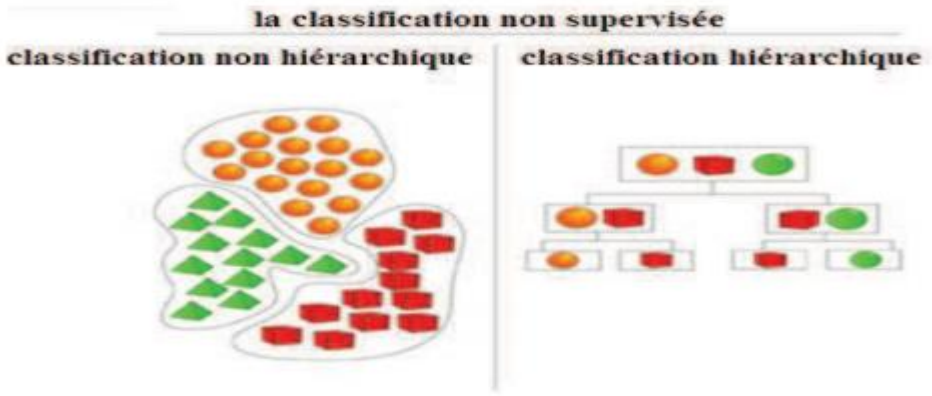


Figure 2.5 : les deux types de clustering hiérarchique/non hiérarchique

2.8 Les Algorithmes de classification non Supervisé :

La classification non supervisée ou « Clustering » est l'une des techniques fondamentales de l'extraction de données structurées ou non structurées.

Plusieurs méthodes ont été proposées : [24]

- Classification hiérarchique: arbre de classes
- Classification hiérarchique ascendante: Agglomérations successives
- Classification hiérarchique descendante: Divisions successives
- Classification à plat: algorithme des k-moyennes: Partition, EM.

2.9 Les Algorithmes de classification Supervisé :

Il existe de nombreuses méthodes d'apprentissage supervisé [25] :

- K plus proches voisins (et ses variantes: Category-based Serch et Cluster-based- Search).
- Arbres de décisions.
- Naive Bayes.
- Réseaux de neurones.
- Machines à support de vecteurs (SVM).
- Programmation génétique.

Ce sont si méthodes-là qui seront utilisé dans notre travail :

Naïve Bayes.

2.9.1 Naïve Bayes :

Le classifieur naïf bayésien est l'une des méthodes les plus simples en apprentissage Supervisé basée sur le théorème de Bayes.[26]

Ce théorème fournit une façon de calculer la probabilité conditionnelle d'une cause sachant la présence d'un effet, à partir de la probabilité conditionnelle de l'effet sachant la présence de la cause ainsi que des probabilités a priori de la cause et de l'effet. [27]

Il est peu utilisé par les praticiens du data mining au détriment des méthodes traditionnelles que sont les arbres de décisions ou les régressions logistiques.

Un avantage de cette méthode est la simplicité de programmation, la facilité d'estimation des paramètres et sa rapidité (même sur de très grandes bases de données). Malgré ses avantages, son peu d'utilisation en pratique vient en partie du fait que ne disposant pas d'un modèle explicite simple (l'explication de probabilité conditionnelle à priori), l'intérêt pratique d'une telle technique et remise en question. [28]

Qui repose sur une hypothèse simplificatrice forte : les descripteurs (Xj) sont deux à deux indépendants conditionnellement aux valeurs de la variable à prédire (Y). Pourtant, malgré cela, il se révèle robuste et efficace. Ses performances sont comparables aux autres techniques d'apprentissage.

Cela, il se révèle robuste et efficace. Ses performances sont comparables aux autres techniques d'apprentissage. [29]

Dans l'approche bayésiennes, on mesure Pr [x|w], c'est-à-dire, la probabilité d'occurrence de l'évènement x si l'évènement w est vérifié : w joue le rôle d'une hypothèse préliminaire que

L'on suppose vérifiée pour estimer la probabilité d'occurrence de l'évènement qui nous intéresse, noté x ici. [30]

Le théorème de Bayes :

Soient A, B et C trois évènements. Le théorème (ou règle) de Bayes démontre que :

Pr [A|B, C] = Pr [B|A, C] P r [A|C] P r [B|C]

Où :

- Pr [B|A, C] est la vraisemblance de l'évènement B si A et C sont vérifiés.
- Pr [A|C] est la probabilité a priori de l'évènement A sachant C.
- Pr [B|C] est la probabilité marginale de l'évènement B sachant C.
- Pr [A|B, C] est la probabilité a posteriori de A si B et C.

Dans cette formulation de la règle de Bayes, C joue le rôle de la connaissance que l'on a. [30]

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

Table 2.1 Table d'abréviations

Algorithme : algorithme de classification par Naïve Bayes

Formation

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

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

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

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

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

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

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

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

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

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

Test

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

Retour$\boldsymbol{c_k}$, où

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

Figure 2.6 Algorithme de Naïve Bayes

2.9.2 K plus proche voisin :

L'algorithme de K plus proche voisin (PPV en bref, K Nearest Neighbors en anglais ou KNN) La méthode des K plus proches voisins est une méthode de l'apprentissage supervisé, dédiée à la classification qui peut être étendue à des tâches d'estimation. Est une méthode d'inférence inductive très efficace pour de nombreux problèmes pratiques. Il est robuste à des données d'entraînement bruyant, et tout à fait efficace lorsqu'il est fourni suffisamment un grand ensemble des données d'apprentissage. [31]

En prenant la moyenne pondérée des K plus proche voisin du point de requête, il peut lisser les effets des exemples isolés d'entraînement bruyants, contrairement à d'autres méthodes statistiques, ne nécessite aucun apprentissage (c'est-à-dire qu'il n'y a aucun modèle à ajuster). C'est l'échantillon d'apprentissage, associé à une fonction de distance et d'une fonction du choix de la classe en fonction des classes voisins les plus proches, qui constitue le modèle.

L'objectif de l'algorithme est de classé les exemples non étiquetés sur la base de leur

Similarité avec les exemples de la base d'apprentissage. Est une méthode de raisonnement à partir de cas.

Son principe est le suivant :

Une donnée de classe inconnue est comparée à toutes les données stockées. On choisit pour la nouvelle donnée la classe majoritaire parmi ses K plus proches voisins (Elle peut donc être lourde pour des grandes bases de données) au sens d'une distance choisie. [32]

Comment identifier le K plus proches voisins ?

- Les instances sont des points dans un espace à *d*-dimensions

d est le nombre d'attributs.

- Une instance xi est définie par son vecteur d'attributs.

< a1(x1), a2(x2),..., ad (xi)>

- Chaque instance a également une catégorie *vi*.
- Identifier les voisins les plus proches de *xi*.
- Trouver les k instances ayant la plus petite distance dis (*xi* , *xj*).
- Similarité : une fonction inverse de la distance.

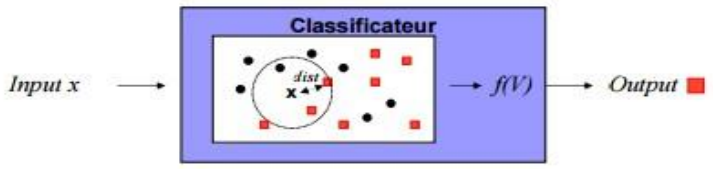


Figure 2.7 classificateur K-PPV **[13]**

Afin de trouver les K plus proches d'une donnée à classer, on peut choisir la distance euclidienne. Soient deux données représentées par deux vecteurs xi et xj, la distance entre ces deux données est donnée par :

$$\boldsymbol{dist(xi,xj)} = \sqrt{\sum} \boldsymbol{(xik - xjk)\ d\ k=1}^{\ 2} \qquad (1)$$

Ou par La distance de Manhattan (valeurs continues) :

$$\boldsymbol{dist(xi,xj)} = \sum |\boldsymbol{(ar(xi) - ar(xj))^2}|\ \boldsymbol{d\ r=1} \qquad (2)$$

Ou par La distance de Hamming (valeurs discrètes) :

$$\boldsymbol{dist(xi,xj)} = \#\ \{\boldsymbol{r \in d : ar(xi) \neq arxj}\} \qquad (3)$$

L'algorithme de K-PPV est utilisé dans de nombreux domaines :

La reconnaissance de formes

- La recherche de nouveaux bio-marqueurs pour le diagnostic.
- Algorithmes de compression.
- Analyse d'image satellite.

Algorithme : algorithme de classification par K-PPV

paramètre : le nombre K de voisin

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

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

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

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

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

5 : **fin pour**

Sortie : le texte t associé à la classe c.

Figure 2.8 Algorithme de K-PPV

2.9.3 Machines à support de vecteurs (SVM) :

Les machines à vecteurs de support, ou SVM (Support Vector Machines), est une méthode de classification binaire par apprentissage supervisé, sont une méthode relativement récente de résolution de problèmes de classification (trier des individus en fonction de leurs caractéristiques), SVM sont un algorithme dont de support dont le but est de résoudre les problèmes de discrimination à deux classes. On appelle problème de discrimination à deux classes dans lequel on tente de déterminer la classe à laquelle appartient un individu (individu ici employé au sens de constituant d'un ensemble) parmi deux choix possibles. [38]

SVM est une méthode de classification qui montre de bonnes performances dans la résolution de problèmes variés. Cette méthode a montré son efficacité dans de nombreux domaines d'applications tels que le traitement d'image, la catégorisation de textes ou le diagnostiques médicales et ce même sur des ensembles de données de très grandes dimensions. [35]

Parmi les modèles des SVM, on constate les cas linéairement séparables et les cas non linéairement séparable. Les premiers sont les plus simples de SVM car ils permettent de trouver facilement le classificateur linéaire. Dans la plupart des problèmes réels il n'y a pas de séparation linéaire possible entre les données, le classificateur de marge maximale ne peut pas

être utilisé car il fonctionne seulement si les classes de données d'apprentissage sont linéairement séparables. [35]

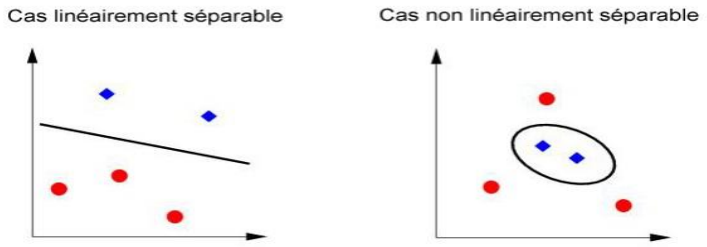


Figure 2.9 les cas linéairement séparables et les cas non linéairement séparable. [13]

2.10 Les critères de mesure des performances des algorithmes

Afin d'évaluer les performances d'algorithmes de recherche d'information, les chercheurs se sont dotés d'outils de mesure auxquels appartiennent les taux de rappel de précision et de f- mesure. Ces critères de performance permettent de quantifier l'aptitude d'un système à trouver des résultats complets et pertinents.

2.10.1 Rappel

Le rappel mesure la capacité du système à identifier tous les documents valides, il donne le pourcentage de réponses correctes renvoyées parmi tous les documents pertinents de la base de données. Ceci implique de connaitre effectivement toutes les réponses pertinentes de la base, ce qui n'est pas réaliste pour des bases quelconques et n'est donc réalisable que sur des bases construites pour évaluer des systèmes de recherche. [36]

$$\text{Rappel}_{\,i} = \frac{\text{le nombre de Documents correctement attribués à la classe}_i}{\text{le nombre de documents appartenant à la classe}_{\,i}} \qquad (4)$$

2.10.2.Précisin:

La précision mesure la capacité du système à trouver des documents valides. Elle donne le pourcentage de réponses correctes parmi les résultats obtenus. [37]

$$\boldsymbol{Précision_i = \frac{Le\ nombre\ de\ documents\ correctement\ attribués\ à\ la\ classe_{\,i}}{le\ nombre\ de\ documents\ classé\ par\ le\ système}} \quad (5)$$

Il est théoriquement possible d'avoir un rappel de 100% en renvoyant la liste de tous les documents de la base, mais la précision sera mauvaise et il sera difficile à l'utilisateur de gérer l'ensemble des résultats retournés. Si un seul résultat pertinent est renvoyé, la précision est excellente. Mais le rappel sera mauvais, Afin de quantifier un compromis, la F-mesure moyenne harmonique a été introduite.

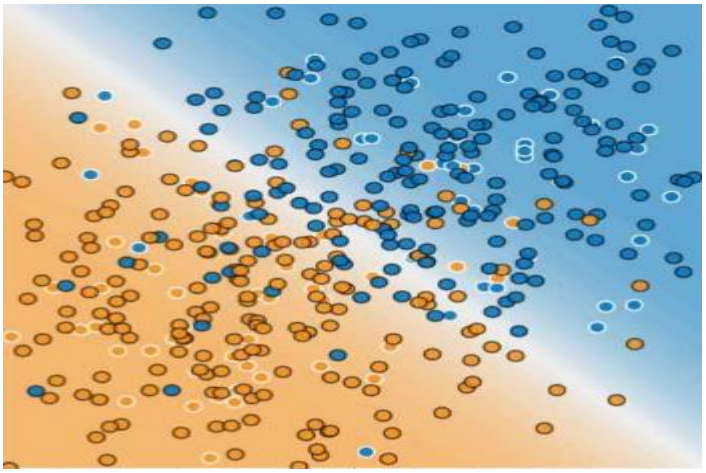
$$Err = \frac{FP+ FN}{VP+VN+FP+FN} = 1 - Acc \qquad (8)$$

2.11 Techniques d'évaluation d'un classificateur :

Après un classificateur est construit, il doit être évalué l'exactitude, il existe de nombreuses façons pour évaluer un classificateur et il y a aussi de nombreuses mesures.

2.11.1 Ensemble des tests :

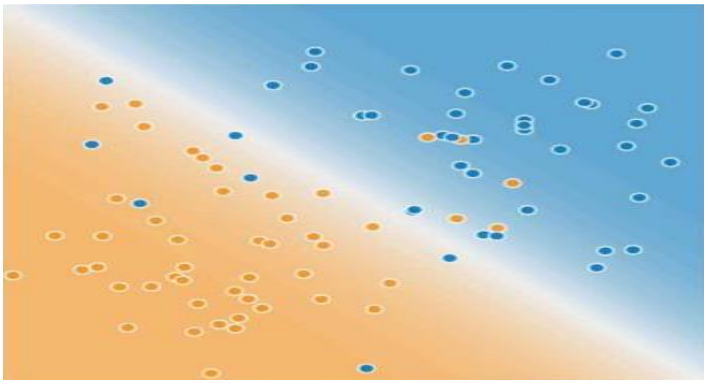
Est un ensemble d'exemples utilisés uniquement pour évaluer la performance d'un classificateur. [39]



Figuer 2.10 : Donne de test

2.11.2 Ensemble d'apprentissage :

Est utilisé pour l'apprentissage d'un classificateur.[39]



Figuer 2.11 : Donne d'apprentissage

2.12. Conclusion :

Dans ce chapitre on a présenté la tâche de la classification et leur importance dans la fouille d’opinion, on a aussi situé les types d’apprentissages automatique qui se résument en deux types, apprentissage supervisé qui inclut plusieur méthode et algorithme de classification nous avons aperçu les méthode le plus connus (Naive Bayes).

3.1 Introduction

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

La classification bayésienne est utilisée comme méthode d'apprentissage probabiliste. Les classificateurs de Naïve Bayes sont parmi les algorithmes connus les plus réussis pour apprendre à classer des documents de texte. Dans ce chapitre, on vise brièvement les outils et les moyens utilisés pour implémenter la classification textuelle Naïve Bayes.

3.2 Opinion de document

Pour démontrer l'utilisation de nos méthodes proposées, nous devons choisir les domaines dotés de certaines fonctionnalités et disponibles sur le site Web avec des critiques en langue arabe. Les données ont donc été sélectionnées dans trois domaines différents, à savoir Mobilis, Djezzy et Ooredoo.

Domain	Positif	Négatif	Source
Mobilis	200	200	www.Mobilis.dz(Facebook)
Djezzy	200	200	www.djezzy.dz(Facebook)
Ooredoo	200	200	www.Ooredoo.dz(Facebook)

Tableau (3.1): Number of positive and negative class with their source

3.3 Traitement

Le prétraitement est une étape nécessaire pour notre méthode. Il existe des données non pertinentes et incorrectes, car nous appliquons un certain nombre de techniques de prétraitement pour atteindre notre objectif. Les étapes que nous avons utilisées sont les suivantes :

- La définition du problème et identification des buts
- La préparation des données
- Le traitement linguistique
- Nettoyage des données
- Lemmatisation
- L'étude lexico étrique
- • Le traitement des données (techniques de Data Mining)

3.4 Présentation de l'approche de classification

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

3.4.1 Machine Learning

L'utilisation d'algorithmes d'apprentissage machine implique souvent un réglage minutieux des paramètres d'apprentissage et des paramètres hyper de modèle. Malheureusement, cet accord est souvent un « art noir» nécessitant une expérience experte, des règles empiriques, ou parfois une recherche en force brute. Il existe donc un grand intérêt pour les approches automatiques capables d'optimiser les performances d'un algorithme d'apprentissage donné au problème considéré. Dans ce travail, nous considérons ce problème à travers le cadre de l'optimisation bayésienne[19]

3.4.2L'apprentissage supervisé (*supervised Learning* en anglais)

Est une tâche d'apprentissage automatique consistant à apprendre une fonction de prédiction à partir d'exemples annotés, au contraire de l'apprentissage non supervisé. On distingue les problèmes de régression des problèmes de classement[1]. Ainsi, on considère que les problèmes de prédiction d'une variable quantitative sont des problèmes de régression tandis que les problèmes de prédiction d'une variable qualitative sont des problèmes de classification.

Les exemples annotés constituent une base d'apprentissage, et la fonction de prédiction apprise peut aussi être appelée « hypothèse » ou « modèle ». On suppose cette base d'apprentissage représentative d'une population d'échantillons plus large et le but des méthodes d'apprentissage supervisé est de bien *généraliser*, c'est-à-dire d'apprendre une fonction qui fasse des prédictions correctes sur des données non présentes dans l'ensemble d'apprentissage.

3.4.3 naive Bayes

"In machine learning, **naive Bayes classifiers** are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features".

La classification naïve bayésienne est un type de classification Bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses. Elle met en œuvre un classificateur bayésienne naïf, ou classificateur naïf de Bayes, appartenant à la famille des classificateurs linéaires [40].

Un terme plus approprié pour le modèle probabiliste sous-jacent pourrait être « modèle à Caractéristiques statistiquement indépendantes » [40].

En termes simples, un classificateur 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 classificateur bayésien naïf déterminera que le fruit est une pomme en considérant indépendamment ces caractéristiques de couleur, de forme et de taille [13]. Selon la nature de chaque modèle probabiliste, les classificateurs bayésiens naïfs peuvent être entraînés efficacement dans un contexte d'apprentissage supervisé [40].

Dans beaucoup d'applications pratiques, l'estimation des paramètres pour les modèles bayé siens naïfs repose sur le maximum de vraisemblance. Autrement dit, il est possible de travailler avec le modèle bayésien naïf sans se préoccuper de probabilité bayésienne ou utiliser les méthodes bayésiennes [40].

Malgré leur modèle de conception « naïf » et ses hypothèses de base extrêmement simplistes, les classificateurs bayésiens naïfs ont fait preuve d'une efficacité plus que suffisante dans beaucoup de situations réelles complexes. En 2004, un article a montré qu'il existe des raisons théoriques derrière cette efficacité inattendue [42]. Toutefois, une autre étude de 2006 montre que des approches plus récentes (arbres renforcés, forêts aléatoires) permettent d'obtenir de meilleurs résultats [41].

L'avantage du classificateur bayésien naïf est qu'il requiert relativement peu de données d'entraînement pour estimer les paramètres nécessaires à la classification, à savoir moyennes et variances des différentes variables. En effet, l'hypothèse d'indépendance des variables permet de se contenter de la variance de chacune d'entre elle pour chaque classe, sans avoir à calculer de matrice de covariance [40].

3.4.4 Description du modèle Bayésienne

Le modèle probabiliste pour un classificateur est le modèle conditionne (\ ,) où C : est une variable de classe dépendante dont les instances ou classes sont peu nombreuses, conditionnée par plusieurs variables caractéristiques , [40].

Lorsque le nombre de caractéristiques est grand, ou lorsque ces caractéristiques peuvent prendre un grand nombre de valeurs, baser ce modèle sur des tableaux de probabilités devient impossible [40].

Par conséquent, nous le dérivons pour qu'il soit plus facilement soluble. À l'aide du théorème de Bayes, nous écrivons :

$$p(C\backslash F1, \dots\dots, Fn) = \frac{P(C)P(F1,\dots\dots\dots Fn\backslash C)}{P(F1,\dots\dots\dots Fn)} \qquad (3.1)$$

$$Postérieure = \frac{antérieure \times vaaisemblance}{évidence} \qquad (3.2)$$

En langage courant, cela signifie :

En pratique, seul le numérateur nous intéresse, puisque le dénominateur ne dépend pas de C et les valeurs des caractéristiques sont données. Le dénominateur est donc en réalité constant. Le numérateur est soumis à la loi de probabilité à plusieurs variables [40].

(c ,F_1,F_2 ,.....F_n) (3.3)

Et peut-être factorisé de la façon suivante, en utilisant plusieurs fois la définition de la probabilité conditionnelle :

P(C, F1, …, Fn) = p (C) p(F1,….,Fn\C)

= p(C) p(F1\C)p(F2,…….,Fn\C,F1)

= p(C) p(F1\C) p(F2\C,F1) p (F3,…..,Fn\C,F1,F2)

= p(C) p(F1\C) p(F2\C,F1) p(F3\C,F1,F2)p(F4,…,Fn\C,F1,F2,F3)

= p(C) p(F1\C) p(F2\C,F1) p(F3\C,F1,F2) …..p (Fn\C,F1,F2,F3,…) (3.4)

C'est là que nous faisons intervenir l'hypothèse naïve : si chaque est indépendant des autres caractéristiques ≠ R alors : Pour tout ≠ R, par conséquent la probabilité conditionnelle peut s'écrire

P(Fi\C, Fj) = p(Fi\C) ……..(3.5)

$$\mathrm{P}(\mathrm{C}, \mathrm{F1}, \ldots, \mathrm{Fn})\mathrm{p}(\mathrm{C})\mathrm{p}(\mathrm{F1}\backslash\mathrm{C})\mathrm{P}(\mathrm{F2}\backslash\mathrm{C})\mathrm{p}(\mathrm{F3}\backslash\mathrm{C}) = \prod_{i=1}^{n} p(F\mathrm{i}\backslash c) \ldots\ldots\ldots\ldots\ldots\ldots (3.6)$$

(Par conséquent, en tenant compte de l'hypothèse indépendance ci-dessus, la probabilité conditionnelle de la variable de classe C peut être exprimée par où :

$$\mathrm{P}\ (\mathrm{C}\backslash\mathrm{F1}\quad \mathrm{Fn}\ = 1/2p(c)\prod_{i=1}^{n} p(F\mathrm{i}\backslash c) \ldots\ldots\ldots\ldots\ldots\ldots (3.7)$$

où Z(appelé « évidence ») est un facteur d'échelle qui dépend uniquement de ,…, à savoir une constante dans la mesure où les valeurs des variables caractéristiques sont connues [40].

Les modèles probabilistes ainsi décrits sont plus faciles à manipuler, puisqu'ils peuvent être factorisés par l'antérieure P(C) (probabilité a priori de C) et les lois de probabilité indépendantes P(Fi|C). S'il existe K classes pour C et si le modèle pour chaque fonction peut être exprimé selon paramètres, alors le modèle bayésien naïf correspondant dépend de (k − 1) + n r k paramètres [40].

Dans la pratique, on observe souvent des modèles où K=2 (classification binaire) et r=1 (les caractéristiques sont alors des variables de Bernoulli). Dans ce cas, le nombre total de paramètres du modèle bayésien naïf ainsi décrit est de 2n+1, avec n le nombre de caractéristiques binaires utilisées pour la classification [40].

3.4.5Estimation de la valeur des paramètres

Tous les paramètres du modèle (probabilités a priori des classes et lois de probabilités associées aux différentes caractéristiques) peuvent faire l'objet d'une approximation par rapport aux fréquences relatives des classes et caractéristiques dans l'ensemble des données d'entraînement. Il s'agit d'une estimation du maximum de vraisemblance des probabilités. Les probabilités a priori des classes peuvent par exemple être calculées en se basant sur l'hypothèse que les classes sont équiprobables (i.e chaque antérieure = 1 / (nombre de classes)), ou bien en estimant chaque probabilité de classe sur la base de l'ensemble des données d'entraînement (i.e antérieure de C= (nombre d'échantillons de C) / (nombre d'échantillons total) [40].

Pour estimer les paramètres d'une loi de probabilité relative à une caractéristique précise, il est nécessaire de présupposer le type de la loi en question ; sinon, il faut générer des modèles non-paramétriques pour les caractéristiques appartenant à l'ensemble de données d'entraînement. Lorsque l'on travaille avec des caractéristiques qui sont des variables aléatoires continues, on suppose généralement que les lois de probabilités correspondantes sont des lois normales, dont on estimera l'espérance et la variance [13]. L'espérance, μ, se calcule avec :

$$\mu = 1/n \sum_{i=1}^{N} xi \quad (3.8)$$

Où N est le nombre d'échantillons et xi est la valeur d'un échantillon donné. La variance, σ^2, se calcule avec :

$$\sigma^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} (xi - \mu)^2 \quad (3.9)$$

Si, pour une certaine classe, une certaine caractéristique ne prend jamais une valeur donnée dans l'ensemble de données d'entraînement, alors l'estimation de probabilité basée sur la fréquence aura pour valeur zéro. Cela pose un problème puisque l'on aboutit à l'apparition d'un facteur nul lorsque les probabilités sont multipliées. Par conséquent, on corrige les estimations de probabilités avec des probabilités fixées à l'avance [40].

3.4.6 Construire

Un classificateur à partir du modèle de probabilités Jusqu'à présent nous avons établi le modèle à caractéristiques indépendantes, à savoir le modèle de probabilités bayésien naïf. Le classificateur bayésien naïf couple ce modèle avec une règle de décision.

Une règle couramment employée consiste à choisir l'hypothèse la plus probable. Il s'agit de la règle du maximum a posteriori ou MAP. Le classificateur correspondant à cette règle est la fonction classificatrice suivante [40] :

$$Classificateur\ (F_1, \ldots\ldots, F_n) = argmax\ p(C = c) \quad (3.10)$$

$$\prod_{i=1}^{n} (Fi = fi \backslash C = c) \quad (3.11)$$

3.4.7Analyse

Fait étonnant, malgré les hypothèses d'indépendance relativement simplistes, le classificateur bayésien naïf a plusieurs propriétés qui le rendent très pratique dans les cas réels. En particulier, la dissociation des lois de probabilités conditionnelles de classe entre les différentes caractéristiques aboutit au fait que chaque loi de probabilité peut être estimée indépendamment en tant que loi de probabilité à une dimension. Cela permet d'éviter nombre de problèmes venant du fléau de la dimension, par exemple le besoin de disposer d'ensembles de données d'entraînement dont la quantité augmente exponentiellement avec le nombre de caractéristiques.

Comme tous les classificateurs probabilistes utilisant la règle de décision du maximum a posteriori, il classifie correctement du moment que la classe adéquate est plus probable que toutes les autres.

Par conséquent les probabilités de classe n'ont pas à être estimées de façons très précises. Le classificateur dans l'ensemble est suffisamment robuste pour ne pas tenir compte de sérieux défauts dans son modèle de base de probabilités naïves. La documentation citée en fin d'article détaille d'autre raisons pour le succès empirique des classificateurs bayésiens naïfs.

3.4.8. Avantage

Cet algorithme dont le modèle d'apprentissage est très général est utilisé dans de nombreux autres domaines que le texte. Il y'a un ensemble d'avantages du classificateur bayésien naïf, parmi lesquelles

- Algorithme facile et simple à implémenter.
- Basé sur une théorie mathématique précise.
- Efficacité et rapidité dans l'apprentissage et la classification.
- Facile à mettre à jour avec de nouveaux exemples d'apprentissage.
- Equivalent à un classificateur linéaire, dans sa rapidité d'application.
  - L'hypothèse d'indépendance des paramètres assouplit l'algorithme pour qu'il soit
  - favorable pour différents types de données
  - Très efficace avec des petits corpus d'apprentissage
  - Résiste au bruit existant dans les données d'entrée
  - Utile pour la classification déterministe comme pour le Rankin puisque il ordonne les classes par degré d'appartenance pour un texte donné
  - Requiert une petite quantité de données d'apprentissage pour estimer les paramètres

  Enfin, le plus important c'est que les méthodes Naïve Bayes donnent de bons résultats [43].

En revanche, l'inconvénient principal à notre avis, c'est bien l'hypothèse d'indépendance entre les descripteurs qui est loin d'être réaliste, mais nous pensons qu'elle n'est pas un handicap majeur dans un contexte de classification.

Tous les avantages cités auparavant et particulièrement la simplicité des calculs, l'efficacité des résultats et la facilité de l'implémentation de cette méthode, au contraire à

d'autres techniques plus sophistiquées gourmandes en ressources et en temps d'exécution avec des taux d'amélioration des résultats très minimes, ont stimulé et justifié le choix du modèle d'indépendance conditionnelle (Naïve Bayes classifier) pour nos travaux.

3.5 PROPOSITION

La figure 3.1 présente le schéma de notre proposition qui consiste en des étages appliqués pour chaque partie (Training Data, Test Data).

La seule différence entre le corpus Training Data et Test Data est les classes connues dans Traning data au contraire pour Test Data

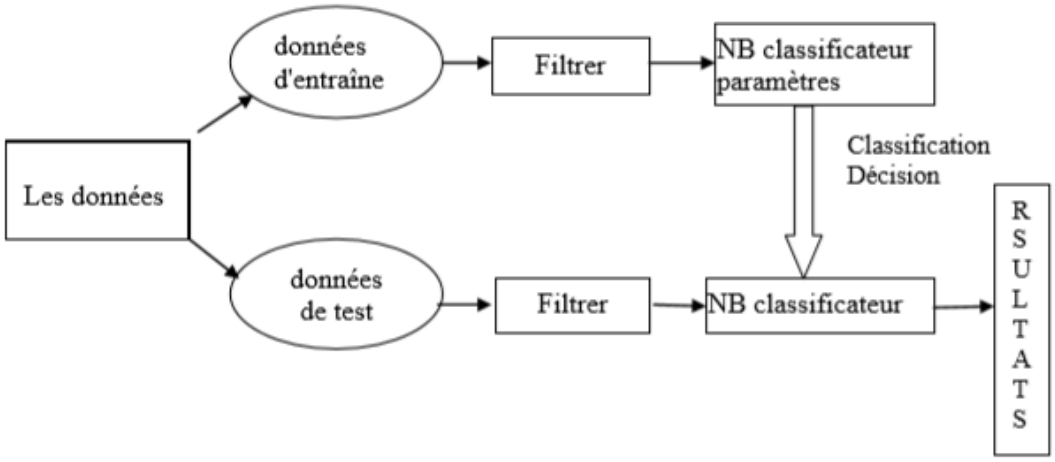


Figure 3.1: Diagramme schématique de l'approche proposée.

3.6 Présentation de notre corpus d'opinion

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

Ils sont souvent faits à la main. Cela entraîne un coût élevé et ne permet donc pas beaucoup d'apprentissage. Au cours de notre travail, nous avons utilisé un ensemble de commentaires sur les trois clients, compilés à partir des pages disponibles sur Internet (Facebook, Instagram, Twitter). Le groupe a abordé différents sujets (services, offres, plaintes, privilèges, ...). Notre cible a été notée pour pouvoir déterminer la polarité des commentaires : positive, négative. C'est ainsi que nous avons appliqué une gamme de traitements pour les mains à ce groupe. Exemples de commentaires : Le tableau suivant fournit un exemple de chaque type d'opinion qu'il faut classifier.

	Polarité	Les commentaires en arabe
01	Positive	شكرا لي موبليس على ما تقدمونه من خدمات
02	Positive Fort	مبروك علينا وعليكم لقد احسنت الاختيار فمتعامل موبليس فهي حقا متمكنة في مجال الاتصالات اعانها الله ووفقها
03	Négative	عروض جازي غير ملائمة مع الشبكة
04	Négative Fort	صراحة لم أحب أسلوب اوريدوا في التعامل مع الزبائن

Tableau 3.2 : Exemples des commentaires.

3.8 Conclusion

Nous avons essayé, tout au long de ce chapitre, de présenter la technique de l'approche linguistique. Cette technique consiste à construire un d'algorithme naïf bayes d'opinion automatique avec l'aide de techniques simples de Traitement Automatique des Langues. Cet algorithme permet ensuite de classer les textes d'opinion selon leur polarité, positive ou négative. Dans le chapitre suivant, nous décrivons enfin le système résultant de notre réalisation par notre travail.

4.1 Introduction :

Nous présentons dans ce chapitre, les outils exploités pour le développement du logiciel tels que le choix du langage de programmation, l'environnement de programmation, ainsi que l'ensemble des résultats des expérimentations par toutes les approches proposées. Nous terminons par une conclusion. Nous sommes Laissez-nous offrir l'aspect pratique de notre demande, et notre objectif est de catégoriser opinion Automatiquement sans ingérence ni connaissance préalable en utilisant L'algorithme de classification est "naïve bayes" car cette technique est produite.

Des groupes homogènes dans un temps très court par rapport à l'autre version des moyens, nous commençons par la description de la règle utilisée, la sélection de l'environnement Ainsi que les étapes de base dans la conception de notre demande. Cette Il porte le nom « Pro classification »

4.2 Le système d'exploitation :

- L'environnement WINDOWS 10 a été choisi comme environnement de travail pour notre logiciel pour les raisons suivantes.
- Une très bonne gestion de mémoire.
- Une architecture orientée évènement.
- Un graphisme indépendant des périphériques.
- La notion de ressources.

4.3 Les Outils de Développements

- ✓ Un PC i5 et 4Go de RAM
- ✓ Microsoft Office 2018 Professionnel
- ✓ NetBeans IDE 8.0.
- ✓ SceneBuilder-8.3.0

4.4 Language de programmation

4.4.1 Introduction à java :

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

4.4.2 SceneBuilder-8.3.0

JavaFX Scene Builder (Scene Builder) vous permet de concevoir rapidement des interfaces utilisateur d'applications JavaFX en faisant glisser un composant d'interface utilisateur à partir d'une bibliothèque de composants d'interface utilisateur et en le déposant dans une zone d'affichage du contenu. Le code FXML de la présentation d'interface utilisateur créée dans l'outil est généré automatiquement en arrière-plan. Pour en savoir plus sur les fonctionnalités de Scene Builder, voir le Guide de l'utilisateur de JavaFX Scene Builder. Scene Builder peut être utilisé comme outil de conception autonome, mais également avec des IDE Java afin que vous puissiez utiliser l'EDI pour

écrire, créer et exécuter le code source du contrôleur que vous utilisez avec l'interface utilisateur de votre application. Bien que Scene Builder soit plus étroitement intégré à l'EDI NetBeans, il est également intégré à un autre IDE Java décrit dans ce document. L'intégration vous permet d'ouvrir un document FXML à l'aide de Scene Builder, d'exécuter les exemples de Scene Builder et de générer un modèle pour le fichier source du contrôleur.

4.5 Stockage de données dans des fichiers txt

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

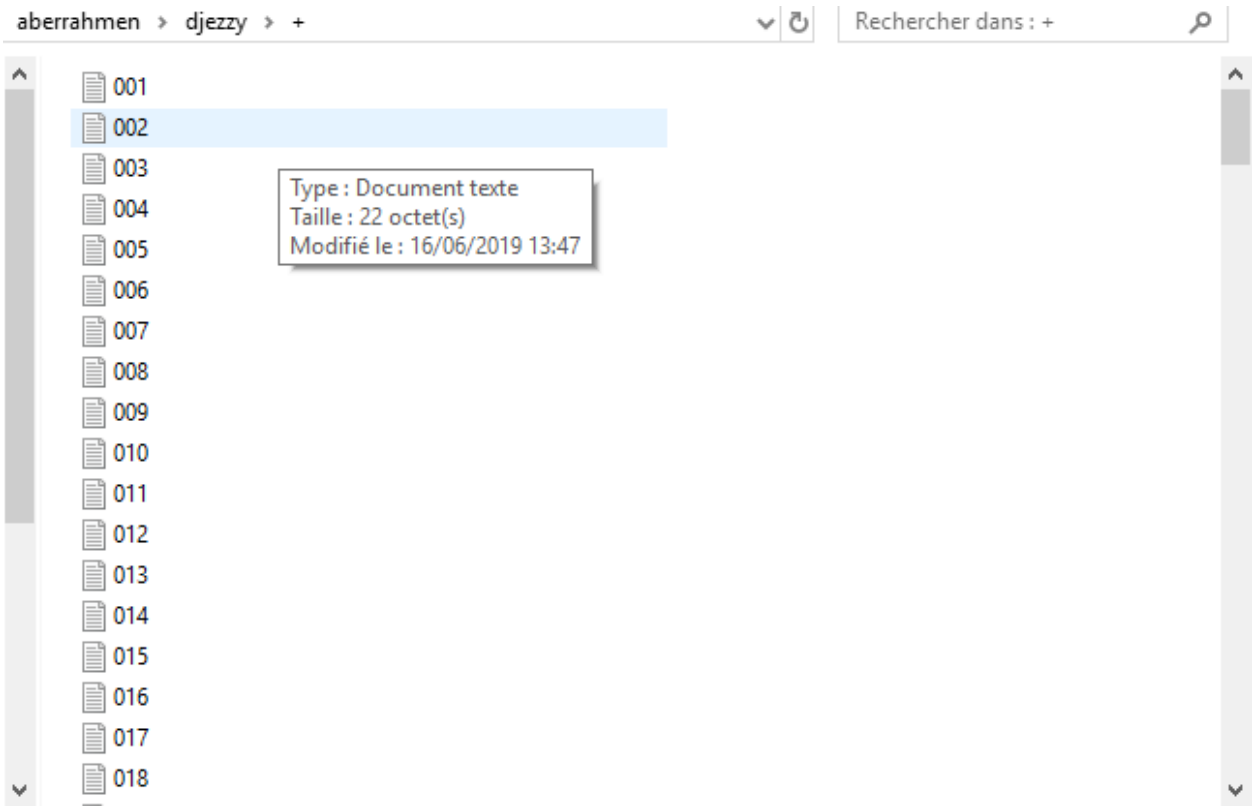
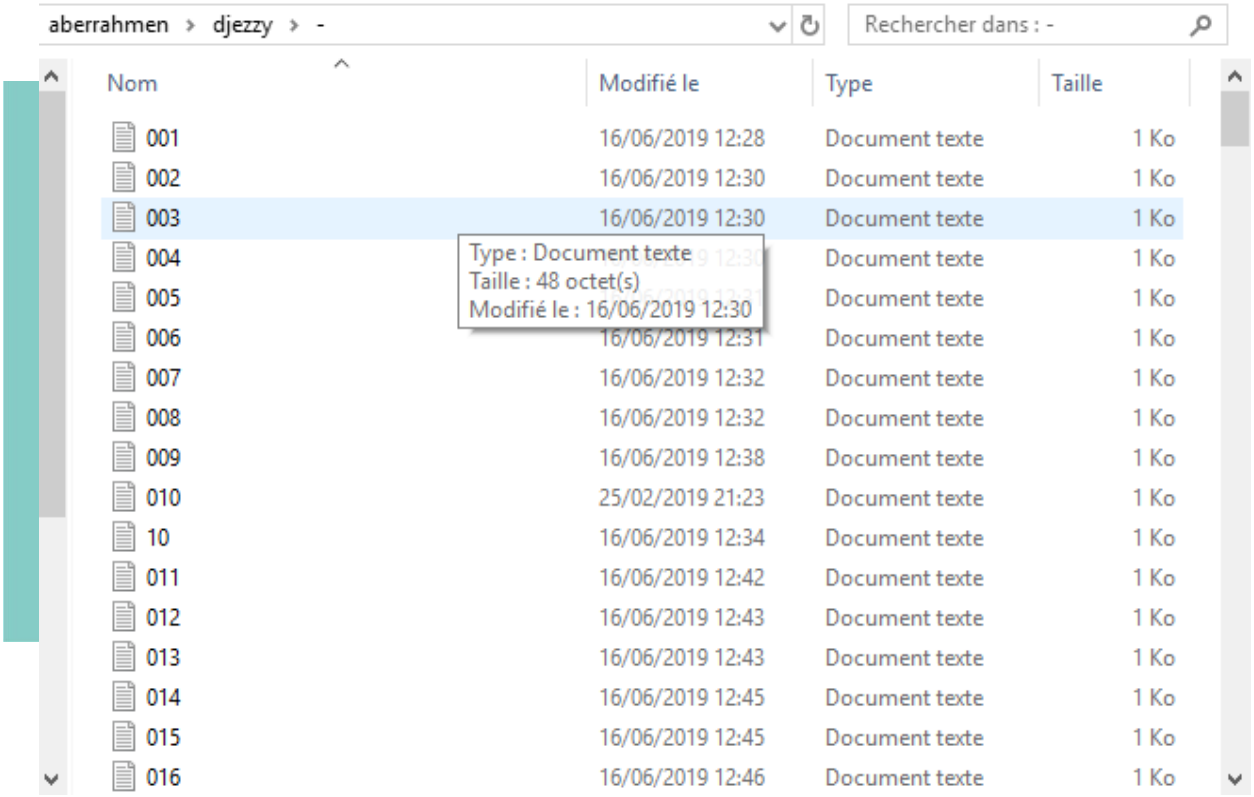


Figure 4.1: Exemple positif de test sur des page official de Mobilis, Ooredoo et Djezzy

Figure 4.2 : Exemple négatif de test sur des page official de Mobilis, Ooredoo et Djezzy



4.6 Diagramme de cas d'utilisation

Figure 4.3 : Diagramme de cas d'utilisation d'application

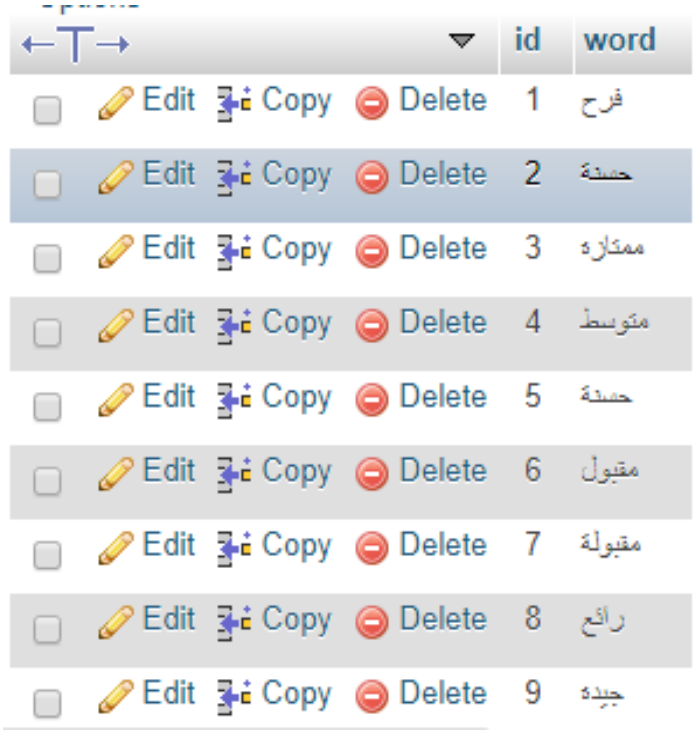
4.7 Interface principale

Figure 4.4 : interface Principale

	id	word
Edit Copy Delete	17	الاستهزاء
Edit Copy Delete	18	الاستخفاف
Edit Copy Delete	19	مهزلة
Edit Copy Delete	20	مهزلة
Edit Copy Delete	21	عير
Edit Copy Delete	22	ليس
Edit Copy Delete	23	مرفوض
Edit Copy Delete	24	لا
Edit Copy Delete	25	هروب
Edit Copy Delete	26	غضب
Edit Copy Delete	27	حزن
Edit Copy Delete	28	انزعاج
Edit Copy Delete	29	انحصار
Edit Copy Delete	30	سارق
Edit Copy Delete	31	انحصار
Edit Copy Delete	32	سرقة

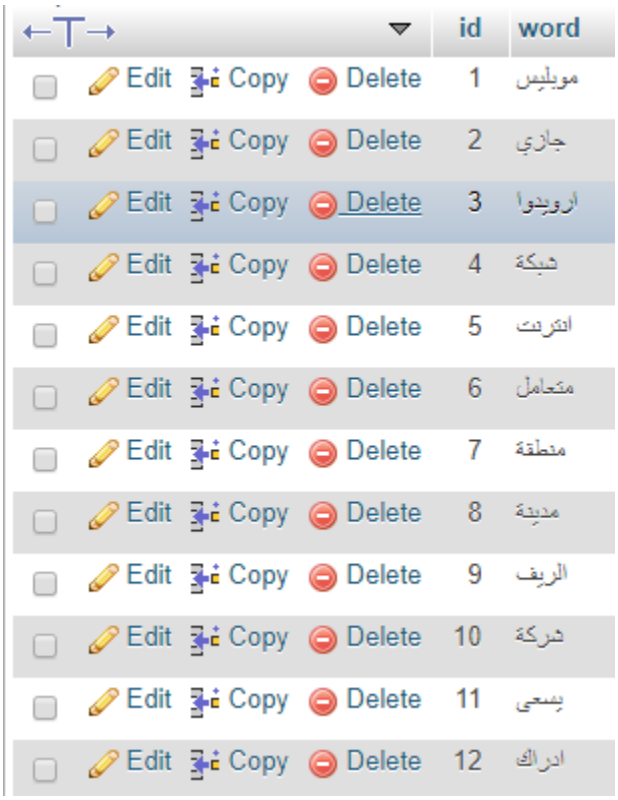
onsole

Table 4.5 des mots négative



	id	word
Edit Copy Delete	1	فرح
Edit Copy Delete	2	حسنة
Edit Copy Delete	3	ممتازه
Edit Copy Delete	4	متوسط
Edit Copy Delete	5	حسنة
Edit Copy Delete	6	مقبول
Edit Copy Delete	7	مقبولة
Edit Copy Delete	8	رائع
Edit Copy Delete	9	جيده

Table 4.6des mots possitive



	id	word
Edit Copy Delete	1	موبليس
Edit Copy Delete	2	جازي
Edit Copy Delete	3	اوريدوا
Edit Copy Delete	4	شبكة
Edit Copy Delete	5	انترنت
Edit Copy Delete	6	متعامل
Edit Copy Delete	7	منطقة
Edit Copy Delete	8	مدينة
Edit Copy Delete	9	الريف
Edit Copy Delete	10	شركة
Edit Copy Delete	11	يسعى
Edit Copy Delete	12	ادراك

Table 4.7des mots neutre

Après avoir créé une base de données dans MY SQL, nous créons des tables pour enregistrer les termes que nous aimerions conserver avec des mots positifs, négatifs ou neutres.

4.8 Ajouter un fichier commentaire :

Après avoir créé la première interface du programme, nous pouvons ajouter un commentaire ou plutôt l'avis d'un client sur l'un des trois revendeurs en Algérie en cliquant sur le bouton Ajouter, après avoir cliqué sur le bouton Ajouter, nous pouvons ajouter deux façons différentes de renouveler le type de client. Nous choisissons l'opinion que nous allons donner parmi une série d'opinions disponibles sur les trois concessionnaires et les renvoyons au type de revendeur, puis nous confirmons et prenons l'opinion du positif ou du négatif.

La deuxième méthode consiste à ajouter notre propre opinion sous forme de récitation et non à écrire

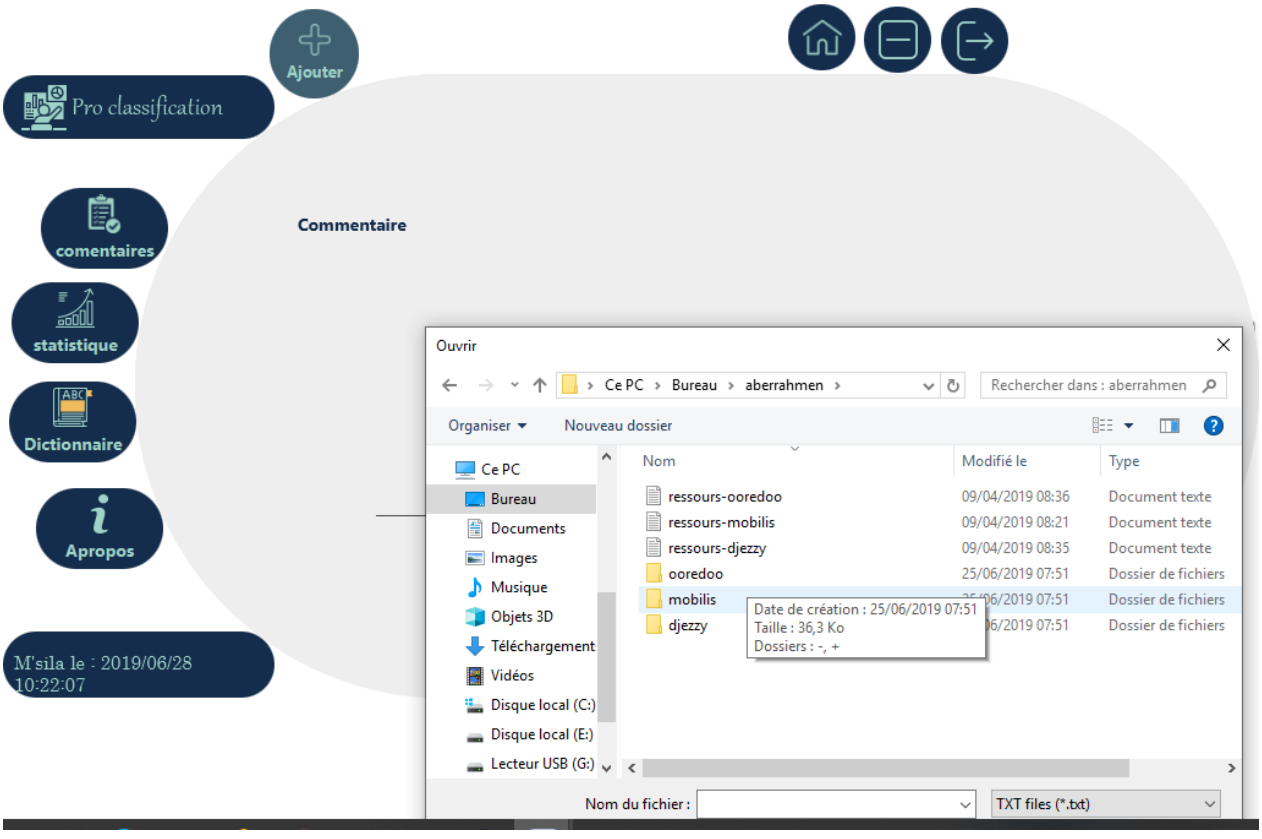


Figure4.5 : Ajouter un commentaire

4.9 Ajouter un mot sur le dictionnaire :

Après avoir créé une base de données dans mon sql, nous la présentons sous l'interface de notre programme sous la forme d'un dictionnaire nous permettant d'ajouter et de supprimer les éléments non significatifs afin que le programme puisse classer les vues en tant que négatif, positif et neutre.

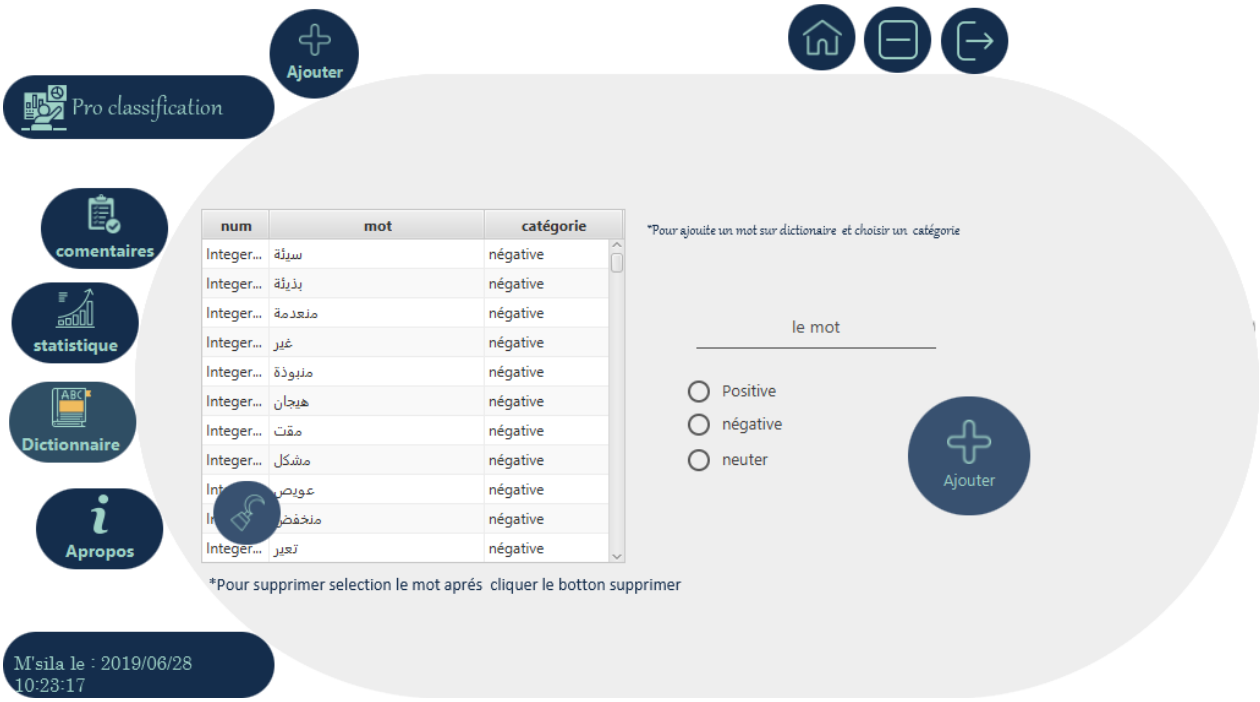


Figure 4.6 : Ajouter un mot sur le dictionnaire

4.10 Analyse les résultats :

4.10.1 Valider et classifier un commentaire (positivé ou négativé) :

Pour valider un commentaire il vaut suivie les images suivantes :

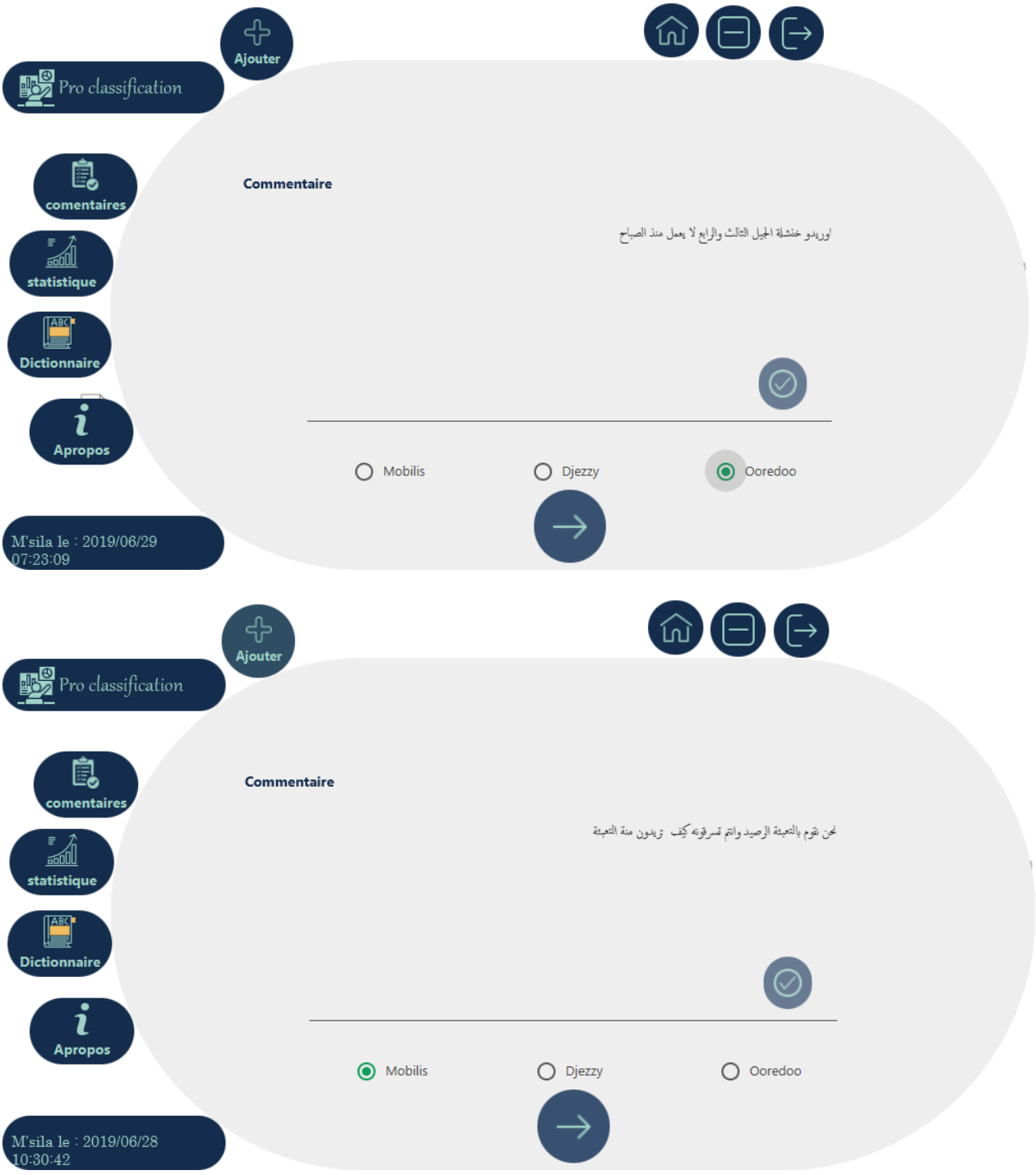


Figure 4.7 : ajoute un commentaire et sélection la société

Figure 4. 8 : pour valider un commentaire

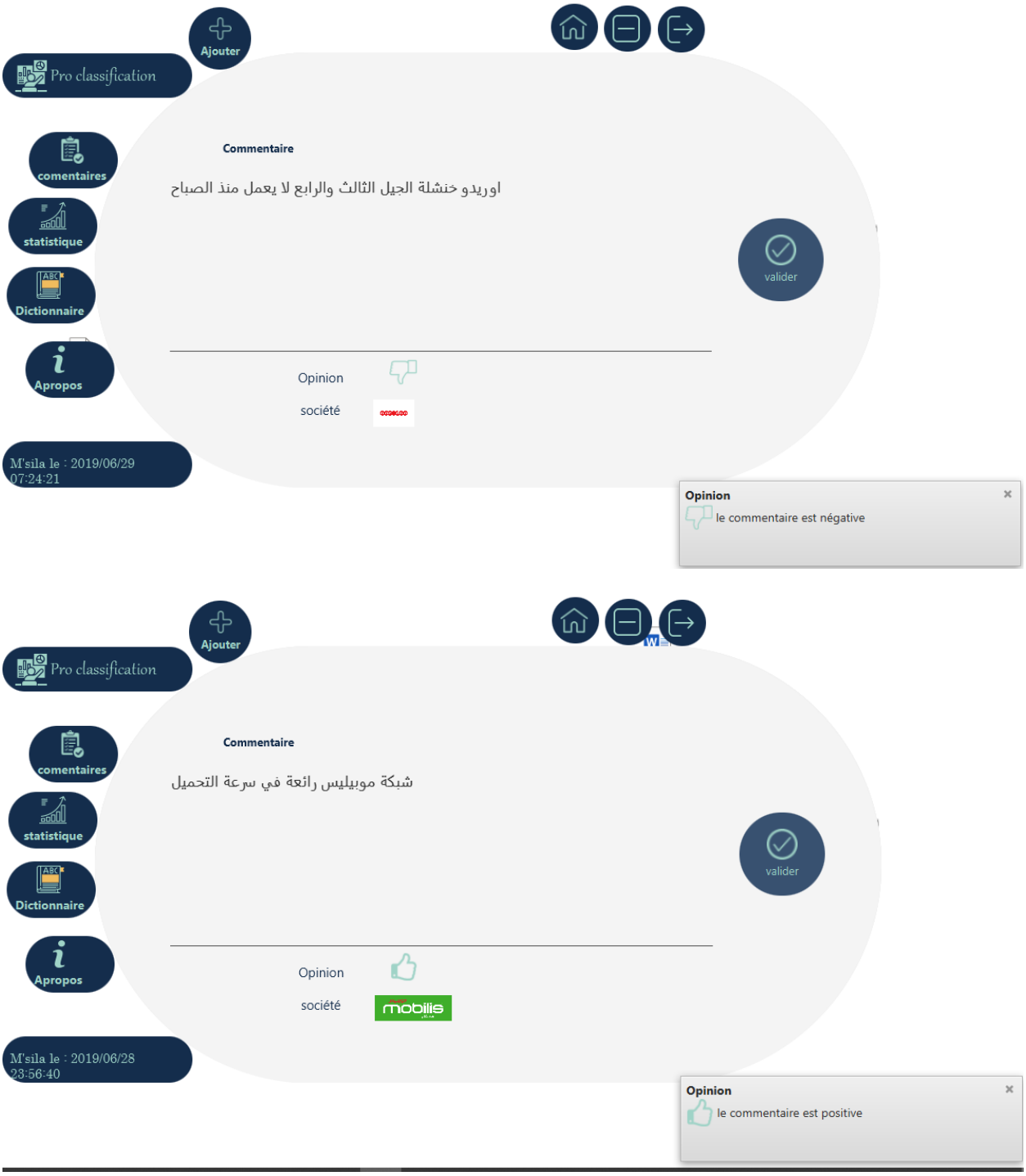


Figure4.9 : les résultats obtenus pour les commentaires

4.11 statistiques les résultats de classification d'opinion de chaque société :

A-Après d'appliquer l'exemple qu'on a créée sur fichier texte dans l'application de commentaire d'opinion positif et négatif concernant la page Facebook official de Ooredoo tel que les résultats sont les même pour 200 commentaire positif et 200 commentaire négatif. Dans la figure 8 montre les résultats avec les pourcentage

Figure4.10 : les résultats obtenus pour les commentaires Ooredoo

B-Après d'appliquer l'exemple qu'on a créée sur fichier texte dans l'application de commentaire d'opinion positif et négatif concernant la page Facebook official de Djezzy tel que les résultats presque les même pour 200 commentaire positif et 200 commentaire négatif. Le figure

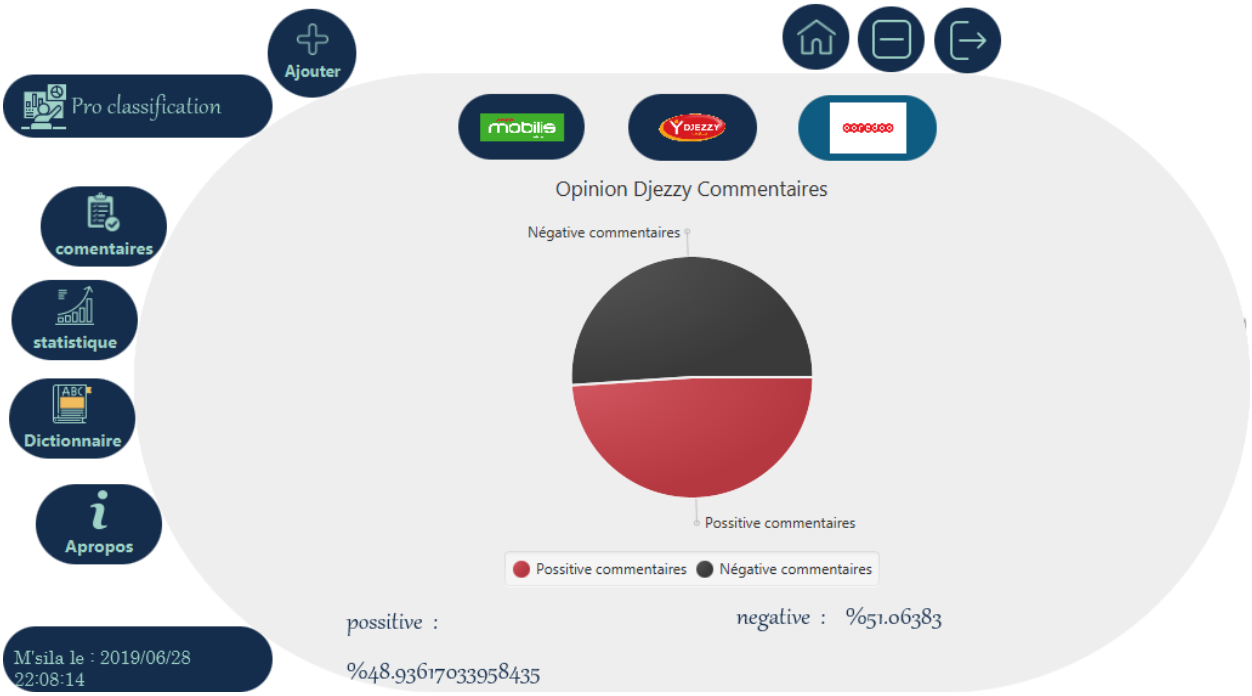


Figure4.11 : les résultats obtenus pour les commentaires Djezzy

C- Après d'appliquer l'exemple qu'on a créée sur fichier texte dans l'application de commentaire d'opinion positif et négatif concernant la page Facebook official de Mobilis tel que les résultats presque les même pour 200 commentaire positif et 200 commentaire négatif. Le figure

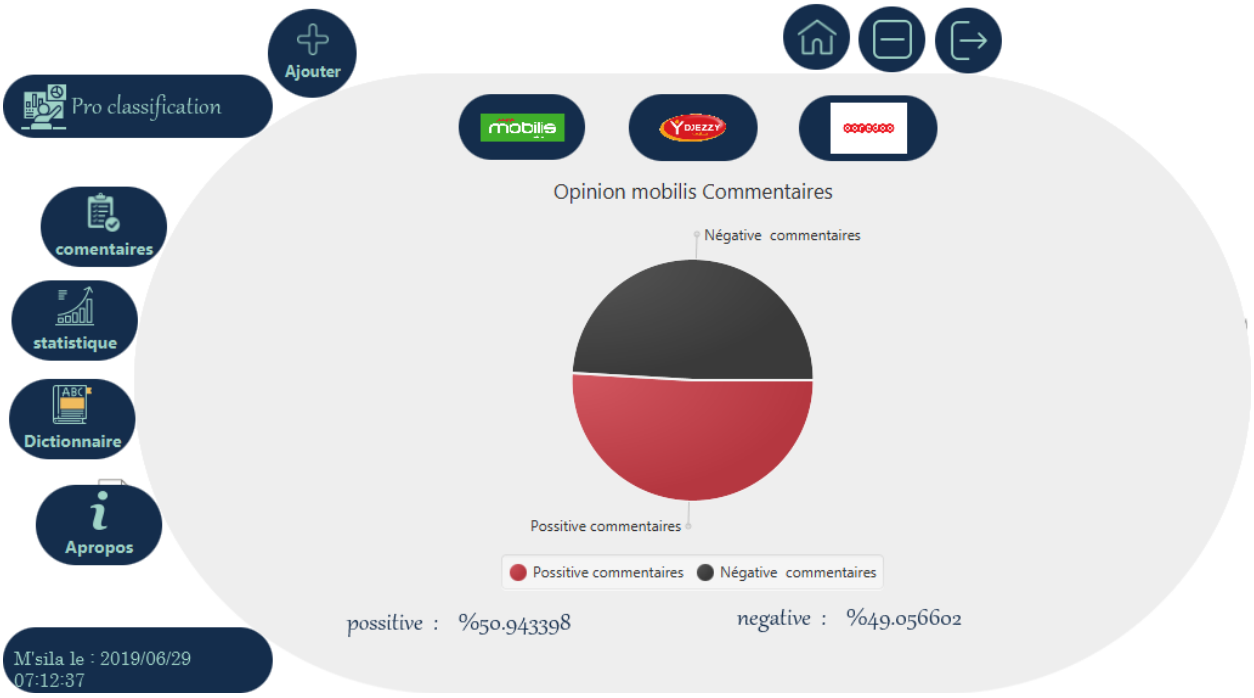


Figure4.12 : les résultats obtenus pour les commentaires Mobilis

4.12 Conclusion :

Dans ce chapitre on a testé un algorithme d'apprentissage supervisé (Naïve Bayes) et appliqué les différentes étapes du prétraitement sur le commentaire des page Sites de réseautage social.

Au cours de ce chapitre, nous avons présenté les expérimentations d'approche proposée avec toutes ses performances, cette approche de classification est l'approche d'apprentissage Machine

Conclusion général

Le domaine de la fouille d'opinion (opinion mining) est un axe de recherche trés active qui vise à faciliter et améliorer l'analyse des commenteires et les critiques des clients concernant un produit ou un service donné. La principale tache est de classifier les textes porteurs d'opinions en catégorie (positive, negative ou neutre).
Nous avons choisis d'étudier les documents d'opinions arabes vu que peu de travaux de recherche sont consacré cette langue.

Durant ce projet, nous pouvons dire que nous avons réaliser notre objectif qui nous a été fixé: concevoir et implementer un systeme pour la classification automatique des documents de critique en langue arabe. Cepandent, il serait interessant d'améliorer quelques aspects comme perspective de travail:

- Utiliser autres algorithmes de classification des textes, à savoir : l'approche Support Vector Machines (SVM) et l'approche de K voisin le plus proche(KNN).
- Etudier l'approche sémantique pour améliorer les performances de la classification, ainsi que l'approche hybride qui combine l'approche supervisée et l'approche sémantique.
- Enrichir le corpus des commentaires en langue arabe.

BIBLIOGRAPHIE

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

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

[3] CHABBOU Fatma Zohra ,BAKHOUCHE Souhaila, « Fouille d'opinions méthodes et outils Étude des méthodes existantes de classification de textes d'opinion» Mémoire de Master, Département de l'informatique, Université de Larbi Tébessi –Tébessa, 2016

[4] Hacéne CHERFI, Etude et réalisation d'un system d'extraction de connaissance à partir d'un texte, Thèse de doctorat, université Henri Poincaré, 15 novembre 2004.

[5] Abdelmalek Amine, Laboratoire Géocode - Université de Saida, cours Data Mining, École d'Hiver sur les applications de l'informatique industrielle, réseaux et génie logiciel, université d'Oran, 09-12 Décembre 2013.

[6] A.Taibi, H.LAZREG, «Utilisation des algorithmes d'apprentissage dans la catégorisation automatique thématique de documents Etude de cas : les algorithmes K_PPV, Naïve Bayes», Mémoire de Licence, Université de M'sila, 2011-2012.

[7] S. RAHEEL, « L'Apprentissage Artificiel pour la Fouille de Données Multilingues : Application à la Classification Automatique des Documents Arabes », Thèse de doctorat en Sciences de l'Information et de la Communication, Université Lumière Lyon 2, 2010.

[8] Taïeb Baccouche L'Information Grammaticale Année 1998 Volume 2 Numéro 1 pp. 49-54 Fait partie d'un numéro thématique : Numéro spécial Tunisie

[09] F. Sebastiano, « Machine learning in automated text categorization », 2002.

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

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

[12] Taghva, K., Elkhoury, R., Coombs, J., "Arabic stemming without a root dictionary", Information Technology: Coding and Computing, ITCC, Vol. 1, pp. 152 – 157, 2005.'

[14] Kanaan G., Al-Shalabi R., Ghwanmeh S., "A comparison of text-classification techniques applied to Arabic text", Journal of the American Society for Information Science and Technology, 60(9), pp. 1836 – 1844, 2009.

[13]LAHRACHE Fatma, « Classification des textes prophétiques », Mémoire de Master, Université Mohamed BOUDIAF– Msila, Algérie, 2015-2016.

[14] Said D., Wanes N., Darwish N., Hegazy N., "A Study of Arabic Text preprocessing methods for Text Categorization", In the 2nd Int. conf. on Arabic Language Resources and Tools, Cairo, Egypt, 2009.

[15] Matallah Hocine ; classification automatique de textes approche orientée agent ; UNIVERSITE ABOUBEKR BELKAID-TLEMCEN FACULTE DES SCIENCES DEPARTEMENT D'INFORMATIQUE .2010-2011

[16] Pio Nardiello Affiliated withMercurioWeb SNC, Fabrizio Sebastiani, Alessandro Sperduti, Discretizing Continuous Attributes in AdaBoost for Text Categorization, Volume 2633 of the series Lecture Notes in Computer Science pp 320-334, 15 April 2003,,» « ch1 » « article ».

[17] http://scholarpedia.org/article/Text_categorization 07/05/2019

[18] Morgane Marchand, « Domaines et fouille d'opinion Une étude des marqueurs multi-

polaires au niveau du texte», Thèse de doctorat en Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), Université Paris-Sud, 2015.

[19]https://en.wikipedia.org/wiki/Machine_learning 07/05/2019

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

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

[22] Gherabi Sara, «CLASSIFICATION AUTOMATIQUE DES TEXTES ARABE (ARABIC OPINION POLARITY) », Mémoire de Master, Université Mohamed BOUDIAF– Msila, Algérie, 2013-2014.

[23] Fatma Karem∗, Mounir Dhibi∗ Arnaud Martin , Combinaison de classification supervisée et non-supervisée par la théorie des fonctions de croyance « ch3 » « article »

[24] LAHRACHE Fatma, « Classification des textes prophétiques », Mémoire de Master, Université Mohamed BOUDIAF– Msila, Algérie, 2015-2016.

[25] Laurent denoue, classification suppervisé de document , 2011,pdf « ch3 » « article ».

[26] Sebastian Raschka ,Naive Bayes and Text Classification I Introduction and Theory/ October 4, 2014

[27] Z Simon, Outils classificatoires par objets pour l'extraction de connaissances dans les bases de donnée. Thèse de doctorat de l'université Henri Poincaré-Nancy 1,Nancy,2000.

[28] http://www.r-bloggers.com/classifieur-naif-bayesien/. 07/05/2019

[29] Tanagra_Naive_Bayes_Classifier_Explained.pdf.

[30] Ph. PREUX, Fouille de données Notes de cours Université de Lille 3 ,26 mai 2011].

[31] Tom M. Mitchell, Machine Learning, (March 1, 1997).

[32] Luc La montagne, Apprentissage a base d'exemple / Concepts avancés pour systèmes intelligents.

[33]Gherabi Sara, «CLASSIFICATION AUTOMATIQUE DES TEXTES ARABE (ARABIC OPINION POLARITY) », Mémoire de Master, Université Mohamed BOUDIAF– Msila, Algérie, 2013-2014.

[34] Gherabi Sara, «CLASSIFICATION AUTOMATIQUE DES TEXTES ARABE (ARABIC OPINION POLARITY) », Mémoire de Master, Université Mohamed BOUDIAF– Msila, Algérie, 2013-2014.

[35] Mohamadally Hasan Fomani, SVM : Machines à Vecteurs de Support ou Séparateurs à Vastes Marges ,16 janvier 2006.

[36] CAMPEDEL Marine, HOOGSTOËL Pierre, Marine Campedel,Pierre Hoogstoël , Sémantique et multimodalité en analyse de l'information] LAVOISIER,2011] .

[37] CAMPEDEL Marine, HOOGSTOËL Pierre, Marine Campedel,Pierre Hoogstoël , Sémantique et multimodalité en analyse de l'information] LAVOISIER,2011] .

[38] Dominiek Francoeur ,Machines A Vecteurs de support une introduction /CaMUS 1 (2010).

[39] Lubing ,Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data ,Second edition ;2011 .

[40] O.CHOAYB, «Classification automatique de textes », Mémoire de Master, Université Mohamed BOUDIAF– Msila, Algérie, 2013-2014.

[41] CARUANA, R. and NICULESCU-MIZIL, A.: "An empirical comparison of supervised learning algorithms». Proceedings of the 23rd international conference on Machine learning, 2006.

[42] HARRY ZHANG,"The Optimality of Naive Bayes". Conference FLAIRS 2004.

[43] H.MATALLAH, «Classificatiion Automatiique de Textes Approche Oriientée Agent», Mémoire de Magister En Informatique, Université Aboubekr Belkaid-Tlemcen, 2011.

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

[45] Encarta Dictionnaire, 2009

[46] Sentiment Lexicons, OpinionFinder: 2006 positive words, 4783 negative words sur l'URL : https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html 07/05/2019

Résumé

La fouille d'opinion est un domaine très intéressent dont le but est d'analyser automatiquement les documents d'opinions et les classer selon leurs contenus en trois catégories (positive, négative, neutre).

Dans notre travail, nous utilisons l'algorithme des naïves bayes NB pour classifier les commentaires. Cette tâche est très intéressante pour beaucoup d'applications tel que e-commerce et le marketing.

Mots clés : Classification d'opinion, analyse de sentiments, fouille d'opinions, naïve bayes.

Abstract

Opinion mining is a very interesting field that aims to automatically analyze opinion documents and classify them according to their content into three categories (positive, negative, neutral).

In our work, we use the naive bayes NB algorithm to classify comments. This task is very interesting for many applications such as e-commerce and marketing.

Keywords: Classification of opinion, sentiment analysis, opinions mining, naive bayes.

ملخص

يعد استخراج الرأي مجالًا مثيرًا للاهتمام للغاية يهدف إلى تحليل مستندات الرأي اليا وتصنيفها وفقًا لمحتواها إلى ثلاث فئات (إيجابية وسلبية ومحايدة).
في عملنا ، نستخدم خوارزمية Naive bayes لتصنيف التعليقات. هذه المهمة مفيدة جدا للعديد من التطبيقات مثل التجارة الإلكترونية والتسويق.

الكلمات المفتاحية: تصنيف الرأي ، تحليل المشاعر ، استخراج الآراء ، السذاجة