REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE: Mathématique et Informatique **DOMAINE: Mathématique et**
DEPARTEMENT: Informatique **FILIERE: Informatique**
N°:…….. **OPTION: S.I.G.L**

Mémoire présenté pour l'obtention
Du diplôme de Master Académique
Par : ABDELSALAM WISSAME
CHIA NASSIMA
Intitulé

Classification des texts arabes

Soutenu devant le jury composé de :

B.BELKACEM	Université de M'sila	Rapporteur

Année universitaire : 2020 /2021

Remerciements

Tous d'abord, nous tenons à remercier premièrement Dieu de nous avoir accordé toute la détermination, la volonté et la force pour qu'on puisse réaliser ce modeste travail.

En déposant ce mémoire, nous tenons tout d'abord à exprimer notre profonde gratitude envers notre encadreur, Mr-BELKACEM BRAHMI, pour sa confiance, ses encouragements continuels, et son suivi de prés de nos travaux en dirigent ce mémoire.

Je voudrai également remercier et exprimer mon profond respect à tous les membres du jury qui ont bien voulu me faire l'honneur d'assister à ma soutenance du Master afin de juger la qualité du travail réalisé.

Mes remerciements vont enfin à toute personne qui a contribué de près ou de loin à l'élaboration de ce travail.

Dédicace

Je dédie Ce travail:
A mes chers parents
À Mon Marie,
À mes enfantes (amine et malek),
À tous ma famille,
À tous ceux qui, de près ou de loin, ont soutenu mon travail par leurs encouragements et conseils.

Chia nassima

Dédicaces

Je dédie ce travail :

A mes chers parents,
Pour leur soutien, leur patience, leur encouragement durant mon parcours scolaire
je vous remercie pour tout ce que vous avez fait pour moi.

A mes sœurs, mes frères de tout mon cœur,

Vous êtes ma joie et mon soutien ;

A ma famille et A mes amis

Qui m'ont toujours encouragé,

Wissam.

Résumé

Avec l'avènement du Web 2.0, nous sommes trouvés plusieurs plates-formes pour l'analyse des sentiments qui permettent aux utilisateurs d'exprimer leurs opinions sur un sujet particulier (services, politique, commercial ou productions). Nous avons également remarqué que les avis des clients sur les produits jouent un rôle primordial dans la décision du client d'acheter un produit ou d'utiliser un service.

Dans ce mémoire nous avons conçu et réalisé un système permet de traiter des commentaires édités en arabe. Nous avons utilisé trois algorithmes pour analyser et classer les commentaires personnels sur un sujet particulier dans différents domaines .Les catégories que nous avons définies sont: positives et négatives.

Mots clés: Fouille de textes, Classification d’opinion, Textes arabes, Fouille d’opinion.

Abstract

With the advent of Web 2.0, we have found several platforms for sentiment analysis that allow users to express their opinions on a particular topic (services, politics, business or productions). We have also noticed that customer reviews of products play a huge role in the customer's decision to buy a product or use a service.

In this thesis we have designed and built a system for processing comments edited in Arabic. We used three algorithms to analyze and classify personal comments on a particular topic in different areas. The categories we defined are: positive and negative. Keywords: Text mining, Opinion classification, Arabic texts, Opinion mining.

ملخص:

مع ظهور Web 2.0 ، وجدنا العديد من المنصات لتحليل المشاعر التي تسمح للمستخدمين بالتعبير عن آرائهم حول موضوع معين (الخدمات أو السياسة أو الأعمال التجارية أو الإنتاج). لقد لاحظنا أيضًا أن مراجعات العملاء للمنتجات تلعب دورًا كبيرًا في قرار العميل بشراء منتج أو استخدام خدمة.

في هذه الرسالة قمنا بتصميم وبناء نظام لمعالجة التعليقات المحررة باللغة العربية. استخدمنا ثلاث خوارزميات وقاموس لتحليل وتصنيف التعليقات الشخصية على موضوع معين في مجالات مختلفة، والفئات التي حددناها هي: إيجابية وسلبية.

الكلمات المفتاحية: التنقيب عن النص ، تصنيف الآراء ، النصوص العربية ، التنقيب عن الرأي.

Liste des figures

Liste des tables :

Table de matière

Introduction générale

Introduction générale

1. Contexte

Avec le développement du Web 2.0, nous sommes devenus dépendants de l'information et de l'analyse dans nos vies. Ces informations ne sont plus disponibles et plus précisément sous forme numérique. Les gens communiquent, partagent du contenu et expriment leurs opinions en ligne sur une variété de sujets, dans des groupes de discussion, des forums, des blogs et d'autres sites liés aux avis sur les produits.

Ces opinions sont devenues une source d'informations importante pour les entreprises à prendre en compte lors du développement de produits et de l'élaboration de plans marketing.

Prenons trois exemples de commentaires des clients sur la caméra:

"Cet appareil photo numérique est mon premier appareil photo et il a été très facile d'apprendre à l'utiliser."

"L'image est superbe et l'obtenir de la bonne exposition est facile."

«Cette caméra a une très grande capacité.»

Dans ces exemples, nous pouvons extraire plusieurs expressions comme « très facile d'apprendre à l'utiliser », «l'image est superbe », « l'obtenir de la bonne exposition est facile », « très grande capacité » qui transmettre l'opinion du client plutôt que des faits. Il est utilisé pour exprimer des opinions positives ou négatives du client concernant les caractéristiques des produits, qui sont dénommées «apprendre à utiliser», «exposition», «capacité» et «image». Sachant que les informations recueillies à partir de plusieurs avis sont plus fiables que les informations d'un seul avis, le tri manuel de grandes quantités d'avis un par un prend du temps et coûte cher aux entreprises et aux clients.

donc, il est plus efficace de fournir les informations nécessaires et traiter automatiquement les différents avis sous une forme résumée.

Si nous devions obtenir le nombre d'avis positifs et négatifs d'un produit particulier par rapport à un exemple, alors classer chaque opinion comme négative ou positive serait la tâche la plus importante. En revanche, si nous voulons montrer les

informations du client sur chacune des différentes caractéristiques d'un produit, il est nécessaire d'analyser le sentiment général de chaque option et d'extraire les caractéristiques des produits.

2. Problématique

Aujourd'hui, la quantité d'informations disponibles sur Internet est énorme et la classification reste à la fois l'une des tâches les plus difficiles et les plus importantes. Une grande partie du travail se concentre actuellement sur l'anglais car c'est la langue dominante sur le Web.

Néanmoins , d'autres langues sont nécessaires car le web devient chaque jour multilingue. Le besoin est plus urgent sur la langue arabes. Nous avons recherché une classification plus détaillée des textes arabes en classant les articles d'opinion arabes selon leur polarité (positive, négative).

3. Objectifs

Le but de ce travail est de fournir une méthode automatisée pour classer chaque commentaire selon l'opinion qu'il exprime un avis positif ou un avis négatif trouvé dans les commentaires rédigés en arabe.

4. Structure de mémoire

Ce mémoire est organisé en quatre chapitres qui peuvent être résumés comme suit :

- ✓ Le **premier chapitre** dans ce chapitre, nous avons parlé principalement sur les notions de la fouille de donnée textuelle et la fouille d'opinion.
- ✓ Le **second chapitre** fait le point sur la présentation des méthodes et approches existantes et generalites sur lange arabe.
- ✓ Le **troisième chapitre** nous présentons en détail la modélisation de notre système.
- ✓ Le **quatrième chapitre** on trouve la partie de réalisation qui est consacré à l'implémentation et l'expérimentation de notre application.

Ce mémoire se conclut par un dernier chapitre contenant une conclusion générale, et quelques perspectives.

Chapitre1
Fouille D'opinion

Chapitre1:Fouille D’opinions

1.1. Introduction

La Fouille de données d’opinions (Opinion Mining) et l’Analyse des Sentiments (Sentiments Analysis), font partie d’un domaine émergent. C'est un sous-domaine de l'informatique qui est considéré comme faisant partie du traitement automatique du langage naturel et a pour but de classifier les sentiments exprimés dans des textes

Nous présentons deux parties principales dans ce chapitre. Dans la première partie du chapitre sur les définitions et les techniques de text mining, nous expliquons la chaîne de traitement de l'extraction de données textuelles. La deuxième partie traite des définitions, du processus d'exploration de données, de l'importance de son utilisation, des types et des tâches d'extraction de diverses données, et des domaines d'application, puis nous mentionnons certaines applications de cette technologie.

1.2.Un peu d’historique

On assiste, Ces dernières années, à une pris de conscience de l'importance de l'opinion sur le web, ce qui explique les nombreux et récents travaux dans ce domaine

les sociétés ont pris en compte les opinions et les préoccupations sentimentales dans les textes, en utilisant les termes «analyse des sentiments» et «recherche d'opinion». Le terme «recherche d'opinions» est apparu pour la première fois dans l'article de Dave et al. (2003).

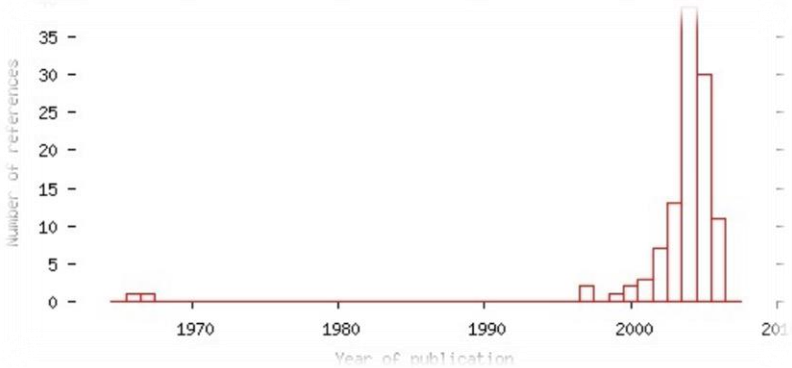


Figure 1.1 : Graphe indiquant l’importance de la détection d’opinions [1].

1.3 Fouille de textes (Text Mining)

1.3.1. Text Mining et Data Mining

Le Data Mining est à la base du Text Mining c'est a dire où celui-ci est l'extension du même but et du même processus vers des données textuelles.

Cependant, les deux technologies se distinguent dans la nature des données à traiter. Le Data Mining s'intéresse aux données numériques et factuelles qui sont bien structurées dans des bases de données, alors que le Text Mining s'intéresse aux données textuelles non structurées, généralement exprimées en langage naturel [2].

1.3.2. Définition de Text Mining

Le Text Mining, également appelé fouille de textes ou extraction de connaissances, est un domaine de recherche dont la première définition est donnée par (R.Feldman, 1995). Le Text Mining, qui est une des disciplines du traitement automatique du langage naturel. L'objectif du Text Mining est d'exploiter l'information à partir de plusieurs documents textuels de différentes manières et de rechercher des relations entre entités textuelles ou entre documents et de découvrir des tendances, des concepts, etc.

Le processus du Text Mining s'effectue en trois étapes :

- Le prétraitement des données.
- L'indexation ou représentation formelle.
- L'analyse des données indexées [3].

1.3.3. Approches du Text Mining

Deux approches non antinomiques sont par la suite envisagées :

1.3.3.1. Approche statistique

Consiste à ne voir le document que via le prisme du nombre et des chiffres. De cette manière, l'outil de fouille de texte statistique génère des informations sur le nombre d'occurrences d'un terme, la cooccurrence de plusieurs termes, ainsi que l'occurrence d'un terme dans un document ou un groupe de textes.

1.3.3.2. Approche sémantique

C'est une méthode d'interprétation automatique de textes écrits en langage naturel, cette méthode se base sur un élément externe, appelé le référentiel. Les référentiels peuvent être des listes à plats, des mots clés, des ontologies ou bien des thesaurus ou peuvent être des logiques de type probabilistes.

1.3.4. Chaîne de traitement pour le processus de fouille de données textuelle

Le schéma suivant présent Chaîne de traitement pour représentées, préparées et organisées les informations d'un texte :

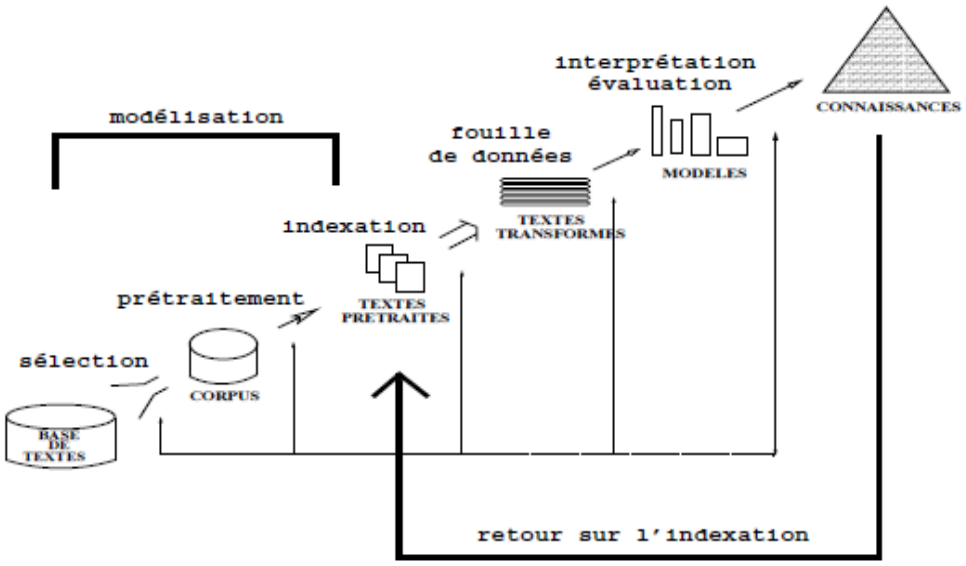


Figure 1.2 La chaîne de traitement pour le processus de fouille de textes [4].

- **Le prétraitement**: C'est une tâche très importante car elle comprend les trois premières étapes du modèle CRISP-DM, de la compréhension du problème et des données à la préparation de ces dernières. Cette phase inclut tous les traitements, les méthodes nécessaires, les processus et pour la préparation des données pour les opérations de bases de la découverte de connaissance du système Text Mining..
- **La modélisation** : Contient les opérations de bases de la fouille de textes qui utilisent les algorithmes de Data Mining pour la découverte de connaissances.
- **L'évaluation** : Pour déterminer ce qu'il faut faire ensuite, on termine le processus dans le cas où les résultats sont bien adaptés à l'application, sinon si le résultat est signifiant mais non satisfaisant, on réitère et le résultat généré sera utilisé comme une partie de l'entrée d'une ou de plusieurs étapes précoces [5].

1.3.5. Tâches principales de la fouille de textes

Nous allons présenter les trois tâches principales gérées par la fouille de textes. Chacune de ces tâches sera un cas particulier du schéma général de la figure1.3, pour lequel nous préciserons :

- Le type des données et des résultats
- Le type des ressources utiles, à titre obligatoire ou facultatif

- Le type des techniques utilisées pour la programmer, et si elle peut être abordée par apprentissage automatique
- les applications concrètes de cette tâche [5].

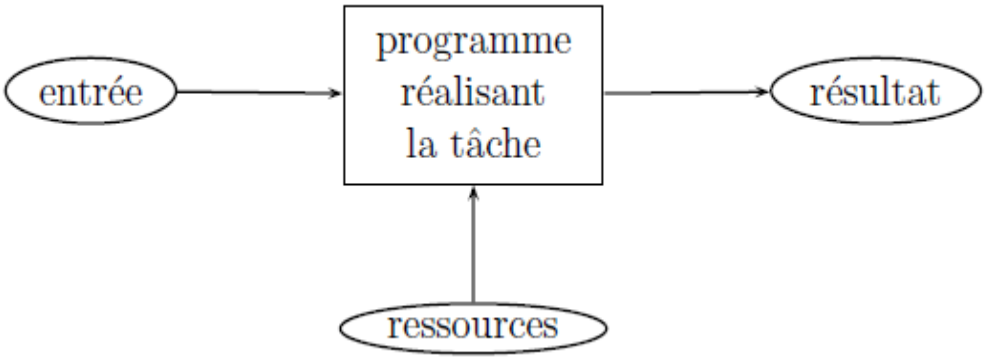


Figure 1.3: schéma général d'une tâche de fouille de textes [6].

1.4.Distinction des deux concepts faits et opinions

Les messages que nous recevons, de proches, d'adultes, de la télévision, des livres, des journaux sont composés d'un ensemble de faits, d'opinions et de sentiments. Il n'est pas toujours facile de savoir quels sont les faits, les opinions et les sentiments. Cette confusion entraîne bien souvent des disputes et peut conduire à des situations de blocage. Toute discussion devrait avoir pour point de départ l'identification des faits, chacun pourrait ensuite exposer ses opinions ou ses sentiments

Il existe deux catégories principales pour classer l'information textuelle:

- Les faits: il s'agit de descriptions objectives (énoncé objectif) sur les entités et les événements dans le monde
- Les opinions: il s'agit d'expressions subjectives d'un individu à propos d'un objet ou d'un sujet particulier.

Certains indices textuels permettre de déterminer l'objectivité du texte informatif qui apporte une information explicite, ses indices sont :

- Une forme, un style et un vocabulaire neutre.
- Eviter les phrases exclamatives et interrogatives directes.

- Eviter l'utilisation de l'impératif.
- L'utilisation de pronoms personnels à la troisième personne, comme «il» ou «on», sauf à l'intérieur des citations où ils ne sont pas obligatoires.
- L'emploi de citations, de références et de statistiques renforcées des affirmations.
- créer des phrases déclaratives [1].

Certains mots énoncent la subjectivité d'un texte, voici quelques indices textuels indiquent la formulation d'opinions personnelles, de jugements, de goûts, de sentiments,

- Un style, un ton et un vocabulaire descriptifs, expressifs, appréciatifs.
- L'utilisation de la phrase exclamative avec de sa justification.
- L'utilisation de pronoms personnels de la première et la deuxième personne à l'intérieur comme à l'extérieur des citations : «je», «tu», «nous» et «vous»
- L'emploi de citations pour renforcer des opinions ou des jugements [1].

1.5.Fouille d'opinions

1.5.1. Définitions

La fouille d'opinions (opinion mining), ou l'analyse de sentiments (sentiment analysis), est un sous-domaine de l'informatique à l'intersection de plusieurs disciplines telles que le traitement automatique du langage naturel, la recherche d'information, la fouille de texte et l'apprentissage automatique. Les termes « fouille d'opinions» et « analyse de sentiments » ont été respectivement introduits dans [14,15] [7]. Le terme « fouille d'opinion » est utilisé pour évoquer le traitement automatique des opinions, des sentiments et de la subjectivité dans les textes. Ce domaine est connu sous les noms d'opinion mining.

1.5.2. Opinion

Il existe plusieurs définitions:

D'après Larousse l'opinion :

- Est un jugement, avis, sentiment qu'un individu ou un groupe émet sur un sujet, des faits, ce qu'il en pense : exprimer son opinion au cours du débat.

L'opinion des critiques synonymes (avis, conviction, façon de penser, idée, impression, jugement, pensée, point de vue, sentiment).[1]

- Est un ensemble des idées d'un groupe social sur les problèmes politiques, économiques, moraux, etc : L'opinion française. synonymes (conviction, croyance, doctrine, position, tendance, thèse, vue).[1]

On peut dire que L'opinion définie comme l'expression des sentiments d'une personne envers une entité. L'opinion est subjective et peut être décrite avec certains attributs. L'attribut d'opinion le plus étudié est la polarité (positive, négative, et éventuellement neutre) qui définit si l'opinion est favorable ou défavorable. D'autres attributs sont l'intensité de l'opinion et le degré de subjectivité [1].

1.5.3. L'objectif de fouille d'opinions

Fouille de l'opinion vise à exploiter les avis sur les divers produits (ex: appareils photo, électronique, voitures, livres, enregistrements musicaux, critiques de films, etc.) en les classant dans un avis positif ou négatif afin d'aider les entreprises dans le marketing. En outre, ces opinions peuvent être résumées afin de fournir aux utilisateurs des informations statistiques [8].

1.5.4. Processus de la fouille d'opinion

Le processus d'un système de fouille d'opinions comprend trois phases : acquisition et analyse du corpus, étude de la pertinence des documents par rapport à un sujet, détection de l'opinion et ré-ordonnancement des documents. La figure 1.4 nous montre les étapes de la fouille d'opinions [1].

[1] https://www.larousse.fr/dictionnaires/francais/opinion/56197#synonyme

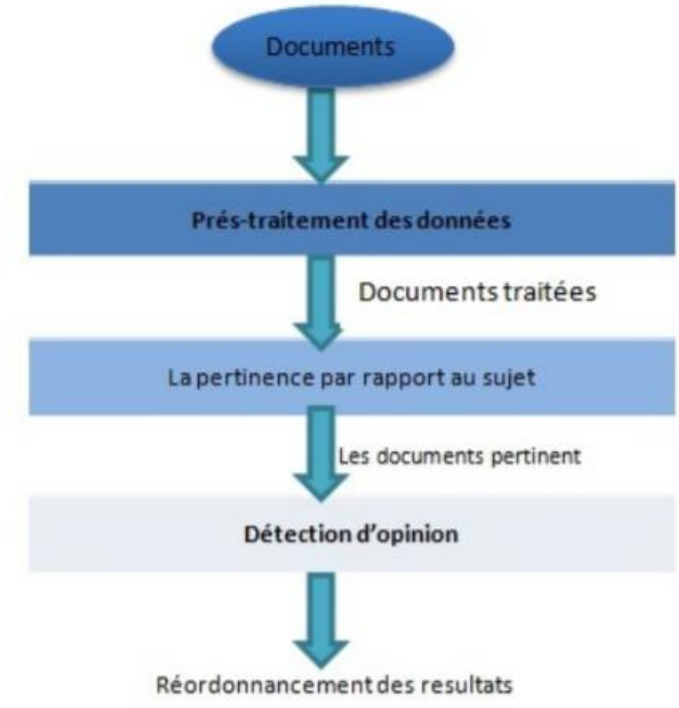


Figure1.4: La fouille d'opinion se compose de plusieurs tâches [1].

1.5.4.1. Acquisition et prétraitement de données

Dans cette étape, les textes sont traités pré-linguistiquement. Les mots vides et les mots qui ne fournissent aucune information sont supprimés, ainsi qu'une analyse lexicale pour supprimer les mots qui ont une signification commune (redondante). En ce qui concerne les blogs, la plupart des travaux se sont concentrés sur le nettoyage des balises HTML inutiles, ainsi que des documents non rédigés en anglais. Certaines techniques de fouille d'opinions utilisent la structure des phrases pour définir l'opinion. Dans cette étape, un étiquetage grammatical est fait (pour identifier l'adjectif, l'adverbe, le verbe, etc.) et des règles de dépendance sont utilisées pour construire la phrase de manière hiérarchique [1].

1.5.4.2. Pertinence par rapport au sujet

Cette étape consiste examiner l'importance des documents par rapport à un sujet particulier « topic » dans Trec « Text Retrieval Conference ». Nous utilisons ces deux termes de manière interchangeable. C'est l'une des méthodes les plus utilisées.

Il existe des modèles de recherche d'information textuelle qui dépendent de la mise en forme des mots clés. Habituellement, les 1000 premiers documents associés sont extraits et utilisés pour l'étape suivante [1].

1.5.4.3. Détection d'opinion

Plusieurs méthodes ont été utilisées pour la détection d'opinions. pour le but de déterminer la polarité des documents pertinents et les réordonner selon un score d'opinion [1].

1.6. Domaines d'application

L'importance de la détection d'opinion est présente dans plusieurs domaines mais la plus grande application de la fouille d'opinion reste dans le monde du business et du politique:

- **Marketing** : du côté entreprises, permet au fournisseur plus de connaissances à propos des besoins des consommateurs, du côté client il peut donner son opinion, s'inspirer des opinions d'autres clients pour l'aidé à sa décision et aussi comparer les produits avant de les acquérir
- **Politique**: Les jugements politiques dans les articles se retrouvent au cœur des débats avec l'avènement des médias sociaux, (et plus spécialement sur le web).
- **e-commerce:** depuis l'apparition du commerce en ligne la plupart des entreprises l'intègrent.(amazon.com,epinons.com).

1.6.1. Exemple d'application de la fouille d'opinions

Un exemple de site avec les mêmes capacités qu'un moteur de recherche orienté opinion, qui permet de recueillir les avis des internautes et de permettre à ses utilisateurs de poster leurs commentaires, car il permet de mener une recherche d'opinion et de lire les commentaires des internautes. Les autres utilisateurs doivent prendre une décision en fonction des conseils publiés. Nous citons le site populaire Epinions.com, qui intègre un moteur de recherche orienté opinion.

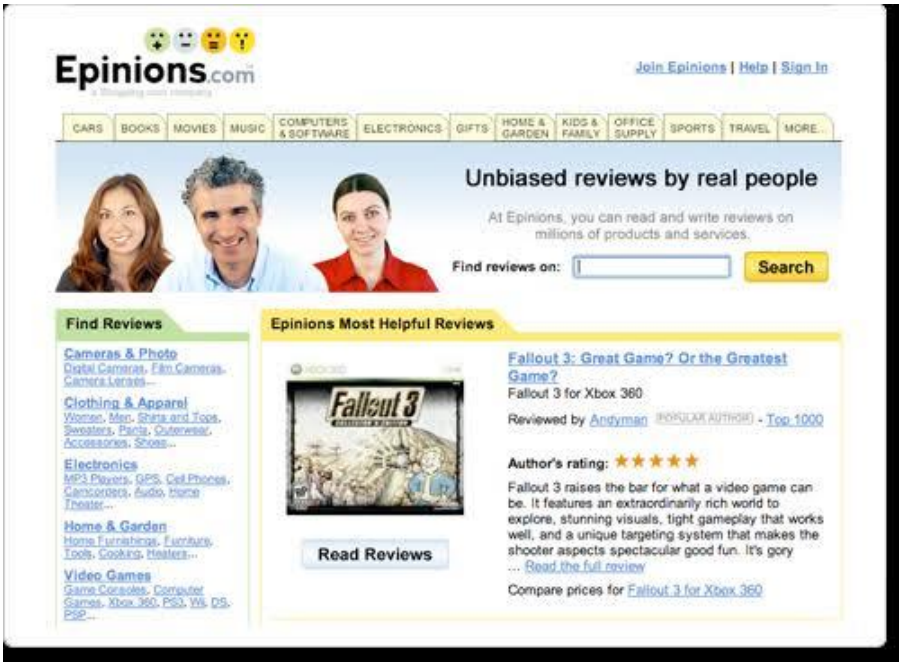


Figure 1.5 : Page d'accueil du site Epinions.com**.**

Ce site offre à ses utilisateurs de poster leurs commentaires sur des produit et services, comme il permet d'effectuer des recherches d'opinions et de lire les commentaires des autres utilisateurs pour prendre une décision fondée sur les recommandations et les conseils postés [10].

1.7. Conclusion

Nous avons étudié dans ce chapitre, les différentes définitions de la fouille de textes , et de la fouille d'opinion, nous avons aborder la distinction entre les textes d'opinion et les textes de faits, nous avons expliquer le processus de la fouille d'opinion et ces étapes . Puis nous avons cité quelques applications pour cette technologie.

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

Chapitre2
Méthodes et Approches

Chapitre2: Méthodes et approches

2.1. Introduction

Avec le développement du Web, et surtout du Web 2.0, le nombre de documents décrivant des opinions sur un produit ou un film devient de plus en plus important. Récemment, les chercheurs de différentes communautés (Fouille de données, Fouille de textes, Linguistique) se sont intéressés à l'extraction automatique de ces données d'opinions sur le Web.

Certaines techniques de détection d'opinions cherchent à déterminer les caractéristiques d'opinions positives ou négatives à partir d'ensembles d'apprentissages.

Dans ce chapitre, nous nous intéresserons en premier axe à la tâche d'identification et de classification d'opinions, et en deuxième axe nous discutons les trois grandes catégories de méthodes peuvent être mises en avant : les approches basées sur l'apprentissage automatique, les approches basées sur la linguistique et les approches hybrides.

2.2. Classification d'opinions

La classification des opinions est une tâche spéciale pour l'exploration de texte et aborde le problème du traitement automatique de texte, qui a pour but d'attribuer une étiquette au texte selon l'opinion qu'il exprime. On considère généralement les classes positive et négative, ou encore neutre. Nous nous intéresserons ici uniquement à la classification d'opinion. Nous cherchons à déterminer les goûts d'utilisateurs à partir de l'analyse de leurs commentaires. Qui concerne le traitement des opinions ou des sentiments et de la subjectivité dans les textes [1].

2.2.1. La classification des textes d'opinion

L'extraction d'opinions consiste à identifier des textes porteurs d'opinions dans un groupe, et c'est une recherche dans les ressources lexicales pour l'identification de la subjectivité des lexiques, la figure suivante montre le processus de détection de l'opinion passant par la détection de la subjectivité, une fois que les mots porteurs d'opinion sont répertoriés, il faut déterminer la polarité [1].

2.2.2. Type de classification

Il existe deux types de classifications :

Binaire : définit deux classes (positive et négative).

Multi-classes : définit trois classes (positive, négative, neutre) ou plus de trois classes (fortement positive, positive, neutre, négative, fortement négative).

2.2.3. Classification de la polarité des opinions

Le résumé d'opinion consiste à fournir un accès rapide et facile aux informations en mettant en évidence les opinions exprimées et les objectifs de ces opinions présentées dans le texte. Ce résumé peut être textuel, chiffré, graphique ou encore imagé. Après la détection d'opinions, il est important de classifier le sentiment et définir sa polarité, puis compter le nombre de mots positifs et le nombre de mots négatifs présents pour attribuer l'opinion a une classe [1].

2.3. Approches de l'analyse des sentiments et la détection des opinions

2.3.1. Approche basées sur l'apprentissage automatique (Machine Learning)

2.3.1.1. Apprentissage supervisé

Il est basé sur les données libellées et donc, les étiquettes sont fournies au modèle au cours du processus d'apprentissage. Ces données libellées sont utilisées par l'algorithme d'apprentissage pour donner un modèle qui sera utilisé lors de la prise de décision. Certains modèles d'apprentissage automatique ont été formulés pour classer le texte, c'est-à-dire que l'on soumet au classificateur des exemples pour s'entrainer à classer correctement les documents futurs.

Il existe de nombreuses méthodes d'apprentissage supervisé :

- K plus proches voisins;
- Arbres de décisions ;
- Naïve Bayes (ou encore Simple Bayes) ;
- Réseaux de neurones ;
- Machines à support de vecteurs (ou SVM) ;
- Programmation génétique [9].

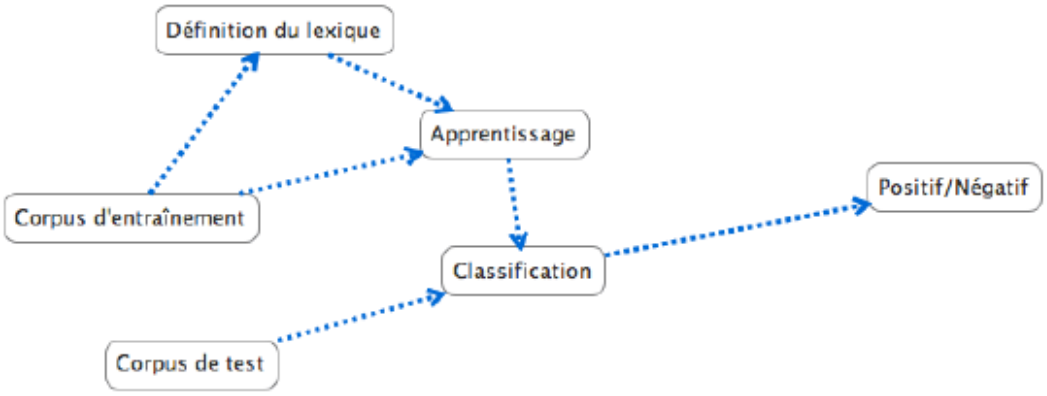


Figure 2.1: Fouille d'opinion avec une méthode d'apprentissage supervisée [4].

a) Le corpus d'apprentissage

L'ensemble d'apprentissage a un effet direct sur l'apprentissage du modèle et donc sur les résultats de la classification, les exemples doivent être aussi représentatifs que possible de toutes les données. Dans les avis de notation, ou plus généralement dans la classification des textes, les groupes étudiés sont très petits car les noms d'exemples sont souvent faits manuellement. Cela coûte cher et ne permet donc pas beaucoup d'apprentissage. Cependant, les données des sites Web 2.0 permettent désormais de traiter des lots beaucoup plus importants [10].

b) Segmentations (Tokenization)

Deux solutions de segmentation ont été retenues.

- **Les mots :** Un document est un ensemble de caractères. La segmentation consiste à découper cet espace de caractères afin d'obtenir un espace de variables. On peut considérer les marques de ponctuation, afin de découper le document en phrases, ou encore les espaces associés à la ponctuation afin d'obtenir des mots.
- **Les n-grammes de lettres :** La deuxième segmentation est le découpage en n-grammes de lettres : On peut par exemple limiter la taille des variables à un nombre n de caractères, on parle alors de n-grammes de lettres. Les n-grammes peuvent également être formés de mots. Ces n-grammes de mots peuvent être construits selon leur ordre dans la phrase, afin de conserver un sens sémantique. Par exemple, dans la phrase « je n'aime pas ce film », le bi-gramme (ou 2-grammes) « aime pas » sera considéré. On peut également construire les n-grammes de mots selon qu'ils

apparaissent dans la même phrase par exemple. Une fois le vocabulaire sélectionné, différents choix sont possibles concernant la représentation du document sur le vecteur [10].

c) La représentation :

Le texte original peut être vu comme une séquence de mots. Ce type de représentation est actuellement incompréhensible pour les algorithmes d'apprentissage automatique qui ont besoin de recevoir des représentations vectorielles numériques des entités à classer. La représentation vectorielle consiste à transformer chaque document en une séquence de nombres, dans laquelle chaque nombre correspond à un mot du vocabulaire de l'ensemble des documents ou corpus.

- **La représentation binaire**

Cette représentation est la plus simple et la plus ancienne, et la moins coûteuse en temps de calcul. elle ne s'intéresse que sur la présence ou la non-présence d'un terme dans le texte, il consiste à utiliser une pondération binaire : 1 si le terme est présent une ou plusieurs fois dans le document, 0 dans le cas contraire [10].

- **La représentation fréquentielle normalisée**

Cette représentation consiste à présenter le texte sous forme de vecteur

Cette représentation consiste à normaliser les vecteurs de représentation des textes par la longueur des textes. C'est-à-dire que les fréquences des variables obtenues avec la représentation fréquentielle sont remplacées par la proportion des variables dans chaque document. La proportion s'obtient en divisant la fréquence de la variable par la taille du document [9].

- **La représentation TF-IDF (Term Frequency - Inverse Document Frequency)**

C'est une représentation vectorielle plus informative que les deux représentations précédentes. La mesure statistique TF-IDF permet l'évaluation d'une variable dans un document à la fois par sa fréquence dans le document concerné, mais également par sa présence dans tous les autres documents du corpus.

- **TF(T, D)** : correspond au nombre d'occurrences de ce terme dans le document considéré. Ainsi, pour le document dj et le terme ti, la fréquence du terme dans le document est donnée par l'équation suivante: [10].

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$\boldsymbol{ni,j}$: est le nombre d'occurrences du terme ti dans di.

$\Sigma\boldsymbol{nk,j}$: est le nombre de termes dans le document.

- **IDF(T)** mesure l'importance du terme dans l'ensemble du corpus. Elle consiste à calculer le logarithme de l'inverse de la proportion de documents du corpus qui contiennent le terme. Elle est définie de la manière suivante : [10].

  D représente le nombre total de documents dans le corpus.

  dj :$ti \epsilon dj$:est le nombre de documents dans lesquels le terme ti apparait.

  $$IDF_i = \log(\frac{D}{d_j : t_i \epsilon d_j})$$

d) Classifieurs :

Ils existent plusieurs méthodes de classement supervisé on peut citer :

➢ **Classification par KNN:**

La méthode des k plus proches voisins ou The k-NN classification (k-Nearest Neighbors) est un classifieur a base d'instances qui fait partie des méthodes géométriques utilisant des mesures de distance, L'idée de K-plus proches voisins est de représenter chaque texte dans un espace vectoriel, dont chacun des axes représente un élément textuel.

Voici son algorithme général :

Paramètres : le nombre k de voisins

Données : un ensemble d'exemples classés (document, classe)

Entrée : un nouveau document D

1. déterminer les k plus proches documents de D

2. Sélectionner la classe majoritaire C des classes de ces k exemples

Sortie : la classe de D est C [11].

➢ **Classification par SVM :**

Les machines à vecteurs de support ou séparateurs à vaste marge (Support Vector Machine, SVM). Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais Support Vector Machine, SVM) sont un ensemble de techniques

d'apprentissage supervisé pour résoudre des problèmes de discrimination et de régression. Les SVM sont une généralisation des classifieurs linéaires effectue le classement en recherchant l'hyper-plan qui différencie les classes que nous avons tracées dans un espace à n dimensions [12].

- **Classification Bayésienne :**

Basée sur l'utilisation des réseaux bayésiens. Naïve Bayes classifier. C'est l'une des méthodes les plus pratiques d'apprentissage, En revanche si les modèles vectoriels fonctionnent bien, leur fondement est entièrement empirique .Le classifieur bayésien naïf reste un des outils de catégorisation de documents les plus pratiques en raison de ses performances reconnues dans ce domaine, et est aujourd'hui intégré à de nombreux produits commerciaux. Est utilisée lorsque l'information que nous disposons est entachée de probabilités, c'est-à-dire incertaine.

2.3.1.2. Apprentissage non supervisé

Cette méthode de classification est utilisée lorsque que l'on possède des documents qui ne sont pas classés et dont on ne connaît pas de classification et c'est un thème de recherche majeur en apprentissage automatique et en fouille de données où l'objectif est la répartition en classes d'un ensemble de données non étiquetées.

Et voici quelques algorithmes de classification non supervisée :

- CURE (Clustering Using REpresentatives).
- BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies).
- ROCK (RObust Clustering using linKs).
- TSVQ (Tree Structured Vector Quantization) [9].

2.3.2. Approche basée lexique

Cette méthode utilise un dictionnaire des sentiments avec des mots d'opinion et les faire correspondre avec les données pour déterminer la polarité. Elle attribue les scores de sentiment aux mots d'opinion décrivant si les mots sont positifs, négatif ou neutre ,les approches fondées sur le lexique reposent principalement sur un lexique de sentiment, à savoir, une collection de termes de sentiment connue et précompilée, des phrases et même des expressions idiomatiques, développés pour les genres traditionnels de communication[12].

Pour construire un dictionnaire, trois genres de techniques sont possible:

- la méthode manuelle ;
- la méthode basée sur les corpus ;
- la méthode basée sur les dictionnaires.

2.3.2.1. Méthode manuelle

Cette méthode demande un effort important en terme de temps, et consiste à enrichir le lexique de mots d'opinions sans aucun outil particulier, seulement les experts font la sélection de mots et expressions porteurs d'opinions. Cet ensemble de mots est appelé graine, construire une première liste de mots et d'expressions utilisée par la suite à trouver, répertorier et classer d'autres mots et expressions porteurs d'opinions [13].

2.3.2.2. Méthode basée dictionnaire

Cette approche consiste à utiliser des dictionnaires de synonymes et antonymes existants tels que WordNet, Où WordNet est une base de données lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton, Dans cette base les mots sont organisés sous forme d'arbres comme l'indique la figure2.1 [13].

Figure 2.2: Exemple d'arbre de synonymes et d'antonymes présents dans WordNet [5].

2.3.2.3 Méthode basée corpus

La méthode basée corpus consiste à établir un ensemble de dictionnaire d'opinion du façons :on peut généré automatiquement à partir du corpus ou les mots qui contiennent une opinion sont extraits directement des informations présentes dans le corpus fondées sur les cooccurrences de mots dans le corpus. Soit utiliser une de conjonctions de coordination suivantes : AND, OR, BUT, EITHER-OR, et NEITHER-NOR. [13].

2.3.3. Approche hybride

La méthode hybride combine les points forts des deux approches précédentes, pour aboutir à des résultats très précis. Elles prennent en compte tout le traitement linguistique des approches basées lexique avant de lancer le processus d'apprentissage comme dans les approches statistique.

La combinaison des approches basée lexique et les approche basées sur l'apprentissage automatique supervisé a donné des résultats plus précis que chacune des approches employées séparément [13].

2.4. les critères d'évaluations utilisées

Ce sont les critères pour effectuer la comparaison entre les Classifieur **:**

- **Compréhensibilité :** montre si le modèle est compréhensible et si le système donne des réponses qui permettent de comprendre pourquoi le document est classé dans une certaine classe ou est-ce une fonction numérique calculée à partir de données qui sert d'exemples (boîte noire).
- **Simplicité :** apprécie le taux de simplicité des résultats d'apprentissage produits par le classifieur.
- **Intelligibilité** : évalue le degré d'intelligence du classifieur.
- **Le temps de réponse et d'indexation** : est aussi un point qui peut être fondamental.
- **L'encombrement du système et les ressources en mémoire requises** : l'espace alloué en mémoire vive et sur le disque dur qui doit être prise en compte dans de nombreux cas [11].

Pour mesurer l'efficacité d'un classificateur dans un problème à n classes (en l'occurrence deux : positif et négatif), trois mesures sont utilisées : la précision, le rappel et le F-score**.**

2.4.1. F-score

La mesure la plus utilisée en classification d'opinions c'est Le F-score. Il se mesure à l'aide de la formule suivante :

F-score = 2 × (Précision × Rappel) / (Précision + Rappel)

2.4.2. Précision

Précision. Proportion d'éléments bien classés pour une classe donnée:

$$Précision_i = \frac{Objets\ correctements\ attribués\ à\ la\ classe\ i}{Nombre\ d'objets\ attribués\ à\ la\ classe\ i}$$

2.4.3. Rappel

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

$$Rappel_i = \frac{Documents\ correctements\ attribués\ à\ la\ classe\ i}{Nombre\ de\ documents\ attribués\ à\ la\ classe\ i}$$

2.5. Langue arabe

La langue arabe est la langue habitants arabes et qui ont occupé les régions du nord de la péninsule arabique. La langue arabe est considérée comme la 5éme langue courante utilisée dans le monde. Avec un nombre de locuteurs estimé entre 315 421 3001 et 375 millions de personnes au sein du monde arabe et de la diaspora arabe [16].

Avec ses propriétés morphologiques et syntaxiques la langue arabe est considérée comme une langue difficile à maitriser dans le domaine du traitement automatique da la langue, [17]. L'arabe doit sa formidable expansion à partir du 7eme siècle grâce à la propagation de l'islam et la diffusion du coran [18]. Les recherches pour le traitement automatique de l'arabe ont débuté vers les années 1970. Les premiers travaux concernaient notamment le lexique et la morphologie.

Avec la diffusion de la langue arabe sur le web et la disponibilité des moyens de manipulation de textes arabes, les travaux de recherche ont abordé des problématiques plus variées comme la syntaxe, la traduction automatique, l'indexation automatique des documents, la recherche d'information, etc.

2.5.1Complexité de la langue arabe

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

- certaines combinaisons de caractères peuvent être écrites de différentes manières, L'orthographe avec diacritiques est plus phonétique et moins ambiguë en arabe [19].

- Les voyelles courtes sont omises des textes arabes écrits, qui à leur tour donnent une prononciation différente, C'est grammaticalement nécessaire [20].
- La langue arabe a une morphologie très complexe par rapport à la langue anglaise.
- Les synonymes des mots de la langue arabe sont très répandus [21].
- De nombreux mots-clés ou fonctionnalités peuvent être trouvés dans le texte arabe, tels que des formes qui peuvent être créées à partir de la racine, ce qui peut entraîner des erreurs de performances en termes de précision et de temps car la classification automatique du texte dépend du contenu du document [21].

2.5.2. La richesse de la langue arabe

La langue arabe s'écrit au moyen de 28 lettres, Elle est une langue très riche, il y aurait 80 termes différents pour identifier le miel, 200 pour le serpent, 500 pour le lion, 1000 pour le chameau et l'épée et jusqu'à 4400 pour définir l'idée de Malheur, elle dispose le plus grand nombre de mots avec plus de 12 millions de mots où le Français 150 000, l'Anglais 600 000 mots, mots, le Russe 130 000 mots. Les grammairiens arabes prétendent que toutes les racines ont été originalement des verbes, où le nombre de ces racines en réalité est de 6000[22].

2.6. Travaux connexes

Dans ce contexte il y a deux genres de travaux :

- le premier sur la langue arabe standard :

Selon les travaux de [26] , ils ont utilisé un corpus contient 500 critiques de films en arabe, collectées à partir de forums et des sites Web. Il est divisé en 250 avis positifs et 250 avis négatifs, ils ont appliqué les deux classificateurs les plus utilisés: Support Vector Machines (SVM) et Naïve Bayes (NB). Les résultats obtenus sont montrés dans les tables suivant :

	Stem	P	R	F1
OCA	Yes	0.8614	0.8800	0.8706
	No	0.8699	0.9480	**0.9073**
EVOCA	Yes	0.9007	0.8680	0.8840
	No	0.8561	0.8840	0.8698

Tableau 2.1 : Résultats avec SVM.

	Stem	P	R	F1
OCA	Yes	0.8106	0.8880	0.8475
	No	0.8274	0.9520	**0.8853**
EVOCA	Yes	0.7100	0.8320	0.7662
	No	0.7323	0.8640	0.7927

Tableau 2.3 : Résultats avec NB.

- le deuxième sur les livres arabe :

Le travail choisis est de [27], ils ont collecté le plus grand ensemble de données d'analyse à ce jour pour la langue arabe. ils ont commencé par la première phase (collection de données), leur corpus était collecté par eux-mêmes à partir du site de critique et d'évaluation de livres, cette collection de données se compose de plus de 63000 livres avis, chacun noté sur une échelle de 1 à 5 étoiles. En suite le prétraitement passant
à la classification, Le corpus contient 13160 commentaires divisé en deux parties le premier parties pour l'apprentissage et le deuxième pour le test, Les résultats obtenus ont montrés que l'algorithme SVM donne les meilleurs résultats dans le cadre déséquilibrée, tandis que MNB est mieux que SVM dans le cadre équilibré.et les meilleurs scores sont obtenus avec tf-idf.

2.7. Conclusion

Dans ce chapitre, nous avons présenté brièvement la classification des opinions qui est l'un des sujets le plus largement étudié. Nous avons montré la différence entre la classification des textes et la classification de la polarité des opinions, et énuméré toutes les approches existantes: les approches basées sur la linguistique, les approches basées sur l'apprentissage automatique et les approches hybrides qui utilisent les deux précédentes.

Nous nous basons sur l'apprentissage supervisé qui est le plus utilisé, et aux techniques génériques pour l'analyse des opinions à l'aide des différents Classifieurs, et les travaux qui s'intéressaient à l'analyse des sentiments Aussi on a focalisé sur la langue arabe et sa complexité.

Dans le chapitre suivant, nous allons présenter notre modélisation, où nous avons collectées un ensemble de commentaires à partir des réseaux sociaux qui ont été rédigés en langue arabe.

Chapitre 3
Systéme proposée

Chapitre3: Système proposée

3.1. Introduction

Nous présentons, dans ce chapitre, la source de données sur laquelle le modèle est appliqué, et les algorithmes d'apprentissage automatique utilisés (KNN, SVM et NB).la figure 3.1 illustre le système propose pour la classification des commentaires arabe.

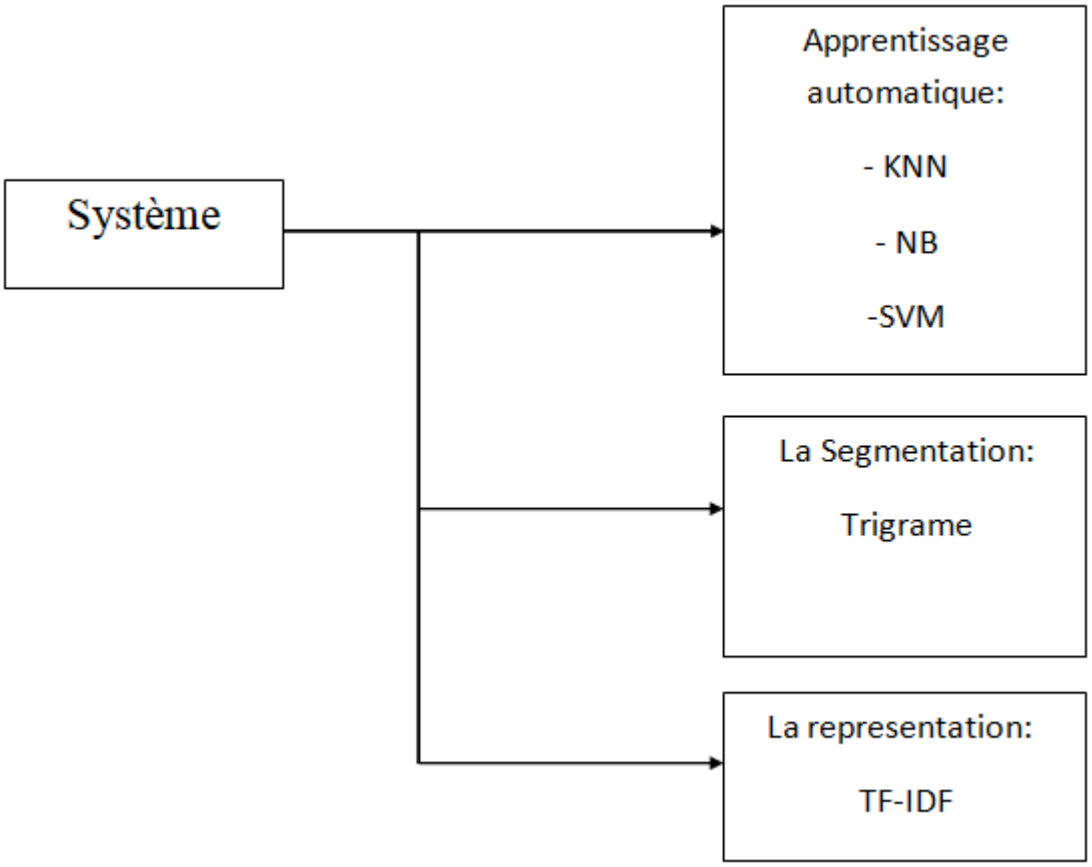


Figure 3.1 : système propose pour la classification des commentaires arabe.

Dans notre travaille en a choisie le découpage en trigrammes de lettres qui comme nous l'avons vu, est une segmentation souvent employée en fouille de textes, cette segmentation consiste en l'ensemble des suites de n caractères présentes dans le texte tell que n égal à 3 (trigrammes). Chaque lettre est un caractère, chaque espace entre deux lettres est un caractère, les ponctuations sont ignorées. Les espaces font donc partie des suites de n caractères représentatives, et nous avons choisie la représentation Inverse Document Frequency TF-IDF la plus populaires dans la littérature.

3.2. Contribution

Nos contributions principales consistent à :

1) Travailler sur l'arabe standard, avec l'exploitation de quatre classificateurs, les k-voisins les plus proches (KNN), Naïve Bayes et les algorithmes de machine à vecteurs de support (SVM).
2) Travailler sur une collection de données qui contient 13160 commentaires.

3.3. Source de données

Dans cette tâche, nous avons utilisé un ensemble de données téléchargé à partir de www.goodreads.com. Il s'agit d'un ensemble de critiques de livres en arabe contenant 13160 critiques divisé en deux parties le premier parties contient 9872 commentaires pour l'apprentissage et le deuxième parties contient 3288 commentaires pour les test . Il s'agit du plus grand ensemble de données d'analyse des sentiments en arabe à ce jour, et en utilisant le langage de programmation Python.

3.4. base de donnees textuelle.

Nous avons utilise toutes les 13160 entrées par l'utilisation de deux polarités positive, négative (qui ont les valeurs 1, 0).

Le tableau 3.1 indique les classes de la base selon les deux valeurs de polarités.

Polarité	Positive	Négative	Total
Nombre de mots	6556	6604	13160

Tableau 3.1 : Nombre de commentaires par polarité.

Dans le tableau 3.2, nous avons décrit quelques exemples avec leurs polarités.

Commentaire	Polarité
رائعة ولن انساها رغم انني قرأتها منذ فترة طويلة	Positive
جميلة و مسلية	Positive
ندمت أنني ابتعت ذلك الكتاب	Négative
للأسف لم يعجبني . لم ترق لي خواطرها أبدا , نجمه واحده تكفي	Négative

Tableau 3.2: Exemple de notation de quelques commentaires.

3.5. Conclusion

Dans ce chapitre, nous avons donné un aperçu détaillé sur notre système et les algorithmes d'apprentissage automatique que nous avons utilisé dans notre travail. Et nous avons présenté la collection de données et des exemples sur cette collection

Chapitre 4
Implémentation

Chapitre 4:Implémentation

4.1. Introduction

Dans ce chapitre, nous allons implémenter les algorithmes d'analyse de sentiments .Ces algorithmes sont : les k-voisins les plus proches (KNN), Naïve Bayes et les algorithmes de machine à vecteurs de support (SVM). Nous présentons les outils exploités pour le développement de l'application, l'environnement de programmation, ainsi que l'ensemble des résultats des expérimentations par toutes les approches proposées .L'objectif est de faire une étude comparative entre les algorithmes testes concernant leur exactitude.

4.2. Ressources utilisées

- ✓ Un PC Intel Core i5 à 3GHZ et 4Go de RAM.
- ✓ Pycharm IDE

4.3.Création des programmes avec python

4.3.1. Définition

Python est le langage de programmation le plus utilisé dans le domaine du Machine Learning, du Big Data et de la Data Science. En tant que langage de programmation de haut niveau, Python permet aux programmeurs de se focaliser sur ce qu'ils font plutôt que sur la façon dont ils le font. Ainsi, écrire des programmes prend moins de temps que dans un autre langage. Il s'agit d'un langage idéal pour les débutants [25].

Python érigé comme le meilleur langage de programmation pour le Big Data, c'est grâce à ses différents packages et bibliothèques de science des données. Voici les plus populaires : Pandas, Agate, Bokeh, Scikit-learn, Scipy, NumPy, PyBrain.

4.3.2. Caractéristiques du langage python

- «open-source» : son utilisation est gratuite et les fichiers sources sont disponibles et modifiables ;
- simple et très lisible ;
- doté d'une bibliothèque de base très fournie ;
- importante quantité de bibliothèques disponibles : pour le calcul scientifique, les statistiques, les bases de données, la visualisation . . . ;

- grande portabilité : indépendant vis à vis du système d'exploitation (linux, windows, MacOS) ;
- orienté objet ;
- typage dynamique : le typage (association à une variable de son type et allocation zone mémoire en conséquence) est fait automatiquement lors de l'exécution du programme, ce qui permet une grande flexibilité et rapidité de programmation, mais qui se paye par une surconsommation de mémoire et une perte de performance ;
- présente un support pour l'intégration d'autres langages.

4.4. Environnement de développement Pycharm

Pycharm est un IDE, Integrated Développent Environment (EDI environnement de développement intégré en français), spécialisé pour les langages de programmation Python et Django. Il offre de riches et nombreuses fonctionnalités en matière d'édition, de débogage, de développement et de tests.

4.5. Exemples de codes sources

Nous allons présenter quelques exemples de codes sources.

Figure 4.1 : présente un morceau de code qui permet d'appeler les bibliothèques nécessaires pour compiler notre application

```
import pandas as pd
from joblib import dump, load
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import cross_val_score
from sklearn.naive_bayes import MultinomialNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
```

Figure 4.1 : Appel des Bibliothèques.

La Figure 4.2 : présente les instructions nécessaires pour lire la collection de données.

```
def pre_processing():
    train = pd.read_csv(os.path.join(DATA_DIR, 'balanced_2classes_train.csv'), encoding='utf-8')
    val = pd.read_csv(os.path.join(DATA_DIR, 'balanced_2classes_val.csv'), encoding='utf-8')
    test = pd.read_csv(os.path.join(DATA_DIR, 'balanced_2classes_test.csv'), encoding='utf-8')
    stops = pd.read_excel(os.path.join(DATA_DIR, 'ar_stops.xlsx'), encoding='utf-8')
    frames = [train, val, test]
```

Figure 4.2 : Lire la collection de données.

La préparation des données **:**

```
data = pd.concat(frames)

data['text'] = data['text'].apply(lambda x: re.sub(r"[0-9]", " ", x))
stop_words = list(stops['"'])
review = data['text']
review = review.apply(lambda x: " ".join(x))
train = data[:train.shape[0]]
val = data[train.shape[0]: (val.shape[0] + train.shape[0])]
test = data[(val.shape[0] + train.shape[0]):]

X_train = train['text']
y_train = train['label']

if not os.path.isfile(os.path.join(JOBLIB_DIR, 'count_vector.joblib')):
    train_df_vectorized = TfidfVectorizer(min_df=2, ngram_range=(1, 3))
    X = train_df_vectorized.fit_transform(X_train)
    dump(train_df_vectorized.fit(X_train), os.path.join(JOBLIB_DIR, 'count_vector.joblib'))
else:
    train_df_vectorized = load(os.path.join(JOBLIB_DIR, 'count_vector.joblib'))
    X = train_df_vectorized.fit_transform(X_train)

return [X, y_train]
```

Figure 4.3 : La préparation des données « pre_processing ».

```
def train_svm():
    data = pre_processing()

    svm = SVC(kernel='linear')
    svm.fit(data[0], data[1])

    cv2 = cross_val_score(svm, data[0], data[1], cv=10)

    dump(svm, os.path.join(JOBLIB_DIR, 'SVMmodel.joblib'))


def predict_svm(text_to_predict):
    if not os.path.isfile(os.path.join(JOBLIB_DIR, 'SVMmodel.joblib')) or not os.path.isfile(
            os.path.join(JOBLIB_DIR, 'count_vector.joblib')):
        train_svm()

    loaded_model = load(os.path.join(JOBLIB_DIR, 'SVMmodel.joblib'))
    loaded_count_vector = load(os.path.join(JOBLIB_DIR, 'count_vector.joblib'))

    return predict(loaded_model, loaded_count_vector, text_to_predict)
```

Figure 4.4:Appel du classificateur SVM.

```
def train_knn():
    data = pre_processing()

    model = KNeighborsClassifier(n_neighbors=5)
    model.fit(data[0], data[1])
    dump(model, os.path.join(JOBLIB_DIR, 'KNNmodel.joblib'))

    cv2 = cross_val_score(model, data[0], data[1], cv=10)
```

Figure 4.5:Appel du classificateur KNN.

```
def train_naive_bayes():
    data = pre_processing()

    clfrNB = MultinomialNB(alpha=0.1)
    clfrNB.fit(data[0], data[1])

    cv = cross_val_score(clfrNB, data[0], data[1], cv=10)

    dump(clfrNB, os.path.join(JOBLIB_DIR, 'NBmodel.joblib'))
```

Figure 4.6:Appel du classificateur Naïve Bayes.

4.6. Expérimentations et resultats

Nous avons testé notre système sur un corpus de textes et cette méthode nous a permis de classer presque tous les commentaires présents dans le corpus de test.

Le tableau suivant contient le résultat de l'analyse de corpus. Nous avons fait le test sur notre collection de donnée, les résultats d'exactitude sont présentés dans le tableau 4.1

Model	SVM	NB	KNN
Test sur notre corpus	Accuracy		
	81.12%	80.67%	50.89%

Tableau 4.1 : Résultats de classification.

Selon les résultats obtenus, il est clair que la performance de l'approche SVM est meilleure que les autres approches par une exactitude de 81.12%, puis l'approche NB par une exactitude de 80.67% et enfin l'approche KNN.

Nous avons testé ces modèles sur touts la collection de donnée, nous pensons que nous pouvons améliorer le processus d'analyse des sentiments par l'implication des autres des autres fonctionnalités et des aspects.

4.7. Interfaces Graphiques

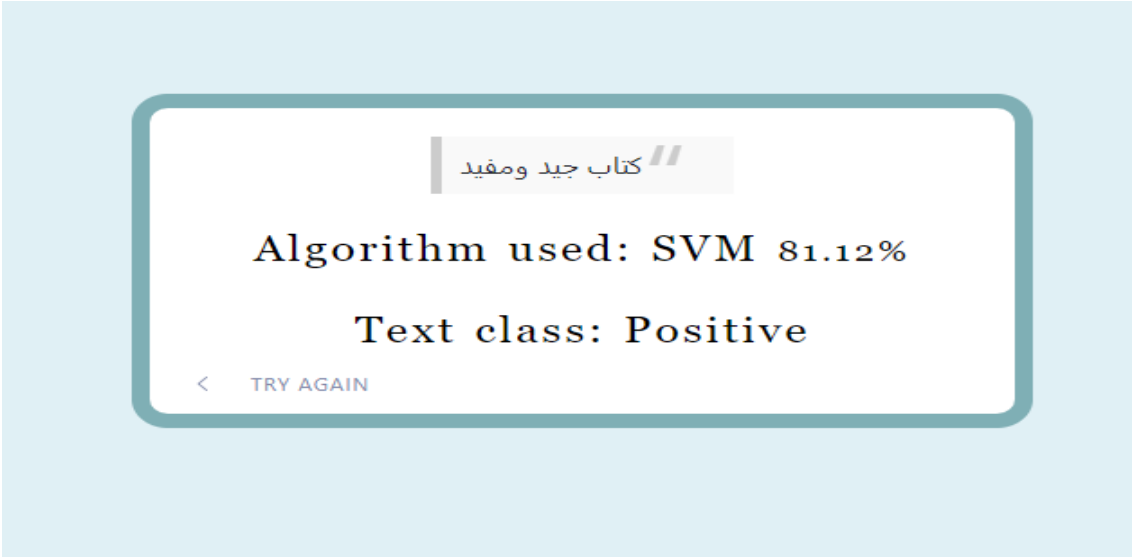
Dans cette partie, nous allons présenter quelques interfaces de l'application, et nous allons implémenter les algorithmes d'analyse de sentiments qui sont quatre algorithmes.

4.7.1. L interface principale de l'application

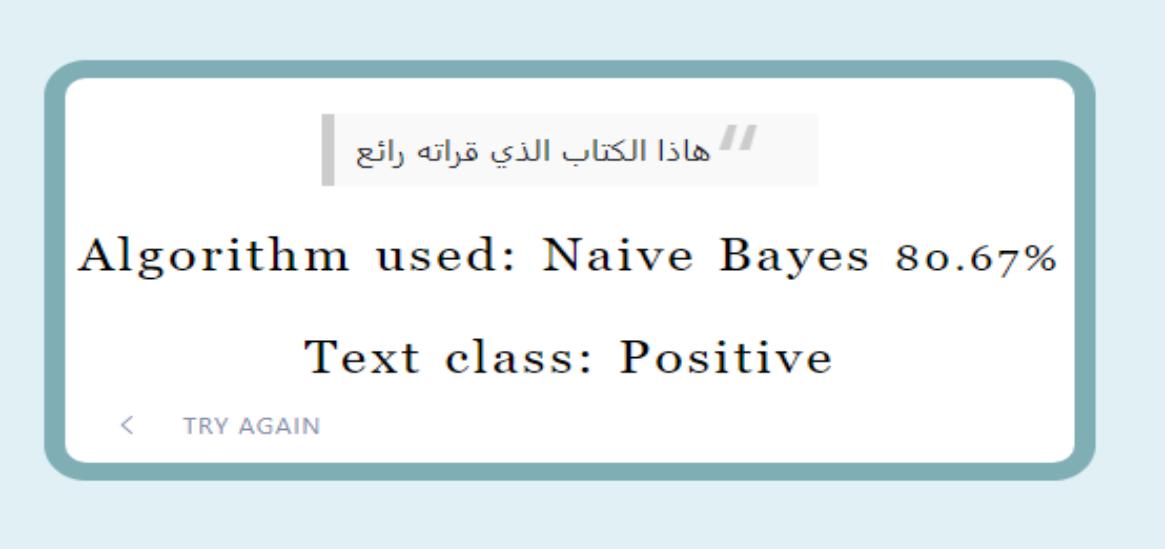
Cette interface vous permet d'écrire un commentaire et de choisir le type de classification.

4.7.2. Appel du classificateur KNN

4.7.3. Appel du classificateur SVM



4.7.4. Appel du classificateur NB



4.8. Conclusion

Au cours de ce chapitre, nous avons présenté les principaux outils utilisés pour réaliser l'application, Nous avons exploité trois classificateurs d'apprentissage automatique qui sont les k-voisins les plus proches (KNN), Naïve Bayes et les algorithmes de machine à vecteurs de support (SVM), ainsi quelques interfaces principales de notre Système.

Conclusion générale et perspectives

Conclusion générale et perspectives

Ce mémoire de fin d'études aborde la détection des polarités des publications dans les réseaux sociaux selon deux voies, une publication positive, une publication négative. Pour mener à bien cette étude, nous avons utilise une collection de données de plus de 13160 commentaires. Chaque commentaire se voit attribuer une étiquette Positive, Négative.

Dans ce rapport, Nous avons commencé par la définition de quelques concepts utilisés dans ce mémoire. Nous avons présenté, également, les différentes techniques que nous avons utilisées dans nos expérimentations. Ensuite, nous avons étudié les particularités de la langue arabe et enfin nous avons fourni une description détaillée de l'application qui applique plusieurs des techniques d'analyse des opinions et des commentaires que nous avons étudiées.

Afin de développer un système automatisé d'analyse d'opinions, nous avons mis en œuvre l'approche Machine Learning, dans laquelle le système reçoit les commentaires et les classe. Les trois algorithmes que nous avons appliqués sont SVM, KNN et NB.

Les résultats que nous avons obtenus sont très performants et très compétitifs et très encourageants car nous avons obtenu une meilleure exactitude « 81,12% »lors de l'utilisation du classificateur SVM.

Perspectives

Pour les perspectives de ce travail. Nous pouvons citer pour l'amélioration de cette étude les points suivants :

- Augmentation de la taille des bases de données des sentiments par plus des mots de la langue arabe standard et du dialecte algérien.
- développer un système d'analyse en utilisant l'analyse d'opinions multilingues, vu que les sites web algériens contiennent multiples langues : dialecte algérien, français, arabe et l'anglais.

- Ajouter plus de classificateurs, plus de paramètres et de fonctionnalités.
- Ajouter l'analyse par l'utilisation de la classe Mixte.

Bibliographie

Bibliographie

[1]CHABBOU Fatma Zohra, BAKHOUCHE Souhaila, 2016, thèse de master Fouille d'opinions méthodes et outils Étude des méthodes existantes de classification de textes d'opinion, Université de Larbi Tébessi –Tébessa.

[2]BOULLIER Dominique, LOHARD Audrey, 2012, Chapitre 5. Détecter les tonalités : opinion mining et sentiment analysis.

[3]DERMOUCHE Mohamed, LOUDCHER Sabine ,VELCIN Julien, FOURBOUL Eric , Analyse et visualisation d'opinions dans un cadre de veillesur le Web ,Université de Lyon (ERIC LYON 2)France.

[4]GILLOT Sébastien, juin 2010, mémoire de master, Fouille d'opinions.

[5]Gherabi Sara, 2014, thèse de master classification automatique des textes arabe (arabic opinion polarity), Université de M'sila.

[6]I. Tellier, Introduction à la fouille de textes, Université de Paris 3, Sorbonne.

[7]THONET Thibaut, doctorat Modèles thématiques pour la découverte non supervisée de points de vue sur le Web, de l'université de Toulouse.

[8]Zhongwu Zhai,Bing Lui,Hua Xu and Hua Xu, February 9-12,2011, Clustering Product Features for Opinion Mining, WSDM'11, Hong Kong, China. Copyright 2011

[9]FAREK Lazhar, 2009 , MEMOIRE Présentation en vue de l'obtention du diplôme de magister Identification d'opinions dans les journaux arabes, Faculté de Sciences de l'ingénieur Annaba.

[10]FAREK Lazhar, 2014, Thése de doctorat, Identification d'opinions dans les textes arabes en utilisant les ontologies.

[11]MATALLAH Hocine, Février 2011, Magister en informatique, classification Automatique de Textes Approche Orientée Agent.

[12]CHRITA Hanae, GAROUANI Moncef, 18.06.2019, Master Sciences et Techniques Systèmes Intelligents & Réseaux, L.S.I.A de la Faculté des Sciences et Techniques, de Fès.

[13]FAIZ Abelbachir, juin2010, Expérimentation de fonctions pour la détection d'opinions dans les blogs, Recherche de l'Université Paul Sabatier Toulouse.

[14] Dave, K., Lawrence, S. et Pennock, D. M. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In Proceedings of the 12th International Conference on World Wide Web, WWW '03, pages 519–528.

[15]Nasukawa, T. N. et Yi, J. (2003). Sentiment Analysis: Capturing Favorability Using Natural Language Processing. In Proceedings of the 2nd International Conference on Knowledge Capture, K-CAP '03, pages 70–77.

[16]Langue arabe https://fr.wikipedia.org/wiki/Arabe, consulté le : 13/10/2020.

[17Larkey L. S., Ballesteeros L. and Connell M., improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis, in proceeding of the 25th annual international conference on research and development in information retrieval (SIGIR 2002), tampere, finland, august 2002,pp.275-282.

[18]J. Leclerc, l'aménagement linguistique dans le mond.

[19]Taghva, K., Elkhoury, R., Coombs, J., "Arabic stemming without a root dictionary", Information Technology: Coding and Computing, ITCC, Vol. 1, pp. 154 , 2005.'

[20]Said D., Wanas 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.

[21] Flora Even, l'influence de facebook sur les idées politiques, https://www.rtbf.be/culture/dossier/chroniques-culture/detail_l-influence-de facebook-sur-les-idees-politiques-flora-eveno?id=9458372, Visité le 16/02/2019.

[22] http://www.agoravox.fr/actualites/religions/article/la-langue-arabe-son-histoire-son-77459, consulté le: 13/10/2020.

[25] https://www.lebigdata.fr/python-langage-definition, consulté le: 10/05/2020.

[26] Mohamed Aly , Amir Atiya, 2013,Computer Engineering Department Cairo University Giza, Egypt.

[27] Mohammed Rushdi-Saleh,September2011, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López , José M. Perea-Ortega, , Bilingual Experiments with an Arabic-English Corpus for Opinion Mining, SINAI research group University of Jaén,September2011.