

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad Dahleb Blida -1-

Faculté des sciences

Département informatique



Mémoire de fin d'étude présenté par :

Mosteghanemi Samira et Feroukhi Afaf

Pour l'obtention du diplôme de master en informatique

Spécialité : Traitement automatique de la langue

Thème :

**Systeme de recommandation multilingue basé sur l'analyse des
sentiments des opinions dans les commentaires en ligne**

Soutenu le : 06/12/2020, devant le jury composé de :

Mme.Boumahdi	Université de Blida 1	Présidente
Mme.Bacha	Université de Blida 1	Examinatrice
Mme. Oukid Lamia	Université de Blida 1	Promotrice
Mr. Abbas Mourad	CRSTDLA	Encadreur

Résumé

Connaitre l'opinion ou l'avis des autres personnes sur des produits ou des services a toujours été un élément d'information important durant le processus de décision. Les gens s'intéressent énormément aux avis des autres personnes dans différents domaines, ils consultent les avis des autres consommateurs avant d'effectuer un achat, ou d'utiliser un service. Notre travail a pour objet de concevoir une application permettant d'analyser des commentaires et de donner une recommandation en s'inspirant des techniques de classification automatique de données textuelles. Dans un premier temps nous proposons d'identifier les commentaires qui expriment un avis ou représentant un bruit. Par la suite, pour les vrais avis, nous proposons une évaluation de 1 à 5 (de forte non-recommandation vers forte recommandation). Enfin, après calculer la moyenne des évaluations des avis pour sortir par une recommandation finale. Nous avons effectué une étude expérimentale sur différents algorithmes de classification avant de faire notre choix. Ainsi, les résultats expérimentaux obtenus pour notre solution sont satisfaisant

Mots-clés :

Commentaire en ligne, données textuelles, recommandation finale, apprentissage automatique, SVM, analyse des sentiments.

Abstract

Knowing the opinion or opinion of other people about products or services has always been an important piece of information during the decision-making process. People are very interested in the opinions of other people in different areas, they consult the opinions of other consumers before making a purchase or using a service. Our work aims to design an application to analyze comments and give a recommendation based on automatic classification techniques for textual data. First, we propose to identify the comments that express an opinion or represent noise. Subsequently, for real reviews, we suggest a rating of 1 to 5 (from strong non-recommendation to strong recommendation). Finally, after calculating the average of the ratings of the reviews to come out with a final recommendation. We carried out an experimental study on different classification algorithms before making our choice. Thus, the experimental results obtained for our solution are satisfactory.

Keywords :

Online commentary, textual data, final recommendation, machine learning, SVM, sentiment analysis.

ملخص

لطالما كانت معرفة رأي أو رأي الآخرين حول المنتجات أو الخدمات جزءاً مهماً من المعلومات أثناء عملية صنع القرار. يهتم الناس كثيراً بأراء الآخرين في مناطق مختلفة ، فهم يستشيرون آراء المستهلكين الآخرين قبل الشراء أو استخدام الخدمة. يهدف عملنا إلى تصميم تطبيق لتحليل التعليقات وإعطاء توصية بناءً على تقنيات تصنيف البيانات النصية التلقائية. أولاً ، نقتراح تحديد التعليقات التي تعبر عن رأي أو تمثل ضوضاء. بعد ذلك ، بالنسبة للمراجعات الحقيقية ، نقتراح تصنيفاً من 1 إلى 5 (من عدم التوصية القوية إلى التوصية القوية) أخيراً ، بعد حساب متوسط تقييمات المراجعات للخروج بتوصية نهائية. لقد أجرينا دراسة تجريبية على خوارزميات تصنيف مختلفة قبل اتخاذ قرارنا. وبالتالي ، فإن النتائج التجريبية التي تم الحصول عليها لحلنا مرضية.

كلمات مفتاحية :

التعليق عبر الإنترنت، البيانات النصية ، التوصية النهائية ، التعلم الآلي ، SVM ، تحليل المشاعر.

REMERCIEMENT

Ce mémoire a été réalisé dans le cadre de fin d'études pour l'obtention du diplôme d'ingénieur en informatique. Nous tenons à remercier d'abord dieu de nous avoir donné force et santé pour mener ce travail à terme. Nous tenons à remercier nos parents pour leurs sacrifices, soutien et compréhension durant toutes nos années d'études. Nous tenons à exprimer nos profonds remerciements à **Mme Oukid.L, Mr Abbas.M et Mr Lichouri.M** Pour leurs conseils précieux, leurs remarques et leurs disponibilités lors de l'élaboration de cette étude.

Nous souhaitons exprimer nos remerciement à tous les membres de jury pour avoir bien voulu accepter de participer à ce jury, pouvant ainsi l'intérêt qu'ils portent à ce travail. Nous exprimons notre gratitude à l'ensemble du corps enseignant, techniques et administratif du département d'informatique à l'université de Blida, pour leur disponibilité et leur gentillesse. Et finalement, mes sincères remerciements à tous ceux qui ont contribué de près ou de loin à la réalisation de ce modeste travail.

Dédicace . . .

Je tiens à remercier en premier lieu mes parents

A mon symbole de sacrifice, écoles de mon enfance
qui était mon ombre durant toutes mes années d'études, toi
mon cher **père** qui a tant investi pour assurer mon Avenir. ;

À la source d'amour et tendresse à celle qui a tout donné à
toi ma chère **maman**.


A mon frère **Khaled**, A mes sœurs **Samia et Souad**

Et à tous les membres de ma famille

A tous mes collègues de deuxième année master. Et à tous
mes collègues de l'université.

Samira

<http://www.myhouseonweb.eu/>
MY HOUSE ON WEB



Dédicace . . .

A mon symbole de sacrifice, écoles de mon enfance qui
était mon ombre durant toutes mes années d'études, toi
mon cher **père** qui a tant investi pour assurer mon Avenir.

;

À la source d'amour et tendresse à celle qui a tout donné à
toi ma chère **maman**.

A mes frères, A mes sœurs

Afaf

<http://www.myhouseonweb.eu/>
MY HOUSE ON WEB

Table des matières

<i>Introduction GÉNÉRALE</i>	13
1 Chapitre I: Analyse des sentiments	16
1.1 Introduction.....	17
1.2 Définition de l'analyse des sentiments	17
1.3 Les niveaux d'analyse des sentiments	20
1.4 Les Domaines d'application de l'analyse des sentiments.....	22
1.5 Les problèmes d'analyse des sentiments	24
1.6 Sélection des fonctionnalités dans la classification des sentiments.....	25
1.7 Les approches d'analyse des sentiments.....	26
1.7.1 Apprentissage automatique	28
1.7.1.1 Apprentissage non supervisée	28
1.7.1.2 Apprentissage supervisée	29
1.7.1.3 Apprentissage semi-supervisé.....	34
1.7.2 Approche basée lexicque.....	35
1.7.2.1 Méthode basée dictionnaire.....	35
1.7.2.2 Méthode basée corpus	35
1.7.3 Approche hybride	36
1.7.4 Les travaux connexes	37
1.8 Conclusion	41
2 Chapitre II : Système de recommandation	42
2.1 Introduction.....	43
2.2 La recherche d'information et Filtrage d'information	43
2.2.1 La recherche d'information	43
2.2.2 Le filtrage d'information	44
2.3 Fonctionnalités des systèmes de recommandations	44
2.4 Les systèmes de recommandation	45
2.4.1 Définition de Système de recommandation	45
2.4.2 Les types des systèmes de recommandation	45
2.4.2.1 Recommandation basée sur le Filtrage collaboratif	45
2.4.2.2 Recommandations basées sur le filtrage par contenu	49
2.4.2.3 Recommandation basée sur le filtrage hybride.....	52
2.5 La recommandation via l'analyse des sentiments	54
2.6 Efficacité d'un algorithme de recommandation	55
2.7 Conclusion.....	57

3	Chapitre III : Conception.....	58
3.1	Introduction.....	59
3.2	Architecture du système	59
3.2.1	Les Corpus	61
3.2.2	Prétraitements	61
3.2.3	Vectorisation (TF-IDF)	65
3.2.4	La phase d'apprentissage.....	66
3.2.5	La Phase de Test et Evaluation.....	70
3.2.6	La détection et classification des avis	70
3.2.7	La recommandation finale	71
3.3	Conclusion	71
4	Chapitre IV : Implémentation et Expérimentation	72
4.1	Introduction.....	73
4.2	Environnement de développement	73
4.2.1	Environnement Matériel.....	73
4.2.2	Environnement et langages de programmation utilisés	74
4.2.2.1	Langages de programmation	74
4.2.2.2	Environnement	76
4.3	Réalisation du projet	77
4.3.1	Description de page d'accueil	77
4.3.2	Description de page de la langue anglaise.....	78
4.3.3	Description de page de la langue arabe	80
4.3.4	Description de page d'affichage des résultats	82
4.4	Expérimentations	83
4.4.1	Jeux de données	83
4.4.2	Résultats obtenus	86
4.4.2.1	Résultats obtenus de corpus anglais	86
4.4.2.2	Résultats obtenus de corpus Arabe.....	96
4.4.3	Evaluation	97
4.4.4	Discussion	97
4.5	Conclusion.....	98
	<i>Conclusion générale</i>	99

Table des tableaux

Tableau 1 : Utilisateur x Produit	46
Tableau 2:Les résultats obtenus de stage 1 de l'expérience 1.	87
Tableau 3: Les résultats obtenus de stage 4 del'expérience 1.	89
Tableau 4: Les résultats obtenus de stage5 de l'expérience 1.	90
Tableau 5: Les résultats obtenus de stage2 de l'expérience 2.	91
Tableau 6: Les résultats obtenus de stage4 de l'expérience2.	94
Tableau 7: Les résultats obtenus de stage5 de l'expérience 2.	95
Tableau 8:Les résultats obtenus del'expérience 3.	96
Tableau 9:Les résultats obtenus de la langue arabe.	97

Table des figures

Figure 1 :L'analyse des sentiments au niveau de la phrase .	21
Figure 2:Les niveaux de l'analyse des sentiments .	22
Figure 3:Domaines d'application d'analyse des sentiments .	23
Figure 4: Hiérarchie des approches de classification des sentiments .	26
Figure 5:Le principe de SVM .	30
Figure 6:L'arbre de décision .	32
Figure 7: Architecture générale d'un réseau de neurones artificiels.	33
Figure 8: Le processus de recherche d'informations.	44
Figure 9: L'architecture d'un système de recommandation .	48
Figure 10:L'architecture d'un système de recommandation basé sur le contenu .	51
Figure 11 :L'architecture d'un système de recommandation hybride .	53
Figure 12:Architecture globale du système.	60
Figure 13: Le commentaire de la figure13 sans ponctuations .	62
Figure 14: Exemple d'un commentaire avec la ponctuation.	62
Figure 15:Le commentaire de la figure14 avec les mots vides .	62
Figure 16: Le commentaire de la figure15 sans les mots vides.	63
Figure 18: Le commentaire de la figure16 après la lemmatisation .	64
Figure 17:Exemple de Pos tag pour une phrase en anglais.	64
Figure 19:Page d'accueil partie 1.	77
Figure 20:Interface de la langue anglaise prtie1 .	78
Figure 22:Interface de la langue anglaise prtie3 .	79
Figure 21 : Interface de la langue anglaise prtie2.	79
Figure 23:Interface de la langue arabe partie 1 .	80
Figure 24:Interface de la langue arabe partie 2 .	81
Figure 25:Interface de la langue arabe partie 3 .	81
Figure 26:La page de l'affichage des résultats.	82
Figure 27:Découpage d'un ensemble de données en un ensemble d'apprentissage et un ensemble de test.	83
Figure 28:Opérateur de prétraitement .	84
Figure 29:Opérateur d'analyseur.	84
Figure 30: Les résultats obtenus de stage 2 de l'expérience 1 .	87

Figure 31 : Les résultats obtenus de stage3 de l'expérience 1.	88
Figure 32: Les résultats obtenus de stage 2 de l'expérience 2.	92
Figure 33: Les résultats obtenus de stage 3 de l'expérience2.	93

Liste des Abréviations

SVM : Support Vector Machine
SA : Sentiment Analysis
Ei : le nom d'une entité
Aij : un aspect d'Ei
Hk : détenteur de l'opinion
Tl : le temps où l'opinion a été exprimée
RF: Random Forest
POS: Part-of-speech
NLTK: Natural Language Toolkit
NLP: Natural Language Processing
NB: Naïve Bayes
LSVC: Linear Support Vector Classification
KNN: K Nearest Neighbor
LR: Logistic Regression
DT: Decision Tree
GB: Gradient Boosting
BNB: Bernoulli Naïve Bayes
MNB: Multinomial Naïve Bayes
LSA :Latent Semantic Analysis
RNA : Réseau de neurone Artificiel

Introduction GÉNÉRALE

De nos jours, Internet est devenu plus que jamais un outil essentiel et puissant pour l'échange d'informations dans le monde entier. Elle nous offre une quantité considérable de données et ses services qui s'adaptent de plus en plus aux besoins des internautes. D'un autre côté, les gens périodiquement cherchent des recommandations sur différents services ou produits pour ne pas tomber dans les mauvaises décisions que d'autres personnes ont déjà pris. Les avis des clients sur les produits jouent un rôle important dans la décision du client d'acheter un produit ou d'utiliser un service. Les préférences et les opinions des clients sont influencées par les commentaires des autres clients en ligne, sur les blogs ou sur les plateformes de réseaux sociaux. En analysant et en regroupant les activités et le comportement des autres utilisateurs principalement sous forme d'évaluations, et en faisant les meilleurs recommandations. A cet égard, le système d'analyse des sentiments pourrait être considéré comme un coup de pouce à la performance d'un système de recommandation classique. C'est pourquoi la combinaison d'un système de recommandation classique basé sur la notation des évaluateurs et d'un système d'analyse des sentiments est indispensable.

De nombreuses études ont montré que l'analyse du sentiment est très intéressante pour les personnes qui se concentrent sur l'opinion publique, pour de nombreuses raisons personnelles, commerciales ou politiques. Selon une étude faite par **Pew Research Center**, 20% des utilisateurs de réseaux sociaux ont modifié leur avis politique sur un sujet à cause de ce qu'ils ont vu sur les médias sociaux [60]. Par conséquent, il est devenu important de concevoir des systèmes et des applications automatiques capables d'analyser, de détecter et d'extraire les évaluations des sentiments et les opinions des individus qui sont exprimés sur les web ou réseaux sociaux.

Notre objectif est de réaliser un système de recommandation qui cible automatiquement les commentaires qui expriment réellement des avis de ceux qui représentent un bruit. Et en second lieu, après détection des commentaires portant d'avis, le système lui donnera une évaluation précise de forte non-recommandé vers forte recommandé (1 à 5) pour chaque avis, et calculer la moyenne des évaluations des avis d'un produit pour faire la recommandation finale. Ceci en se basant sur les méthodes de classification supervisées, notre but dans ce travail est également de prendre en compte différents langages. Nous nous intéressons particulièrement aux langages anglais et arabe.

Le présent travail est structuré en quatre chapitres :

Le premier chapitre présente le problème de l'analyse du sentiment, ses différents niveaux, ses approches et quelques travaux connexes.

Le deuxième chapitre présente les familles des systèmes de recommandations et la recommandation via l'analyse des sentiments avec quelques travaux existants.

Le troisième chapitre concerne la conception et l'architecture de notre système.

Le dernier chapitre présente la partie implémentation et expérimentations de notre système, où nous présenterons l'environnement de développement, la structure de notre application et quelques interfaces de celle-ci.

Enfin, nous clôturons ce mémoire par une conclusion générale et quelques perspectives.

1 Chapitre I: Analyse des sentiments

1.1 Introduction

Dans le monde réel, les entreprises et les organisations veulent toujours trouver les opinions des consommateurs ou du public sur leurs produits et services. Les consommateurs individuels souhaitent également connaître les opinions des utilisateurs existants d'un produit avant de l'acheter pour prendre une décision. De nos jours, si l'on veut acheter un produit, il existe de nombreux avis d'utilisateurs et discussions dans les forums publics sur le Web à propos du produit [1]. D'où la nécessité d'analyser des opinions, ou d'avis afin de savoir ce que pensent les internautes. Un processus d'analyse des sentiments, qui peut analyser si un utilisateur fournit un bon ou une mauvaise opinion sur un certain produit, ou un service.

Dans ce chapitre nous allons présenter quelques concepts et définitions liées à l'analyse des sentiments. Sans oublier de mentionner ses problèmes et les domaines d'application, avec certains travaux liés à ce type de tâche.

1.2 Définition de l'analyse des sentiments

L'analyse des sentiments a plusieurs autres noms tels que l'extraction d'opinion, la détection d'émotions, etc. C'est une tâche qui tente de classer les données textuelles avec diverses méthodes.

Bing Liu [2] a donné une définition à cette tâche :

"L'analyse du sentiment, aussi appelée opinion mining, est le domaine d'étude qui analyse les opinions, les sentiments, les évaluations, les attitudes et les émotions des gens envers les entités et leurs attributs exprimés dans un texte écrit."

D'une autre façon, l'analyse du sentiment concerne le traitement d'un texte d'opinion pour extraire la polarité des sentiments généralement exprimée en termes d'opinion positive ou négative (classification binaire), ou d'une classification multiple, où le sentiment peut avoir une étiquette neutre ou même d'autres étiquettes différentes comme très positive, positive, neutre, négative, très négative, les étiquettes peuvent aussi être associées à des émotions comme la tristesse, la colère, le bonheur, etc. Et même des valeurs numériques.

Il est l'un des domaines de recherche les plus actifs dans le traitement du langage naturel et est également largement étudié dans data mining, Web mining, et text-mining. L'importance croissante de l'analyse des sentiments coïncide avec la croissance des médias sociaux tels que les critiques, les discussions de forum, les blogs, les micro-blogs. [4].

Définition 1 :

Le traitement automatique de la langue naturelle (TALN) ou des langues (TAL) est une discipline à la frontière de la linguistique, de l'informatique et de l'intelligence artificielle. Elle concerne la conception de systèmes et techniques informatiques permettant de manipuler le langage humain dans tous ses aspects [5].

Définition 2 :

➤ Définition de sentiment :

Le sentiment est la composante de l'émotion qui implique les fonctions cognitives de l'organisme, la manière d'apprécier. Le sentiment est à l'origine d'une connaissance immédiate ou d'une simple impression. Il renvoie à la perception de l'état physiologique du moment. Le sens psychologique de sentiment qui comprend un état affectif est à distinguer du sens propre de la sensibilité [6].

➤ Définition d'opinion :

Une opinion c'est un Jugement, avis, sentiment qu'un individu ou un groupe émet sur un sujet, des faits, ce qu'il en pense [7].

Une opinion est un quintuplet (ei ,aij,oijkl,hk ,tl), où ei est l'entité, aij est un des aspects de (ei , hk) est le titulaire, oijkl est son orientation (polarité), et tl est la date (time) où elle a été exprimée. Parce que bien sûr, une opinion peut être fluctuante dans le temps.

On peut classer les types d'opinions en opinion régulière et un autre type est appelé opinion comparative .En fait, on peut aussi classer les opinions en fonction de la manière dont ils sont exprimés dans le texte, l'opinion explicite et l'opinion implicite [4].

a) **Opinion régulière ou comparative :**

❖ **Opinion régulière:**

Une opinion régulière est souvent appelée simplement une opinion dans la littérature et il comporte deux sous-types principaux [4]:

Opinion directe: Une opinion directe fait référence à une opinion exprimée directement sur une entité ou un aspect de l'entité, par exemple, “**The picture quality is great.**”

Opinion indirecte: Une opinion indirecte est une opinion exprimée indirectement sur une entité ou un aspect d'une entité en fonction de ses effets sur d'autres entités. Ce sous-type se produit souvent, dans le domaine médical. Par exemple, “**After injection of the drug, my joints felt worse** ” décrit un effet indésirable du médicament sur« mes articulations », une opinion ou un sentiment négatif à l'égard du médicament. Dans le cas, l'entité est le médicament et l'aspect est l'effet sur les articulations.

❖ **Opinion comparative :**

Une opinion comparative exprime une relation de similitudes ou de différences entre deux ou plusieurs entités et / ou une préférence du porteur d'opinion basée sur certains aspects communs des entités, Par exemple, les phrases, “**Coke tastes better than Pepsi** ” and “**Coke tastes the best** ”

Expriment deux comparatifs des avis. Une opinion comparative est généralement exprimée en utilisant la forme comparatif ou le superlatif d'un adjectif ou d'un adverbe, mais

pas toujours (par exemple, préférer). Des opinions comparatives également ont de nombreux types.

b) **Opinion explicite ou implicite :**

Opinion explicite: Une opinion explicite est une déclaration subjective qui donne une opinion régulière ou comparative, par exemple “**Coke tastes great,**” and “**Coke tastes better than Pepsi.**”.

Opinion implicite : Expressions d'aspect qui ne sont pas des noms ou des phrases nominales mais indiquant que certains aspects sont appelés expressions d'aspect implicites. Par exemple, « **expensive** » est une expression d'aspect implicite dans “**This camera is expensive.**” Cela implique l'aspect prix. De nombreuses expressions d'aspect implicites sont adjectifs et adverbes utilisés pour décrire ou qualifier certains aspects spécifiques, par exemple, cher (prix) et fiable (fiabilité). Ils peuvent également être des phrases verbales et verbales, par exemple, “**I can install the software easily.**” Install indique l'aspect installation.

1.3 Les niveaux d'analyse des sentiments

➤ **Niveau du document :**

La tâche à ce niveau est de classer si un document d'opinion exprime un sentiment positif ou négatif. Par exemple, étant donné un avis de produit, le système détermine si l'avis exprime une opinion globale positive ou négative sur le produit. Cette tâche est communément appelée classification des sentiments au niveau du document [4]. C'est la forme la plus simple de classification, la totalité de document de texte d'opinion est considéré comme élémentaire unité d'information. On suppose que le document est avoir une opinion sur un seul objet (film, livre ou Hôtel). Cette approche ne convient pas si le document contient des opinions sur différents objets comme dans les forums et les blogs [8].

➤ **Niveau de la phrase :**

La tâche à ce niveau va aux phrases et détermine si chaque phrase a exprimé une opinion positive, négative ou neutre. Neutre ne signifie généralement pas d'opinion. Ce niveau

d'analyse est étroitement lié à la classification de la subjectivité qui distingue les phrases qui expriment des informations à partir de phrases qui expriment des vues et des opinions subjectives [4]. L'analyse des sentiments au niveau de la phrase est l'analyse la plus fine du document. En cela, la polarité est calculée pour chaque phrase car chaque phrase est considérée comme une unité distincte et chaque phrase peut avoir une opinion différente. L'analyse a deux tâches [8]:

❖ *Classification de la subjectivité*

Une phrase peut être une phrase subjective ou phrase objective. La phrase objective contient les faits. Il n'a pas de jugement ou d'opinion sur l'objet ou entité alors que la phrase subjective a des opinions (par exemple), « L'économie indienne est fortement dépendante du tourisme et de l'informatique industrie », « C'est un excellent endroit où vivre ». La première phrase est factuelle et ne transmet aucun sentiment envers l'Inde. L'avantage de l'analyse au niveau des phrases réside dans la classification subjectivité / objectivité.

❖ *Classification des sentiments*

La phrase peut être classée comme positive, négative ou neutre en fonction des mots d'opinion présents dans la phrase.

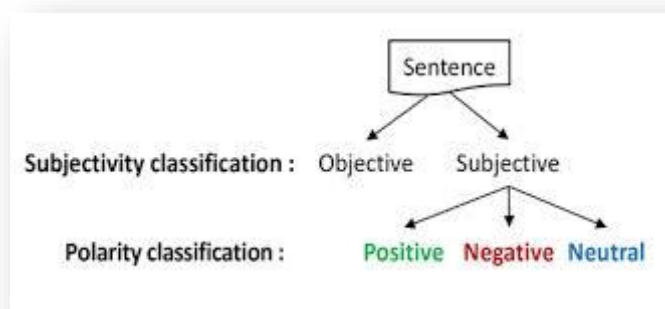


Figure 1:L'analyse des sentiments au niveau de la phrase [56].

➤ **Niveau d'entité et d'aspect :**

Les analyses au niveau du document et de la phrase ne découvrent pas exactement ce que les gens ont aimé et n'a pas aimé. Le niveau d'aspect effectue une analyse détaillée. Le niveau d'aspect était auparavant appelé niveau de fonctionnalité. Au lieu de regarder les constructions de langage, le niveau d'aspect regarde directement l'opinion elle-même. Elle repose sur l'idée qu'une opinion est constituée d'un sentiment et d'une cible. Une opinion sans que sa cible soit identifiée est d'une utilité limitée. Réaliser l'importance des cibles d'opinion nous aide également à mieux comprendre le problème de l'analyse des sentiments [4].



Figure 2:Les niveaux de l'analyse des sentiments [56].

1.4 Les Domaines d'application de l'analyse des sentiments

L'analyse du sentiment a retenu beaucoup l'attention ces dernières années en raison de ses nombreuses applications passionnantes dans une variété de domaines [9].

Les applications d'analyse du sentiment se sont étendues à presque tous les domaines possibles, des produits de consommation, des soins de santé, du tourisme, de l'hôtellerie, des services financiers et des événements sociaux. En plus des intérêts commerciaux, de multiples projets réels tentent d'implémenter des modèles d'analyse du sentiment comme les restaurants, les cosmétiques, les films, l'automobile, les compagnies de téléphonie mobile, etc. Comme le montre l'image ci-dessous [3].

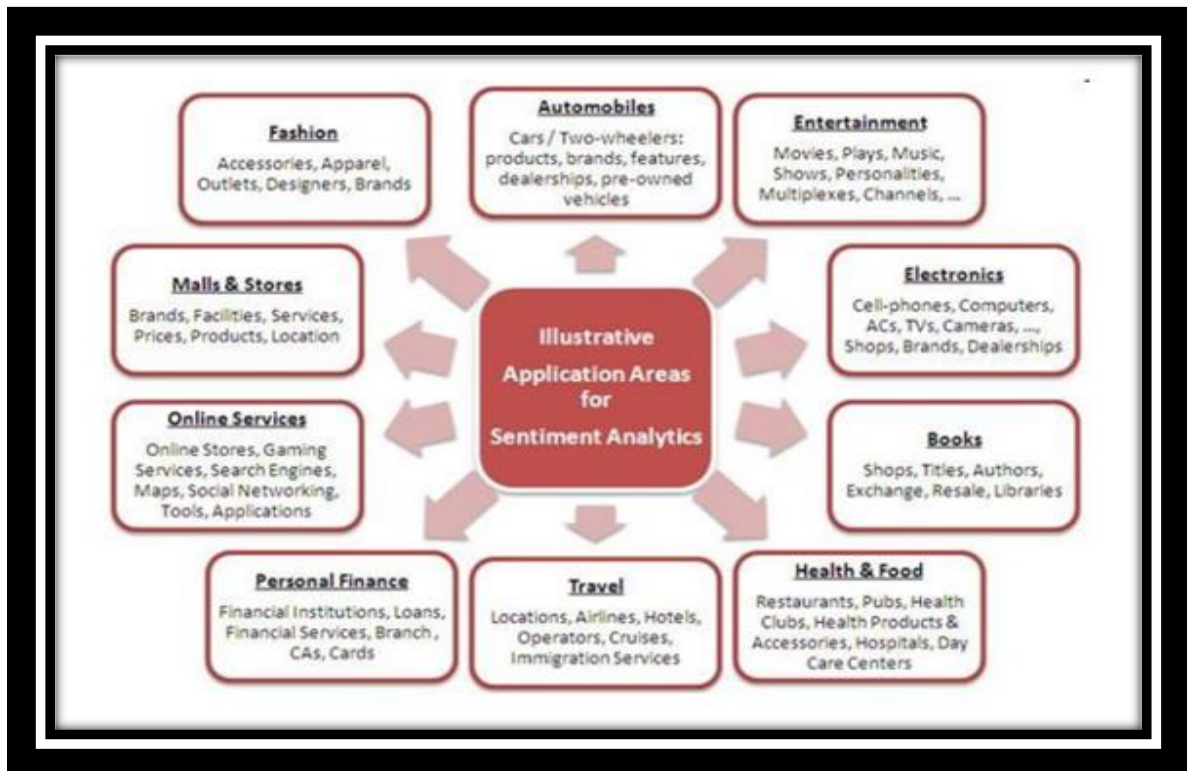


Figure 3: Domaines d'application d'analyse des sentiments [57].

Nous verrons certaines de ces applications [10] :

➤ **Applications en tant que technologie de sous-composant :**

Un système de prédiction de sentiment peut être utile pour les systèmes de recommandation aussi. Le système de recommandation ne recommandera pas les articles qui reçoivent beaucoup de commentaires négatifs ou moins de notes. Dans la communication en ligne, nous rencontrons un langage abusif et d'autres éléments négatifs. Ceux-ci peuvent être détectés simplement par identifier un sentiment très négatif et en conséquence prendre des mesures contre elle.

➤ **Applications en Business Intelligence :**

Il a été observé que les gens ont aujourd'hui tendance à regarder avis sur les produits disponibles en ligne avant leur achat. Et pour de nombreuses entreprises, l'opinion en ligne

décide du succès ou échec de leur produit. Ainsi, l'analyse des sentiments joue un rôle important dans les entreprises. Les entreprises souhaitent également extraire le sentiment des critiques en ligne afin de s'améliorer leurs produits et à leur tour leur réputation et leur aide à la clientèle la satisfaction.

➤ **Applications dans les maisons intelligentes :**

Les maisons intelligentes sont censées être la technologie du futur. À l'avenir, des maisons entières seraient mises en réseau et les gens être en mesure de contrôler n'importe quelle partie de la maison à l'aide d'une tablette. Récemment, de nombreuses recherches ont été menées sur Internet sur Choses (IoT). L'analyse des sentiments trouverait également son chemin dans IoT. Comme par exemple, basé sur le sentiment actuel ou émotion de l'utilisateur, la maison pourrait modifier son ambiance pour créer un environnement apaisant et paisible.

1.5 Les problèmes d'analyse des sentiments

➤ **Construire un classificateur pour les tweets subjectifs vs. Objectifs :**

Les travaux en cours se concentrent principalement sur la classification correcte du positif et du négatif. Il est nécessaire d'examiner la classification des tweets avec ou sans sentiment [12].

➤ **La détection automatique des sarcasmes :**

Le sarcasme [27] est généralement vu comme un sous type de l'ironie verbale ou sous forme des blagues, le sarcasme est une ironie mordante parce qu'elle s'exprime avec aigreur et emportement. Peuvent conduire à une mauvaise compréhension du sens. Comme exemple :

" Nice play, you should singe with Real Madrid."

Que des mots positifs, mais en réalité signifie un sentiment négatif.

➤ Identifier les parties subjectives du texte:

Les parties subjectives représentent un contenu porteur de sentiment. Le même mot peut être traité comme subjectif dans un cas, ou comme objectif dans un autre. Cela rend difficile l'identification du subjectif portions de texte, Par exemple:

“The language of the Mr.Dennis was very crude.”

“Crude oil is obtained by extraction from the sea beds.”

Le terme "Crude" est utilisé dans le premier exemple sous forme d'opinion, alors qu'il est tout à fait objectif dans le second [10].

➤ Dépendance de domaine:

La même phrase ou expression peut avoir des significations différentes dans différents domaines. Par exemple, le mot « **unpredictable** » est positif dans le domaine des films, des drames, etc., mais si le même mot est utilisé dans le contexte de la direction d'un véhicule, alors il a un avis négatif [10].

1.6 Sélection des fonctionnalités dans la classification des sentiments

La tâche d'analyse des sentiments est considérée comme un problème de classification des sentiments. La première étape du problème SC consiste à extraire et à sélectionner des fonctionnalités de texte. Certaines des fonctionnalités actuelles sont [13]:

➤ Présence et fréquence des termes

Ces caractéristiques sont des mots individuels ou des n-grammes de mots et leur nombre de fréquences. Il donne soit la pondération binaire des mots (zéro si le mot apparaît, soit un sinon) ou utilise des pondérations de fréquence de terme pour indiquer l'importance relative des caractéristiques.

➤ Parts of speech (POS)

Trouver des adjectifs, car ce sont des indicateurs importants des opinions.

➤ **Mots et expressions d'opinion**

Ce sont des mots couramment utilisés pour exprimer des opinions, notamment bonnes ou mauvaises, aime ou déteste. D'un autre côté, certaines phrases expriment des opinions sans utiliser de mots d'opinion.

➤ **Négations**

L'apparition de mots négatifs peut changer l'orientation de l'opinion comme pas bon est équivalent à mauvais.

1.7 Les approches d'analyse des sentiments

Les techniques de classification des sentiments peuvent être grossièrement divisées en une approche d'apprentissage automatique, une approche basée sur le lexique et une approche hybride. L'approche Machine Learning (ML) applique les célèbres algorithmes ML et utilise des fonctionnalités linguistiques. L'approche basée sur le lexique repose sur un lexique des sentiments, une collection de termes de sentiments connus et précompilés. Il est divisé en une approche basée sur un dictionnaire et une approche basée sur un corpus qui utilisent des méthodes statistiques ou sémantiques pour trouver la polarité des sentiments. L'approche hybride combine les deux approches et est très courante, les lexiques de sentiment jouant un rôle clé dans la majorité des méthodes.

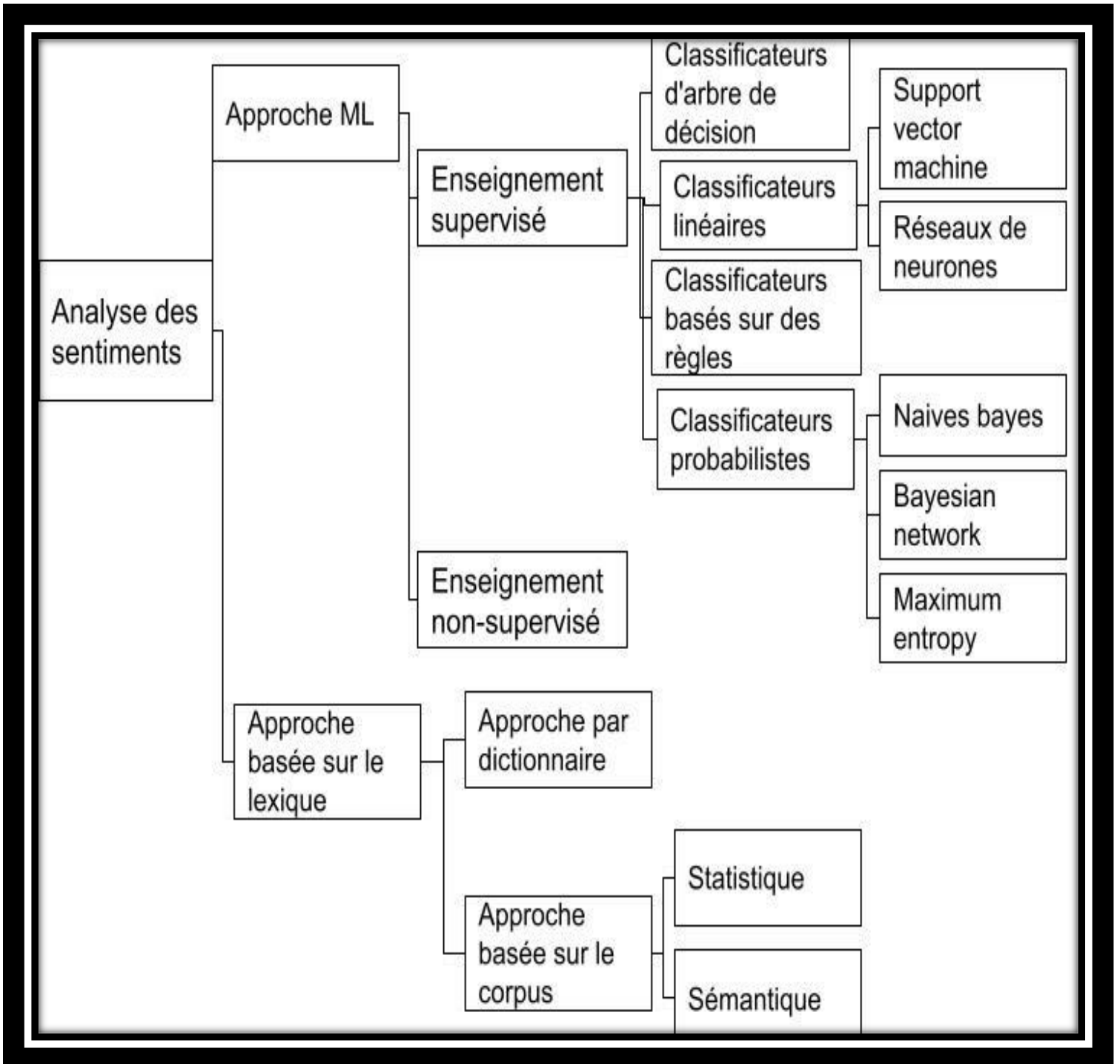


Figure 4: Hiérarchie des approches de classification des sentiments [13].

1.7.1 Apprentissage automatique

L'approche de l'apprentissage automatique s'appuie sur les célèbres algorithmes ML pour résoudre la SA en tant que problème de classification de texte régulier qui utilise des fonctionnalités syntaxiques et / ou linguistiques [13].

Il s'agit de concevoir des algorithmes capables, à partir d'un nombre important d'exemples, d'en assimiler la nature afin de pouvoir appliquer ce qu'ils ont ainsi appris aux cas futurs. Ainsi, le but essentiel de l'apprentissage automatique est de déterminer la relation entre les objets et leurs catégories pour la prédiction et la découverte des connaissances.

On distingue ainsi trois types d'apprentissage: l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage semi-supervisé.

1.7.1.1 *Apprentissage non supervisé*

L'apprentissage non supervisé [15] a été largement étudié dans la communauté du Machine Learning, et les algorithmes de clustering, de réduction de dimensionnalité ou de densité sont régulièrement utilisés dans plusieurs domaines :

- **Médecine**
- **Le traitement de la parole**
- **Traitement d'images.**
- **Classification de documents.**
- **l'analyse des opinions.**

Est qu'ils peuvent être aussi appliqués sur n'importe quel domaine ou ensemble de données spécifique, comme un satellite.

❖ ***Algorithmes de classification non supervisée :***

- **K-moyennes (K-Means)** : Un algorithme de partitionnement des données en K groupes ou clusters. Chaque objet sera associé à un seul cluster. Le K est fixé par l'utilisateur.
- **Fuzzy K-Means** : Il s'agit d'une variante du précédent algorithme proposant qu'un objet ne soit pas associé qu'à un seul groupe.
- **Regroupement hiérarchique.**

1.7.1.2 *Apprentissage supervisé*

La classification supervisée consiste à partir de la description de l'élément détermine sa classe avec le plus faible taux d'erreurs. La performance de la classification dépend notamment de l'efficacité de la description. De plus, si l'on veut obtenir un système d'apprentissage, la procédure de classification doit permettre de classer efficacement tout nouvel exemple (pouvoir prédictif).

❖ *Algorithmes de classification supervisée :*

Il existe de nombreuses méthodes d'apprentissage supervisé:

➤ *Les machines à support de vecteurs (SVM) :*

Les « Supports Vectors Machines » appelés aussi « Maximum Margin Classifier» est des techniques d'apprentissage supervisé basées sur la théorie de l'apprentissage statistique ou automatique, est très rapide et efficace pour les problèmes de classification de texte [13].

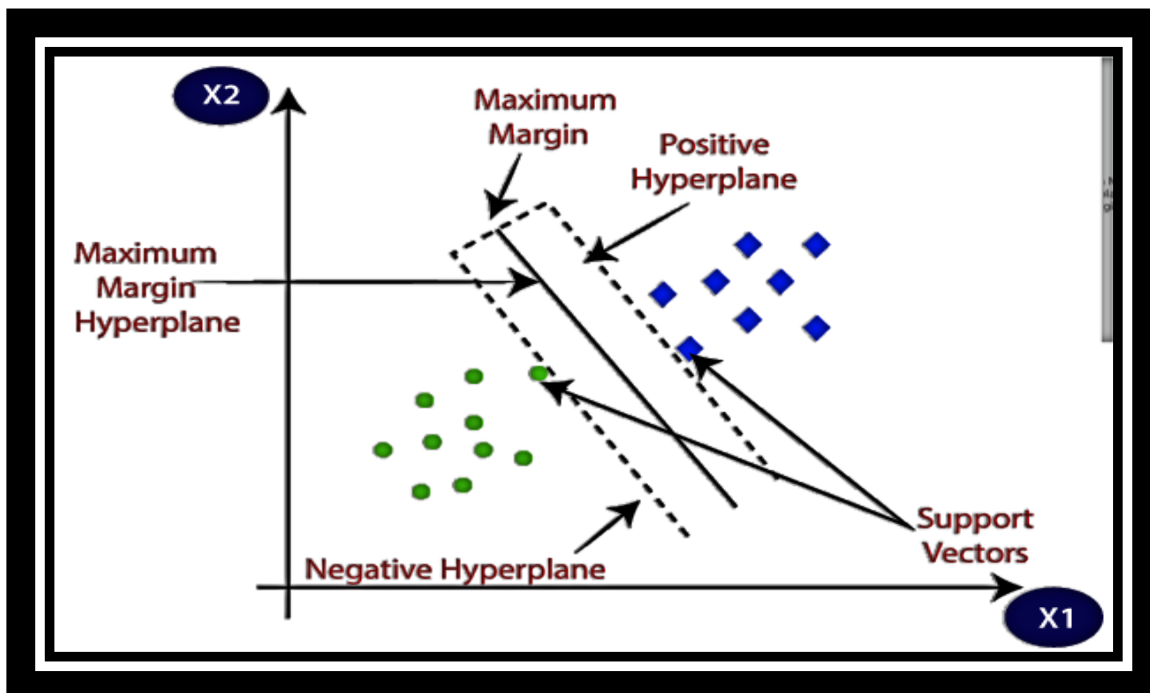


Figure 5:Le principe de SVM [58].

➤ *Naïve Bayes* :

Un classificateur Naïve Bayes est un modèle d'apprentissage automatique probabiliste utilisé pour la tâche de classification. Le nœud du classificateur est basé sur le théorème de Bayes [17]. Certains utilisent un modèle multivarié de Bernoulli (BNB), c'est-à-dire un réseau bayésien. D'autres utilisent un modèle multinomial (MNB), c'est-à-dire un modèle de langage uni-gramme avec des nombres de mots entiers.

➤ **K plus proches voisins (KPPV) :**

L'algorithme KNN est une approche basée sur des instances bien connue qui a été largement appliquée à la catégorisation des textes en raison de sa simplicité et de sa précision. Pour classer un document inconnu, le KNN classifie les voisins du document parmi les documents de formation et utilise les libellés de classe du k voisins les plus similaires. La similitude entre deux documents peut être mesurée par la distance euclidienne, mesure cosinus, etc. Le score de similarité de chaque document le plus proche voisin du test

document est utilisé comme poids des classes du document voisin. Si une catégorie spécifique est partagée par plus d'un des k voisins les plus proches, alors la somme des scores de similarité de ces voisins est obtenue à partir du poids de cette catégorie partagée particulière, Lorsque la classification est effectuée au moyen du KNN, le paramètre le plus important affectant la classification est le nombre k-plus proche voisin. Habituellement, la valeur optimale de k est déterminée empiriquement. La valeur k est déterminée de manière à donner la moindre erreur de classification [16].

La distance entre un texte et ses voisins se fait via une métrique de distance. Cette métrique peut être comme suit :

- ✓ **Mesure Cosinus** qui consiste à calculer le produit scalaire entre deux vecteurs a et b, que nous divisons par le produit de la norme de ces deux vecteurs. La formule de la mesure Cosinus est :

$$\text{cosinus } (a, b) = \frac{\sum(a * b)}{\sqrt{\sum a^2 * \sum b^2}} \quad (1)$$

- ✓ **Mesure de Distance euclidienne** La formule de la mesure de Distance est comme suivante :

$$D (a, b) = \sqrt{\sum |a - b|^2} \quad (2)$$

- ✓ **Mesure de Jaccard** La formule de la mesure de Jaccard est :

$$J (a, b) = \frac{\sum a * b}{\sum a^2 + \sum b^2 - \sum ab} \quad (3)$$

L'algorithme ci-dessous montre comment classer un nouvel exemple par la méthode K plus proche voisin KPPV :

➤ **Les arbres de décision (DT):**

La popularité de la méthode repose en grande partie sur sa simplicité. Il s'agit de trouver un partitionnement des individus que l'on représente sous la forme d'un arbre de décision. L'objectif est de produire des groupes d'individus les plus homogènes possibles du point de vue de la variable à prédire. Il est d'usage de représenter la distribution empirique de l'attribut à prédire sur chaque sommet (nœud) de l'arbre [18].

Concrètement, chaque nœud d'un arbre de décision contient un test (un IF...THEN) et les feuilles ont les valeurs Oui ou Non. Chaque test regarde la valeur d'un attribut de chaque exemple. En effet, on suppose qu'un exemple est un ensemble d'attributs/valeurs. Pour des textes, chaque attribut peut être un mot et la valeur sera par exemple 0 ou 1 selon que ce mot appartient ou non au document.

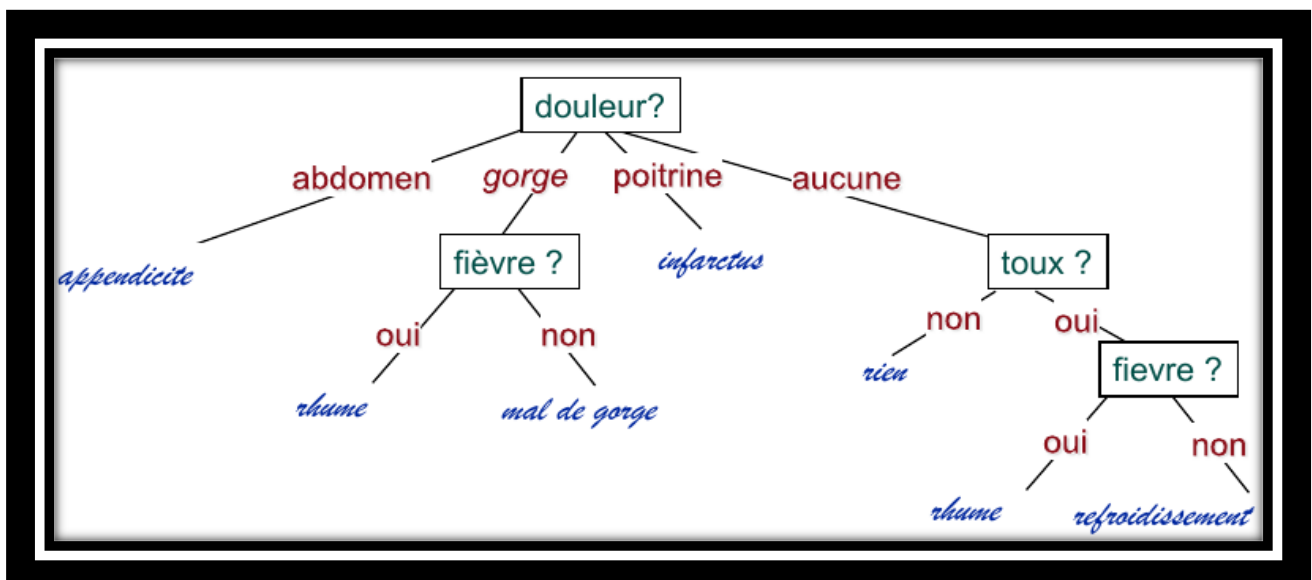


Figure 6:L'arbre de décision [69].

➤ **Réseau neuronal(RNA) :**

Un réseau de neurones formels à temps discret est un système composé de deux types d'éléments, ou unité : les entrées du réseau et les neurones eux-mêmes. Chaque neurone (déterministe) est un processeur non linéaire (généralement simulé sur ordinateur, parfois réalisé sous forme de circuit électronique) qui, a chaque instant discret k , calcule son potentiel $v_i(k)$ et son activité $z_i(k)$ de la façon suivante. [19]:

$$z_i(k) = f_i(v_i(k)) \quad \text{où} \quad v_i(k) = \sum_{j \in P_i} \sum_{\tau=0}^{q_{ij}} c_{ij,\tau} z_j(k-\tau) \quad (4)$$

Un réseau de neurones est conçu pour remplir une tâche que le concepteur définit par un ensemble de valeurs d'entrée, et par un ensemble de valeurs désirées correspondantes pour les activités de certains neurones du réseau que l'on appelle neurones de sortie (les éléments de ces ensembles sont appelés « exemples d'apprentissage »). Les neurones qui ne sont pas des neurones de sortie sont dits cachés.

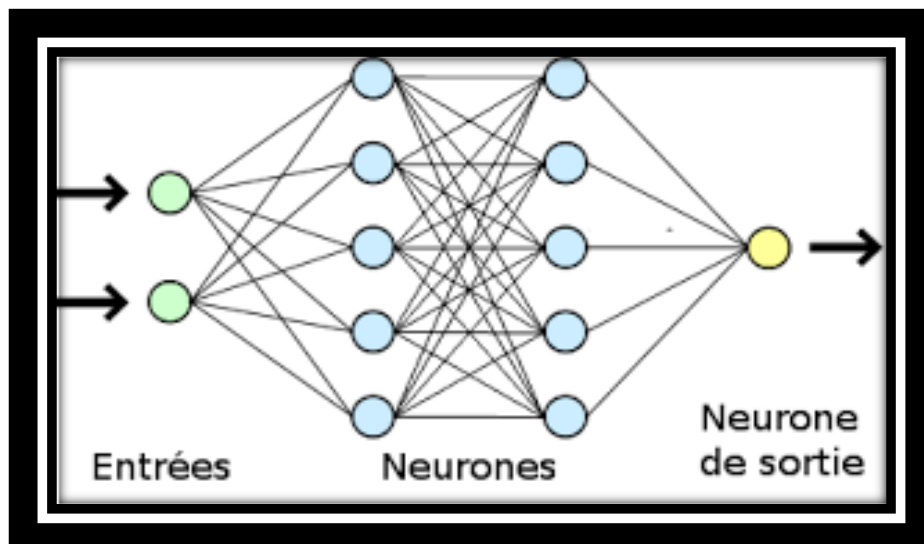


Figure 7: Architecture générale d'un réseau de neurones artificiels.

➤ *Forêts d'arbres décisionnels (Random Forest) RF :*

La forêt aléatoire est un algorithme d'apprentissage supervisé qui est utilisé à la fois pour la classification. Mais cependant, il est principalement utilisé pour les problèmes de classification. Comme nous savons qu'une forêt est composée d'arbres et plus d'arbres signifie une forêt plus robuste. De même, l'algorithme de forêt aléatoire crée des arbres de décision sur des échantillons de données, puis obtient la prédiction de chacun d'eux et sélectionne enfin la meilleure solution au moyen du vote. Il s'agit d'une méthode d'ensemble qui est meilleure qu'un arbre de décision unique car elle réduit le sur-ajustement en faisant la moyenne du résultat [20].

➤ *Logistic Regression(LR) :*

La régression logistique est l'un des algorithmes d'apprentissage automatique les plus simples et les plus couramment utilisés pour la classification à deux classes. Il est facile à mettre en œuvre et peut être utilisé comme référence pour tout problème de classification binaire. Ses concepts fondamentaux de base sont également constructifs dans l'apprentissage en profondeur. La régression logistique décrit et estime la relation entre une variable binaire dépendante et des variables indépendantes [21].

1.7.1.3 *Apprentissage semi-supervisé*

En contexte semi-supervisé [22], on dispose de deux échantillons de données : $S_{lab} = \{(x_1, y_1), \dots, (x_l, y_l)\}$, un échantillon de données étiquetées, et $S_{unl} = \{x_0^1, \dots, x_0^l\}$, un échantillon de données non étiquetées. On modélise la présence de ces deux échantillons par l'existence d'un oracle qui, avec une certaine probabilité β fournit un exemple étiqueté et avec une probabilité $1 - \beta$ procure un exemple non étiqueté. Le paramètre β complète le modèle θ vu précédemment. Dans ce nouveau modèle, que nous notons θ' , les probabilités d'avoir $z = (x, y) \in S_{lab}$ et d'avoir $z = x \in S_{unl}$ se calculent de la manière suivante :

$$\begin{aligned} P(z = (x, y) | \theta', z \in S_{lab}) &= \beta \cdot P(x, y | \theta) \\ P(z = x | \theta', x \in S_{unl}) &= (1 - \beta) \cdot P(x | \theta) \\ \text{avec } P(x | \theta) &= P(y = 1 | \theta) \cdot P(x, y | y = 1, \theta) + P(y = 0 | \theta) \cdot P(x, y | y = 0, \theta) \end{aligned} \quad (5)$$

La vraisemblance de Slab et Sunl pour le modèle θ' s'écrit alors :

$$\begin{aligned}
 L(\theta', S_{lab}, S_{unl}) &= \prod_{s=1}^l \beta P(x_s, y_s | \theta) \prod_{r=1}^{l'} (1 - \beta) P(x'_r | \theta) \\
 &= \beta^l L(\theta, S_{lab}) (1 - \beta)^{l'} \prod_{r=1}^{l'} \left(\alpha \prod_{\substack{1 \leq i \leq m \\ k/x'_r = k}} \lambda_{ik1} + (1 - \alpha) \prod_{\substack{1 \leq i \leq m \\ k/x'_r = k}} \lambda_{ik0} \right)
 \end{aligned} \tag{6}$$

1.7.2 Approche basée lexicque

Les mots d'opinion sont utilisés dans de nombreuses tâches de classification des sentiments. Les mots d'opinion positifs sont utilisés pour exprimer certains états souhaités, tandis que les mots d'opinion négatifs sont utilisés pour exprimer certains états indésirables.

1.7.2.1 Méthode basée dictionnaire

A présenté la stratégie principale de l'approche par dictionnaire. Un petit ensemble de mots d'opinion est collecté manuellement avec des orientations connues. Ensuite, cet ensemble est développé en recherchant dans les corpus bien connus WordNet ou thésaurus leurs synonymes et antonymes.

L'approche par dictionnaire présente un inconvénient majeur qui est l'incapacité à trouver des mots d'opinion avec des orientations spécifiques au domaine et au contexte [13].

1.7.2.2 Méthode basée corpus

L'approche basée sur le corpus [13] aide à résoudre le problème de la recherche de mots d'opinion avec des orientations spécifiques au contexte. Ses méthodes dépendent des modèles syntaxiques ou des modèles qui se produisent les algorithmes et applications d'analyse des sentiments.

L'approche basée sur le corpus est réalisée en utilisant une approche statistique ou une approche sémantique comme illustré dans les sous-sections suivantes:

1.7.2.2.1 Approche statistique :

La recherche de modèles de cooccurrence ou de mots d'opinion peut être effectuée à l'aide de techniques statistiques. Cela pourrait être fait en dérivant les polarités postérieures en utilisant la cooccurrence d'adjectifs dans un corpus. Il est possible d'utiliser l'ensemble complet des documents indexés sur le Web comme corpus pour la construction du dictionnaire. Ceci résout le problème de l'indisponibilité de certains mots si le corpus utilisé n'est pas assez grand.

Latent Semantic Analysis (LSA) est une approche statistique qui est utilisée pour analyser les relations entre un ensemble de documents et les termes mentionnés dans ces documents afin de produire un ensemble de modèles significatifs liés aux documents et aux termes.

1.7.2.2.2 Approche sémantique :

L'approche sémantique donne directement des valeurs de sentiments et s'appuie sur différents principes pour calculer la similitude entre les mots. Ce principe donne des valeurs de sentiment similaires aux mots sémantiquement proches. Par exemple, WordNet fournit différents types de relations sémantiques entre les mots utilisés pour calculer les polarités des sentiments. WordNet pourrait également être utilisé pour obtenir une liste de mots de sentiments en élargissant de manière itérative l'ensemble initial avec des synonymes et d'antonymes, puis en déterminant la polarité des sentiments pour un mot inconnu par le nombre relatif de synonymes positifs et négatifs de ce mot.

1.7.3 Approche hybride

Le concept de méthodes hybrides est très intuitif : combinez simplement le meilleur des deux approches, celui basé sur des règles et celui automatique. Généralement, en combinant les deux approches, les méthodes peuvent améliorer la précision [23].

1.7.4 Les travaux connexes de l'analyse des sentiments

Divers travaux anciens et récents ont été faits dans ce domaine de l'analyse du sentiment. Dans ce qui suit, nous avons mentionné quelques-unes :

Un classificateur NB amélioré a été proposé par Kang et Yoo [31] pour résoudre le problème de la tendance de la précision de classification positive à apparaître jusqu'à environ 10% plus élevée que la précision de classification négative. Cela crée un problème de diminution de la précision moyenne lorsque les précisions des deux classes sont exprimées en valeur moyenne. Ils ont montré que l'utilisation de cet algorithme avec des critiques de restaurants a réduit l'écart entre la précision positive et la précision négative par rapport à NB et SVM. La précision est améliorée dans le rappel et la précision par rapport à la fois au NB et au SVM.

Les SVM ont été utilisés par Li et Li [30] comme classificateurs de polarité des sentiments. Contrairement au problème de classification binaire, ils ont fait valoir que la subjectivité de l'opinion et la crédibilité de l'exprimeur devraient également être prises en considération. Ils ont proposé un cadre qui fournit un résumé numérique compact des opinions sur les plateformes de micro-blogs. Ils ont identifié et extrait les sujets mentionnés dans les opinions associées aux requêtes des utilisateurs, puis classé les opinions à l'aide de SVM. Ils ont travaillé sur des publications Twitter pour leur expérience. Ils ont découvert que la prise en compte de la crédibilité des utilisateurs et de la subjectivité des opinions est essentielle pour agréger les opinions des micro-blogs. Ils ont prouvé que leur mécanisme peut effectivement découvrir l'intelligence de marché (MI) pour aider les décideurs en établissant un système de suivi pour suivre les opinions externes sur différents aspects d'une entreprise en temps réel.

Il existe une comparaison empirique entre la SVM et les réseaux neuronaux artificiels RNA présentée par Moraes et Valiati [29] concernant l'analyse des sentiments au niveau du document. Ils ont fait cette comparaison parce que SVM a été largement et avec succès utilisé

en AS tandis que les RNA ont attiré peu d'attention en tant qu'approche pour l'apprentissage des sentiments. Ils ont discuté des exigences, des modèles résultants et des contextes dans lesquels les deux approches atteignent de meilleurs niveaux d'exactitude de classification. Ils ont également adopté un contexte d'évaluation standard avec des méthodes supervisées populaires pour la sélection et la pondération des caractéristiques dans un modèle BOW traditionnel. Leurs expériences ont indiqué que RNA a produit des résultats supérieurs à SVM, sauf dans certains contextes de données déséquilibrés. Ils ont testé trois ensembles de données de référence sur des critiques de films, GPS, appareils photo et livres d'Amazon.com. Ils ont prouvé que les expériences sur les critiques de films RNA ont surpassé SVM par une différence statistiquement significative. Ils ont confirmé certaines limites potentielles des deux modèles, qui ont été rarement discutées dans la littérature AS, comme le coût de calcul de SVM au moment de l'exécution et RNA au moment de la formation. Ils ont prouvé que l'utilisation du gain d'information (une méthode de sélection de caractéristiques peu coûteuse en termes de calcul) peut réduire l'effort de calcul de RNA et SVM sans affecter de manière significative la précision de classification résultante.

Concernant l'analyse des sentiments sur Twitter, Barbosa et Feng [3] ont fait valoir que l'utilisation de n-grammes sur les données de tweet peut nuire aux performances de classification en raison du grand nombre de mots peu fréquents sur Twitter. Au lieu de cela, ils ont proposé d'utiliser des fonctionnalités de microblogage telles que les re-tweets, les hashtags, les réponses, les signes de ponctuation et les émoticônes. Ils ont constaté que l'utilisation de ces fonctionnalités pour former les SVM améliore la précision de la classification des sentiments de 2,2% par rapport aux SVM formés uniquement à partir d'unigrammes. Une conclusion similaire a été rapportée par Kouloumpis et al. [11]. Ils ont exploré les fonctionnalités de microblogage, y compris les émoticônes, les abréviations et la présence de intensificateurs tels que les majuscules et les répétitions de caractères pour la classification des sentiments sur Twitter. Leurs résultats montrent que les meilleures performances proviennent de l'utilisation des n-grammes avec les fonctionnalités de microblogage et les fonctionnalités de lexique où les mots sont étiquetés avec leur polarité antérieure. Cependant, l'inclusion des fonctionnalités POS a entraîné une baisse des performances.

Agarwal et al. [14] a également exploré les fonctionnalités POS, les fonctionnalités de lexique et les fonctionnalités de micro-blogging. En plus de simplement combiner diverses

fonctionnalités, ils ont également conçu une représentation arborescente des tweets pour combiner de nombreuses catégories de fonctionnalités dans une représentation succincte.

Dans Ortigosa, Martín et Carro (2014), les auteurs présentent une application appelée SentBuk qui s'exécute sur Facebook et extrait des informations sur le sentiment des utilisateurs de manière automatique et non intrusive. Il classe les opinions des élèves comme positives, négatives ou neutres pour détecter des changements émotionnels importants, ils ont utilisées techniques d'apprentissage automatique deux classifier NB et SVM avec association d'une approche hybride a une approche lexicale. Les résultats obtenus ont montré une robustesse élevée de 83,27% [24].

Dans [25] ils ont comparé deux algorithmes d'apprentissage automatique de Naïve Bayes et KNN pour la classification des sentiments des critiques de films et avis sur les hôtels. Les résultats expérimentaux montrent que les classificateurs ont donné de meilleurs résultats pour le film critiques avec l'approche de Naïve Bayes donnant Précision de 80% et surperformance que l'approche de k-NN. Cependant, pour les critiques de l'hôtel, le les précisions sont beaucoup plus faibles et les deux classificateurs a donné des résultats similaires. Ainsi nous pouvons dire le classificateur Naïve Bayes peut être utilisé avec succès pour analyser le film Commentaires.

Le travail [26] a montré la création de deux classificateurs SVM, l'un pour détecter le sentiment des messages tels que les tweets et l'autre pour détecter le sentiment d'un terme dans un message. Les fonctionnalités du lexique des sentiments (créées manuellement et générées automatiquement) ainsi que fonctionnalités ngram (ngrammes de mots et de caractères) conduit au plus grand gain de performances. Les résultats obtenus ont été très bons, un F-score de 69,02% dans la tâche au niveau du message et de 88,93% dans la tâche au niveau du terme.

Les méthodes sémantiques peuvent être mélangées avec les méthodes statistiques pour effectuer une tâche SA comme le travail présenté par Zhang et Xu [28] qui ont utilisé les deux méthodes pour trouver la faiblesse du produit à partir de revues en ligne. Méthode basée sur l'identification des mots caractéristiques des avis.

1.8 Conclusion

Dans ce chapitre, nous avons présenté ce qu'est l'analyse du sentiment, avec ses trois différents niveaux, et quelques travaux et énuméré toutes les approches de classification des sentiments existantes sans oublier les domaines d'applications, et nous avons aussi parlé des problèmes qui existaient dans ce type de tâche.

La classification de texte a essentiellement progressé ces dix dernières années grâce à l'introduction des techniques héritées de l'apprentissage automatique qui ont amélioré très significativement les taux de bonne classification des sentiments. L'approche basée apprentissage supervisée donne d'excellents résultats grâce aux algorithmes sophistiqués.

L'analyse des sentiments est un domaine devient de plus en plus important car de plus en plus les gens aiment acheter en ligne et donner des commentaires, des critiques et des commentaires sur les produits ou services, où il a été utilisé dans des applications divers surtout le système de recommandation dont nous allons voir dans le deuxième chapitre.

2 Chapitre II : Système de recommandation

2.1 Introduction

Avec l'avènement des réseaux sociaux et les évolutions technologiques du web, la masse de données à exploiter ou analyser est devenue très volumineuse. Si bien qu'il est devenu difficile de savoir quelles sont les données à rechercher et où les trouver, l'un des usages les plus populaires sur ces réseaux ou le web aussi est le magasinage en ligne, où l'achat et la vente des produits et des services sont effectués électroniquement, les clients ont des difficultés à trouver des produits qui leur intéressent. Des techniques informatiques ont été développées pour faciliter cette recherche ainsi que l'extraction des informations pertinentes, par exemple la recommandation, il s'agit de guider l'utilisateur dans son exploration des données afin qu'il trouve des informations pertinentes.

Dans ce chapitre, nous fournissons un aperçu de la terminologie et des techniques liées aux systèmes de recommandation et sa relation avec l'analyse des sentiments. Avant nous commencerons par la recherche d'information et le filtrage d'information.

2.2 La recherche d'information et Filtrage d'information :

2.2.1 La recherche d'information :

La recherche d'information est basée sur un principe d'indexation des données afin de répondre aux requêtes des utilisateurs. Plus spécifiquement, la recherche documentaire consiste à interroger une base de connaissance par le biais de requêtes écrites en langues naturelles ou bien sous forme de mots clés (nommées requêtes ad hoc). Par conséquent, il peut y avoir une surcharge d'informations que l'utilisateur doit filtrer [32].

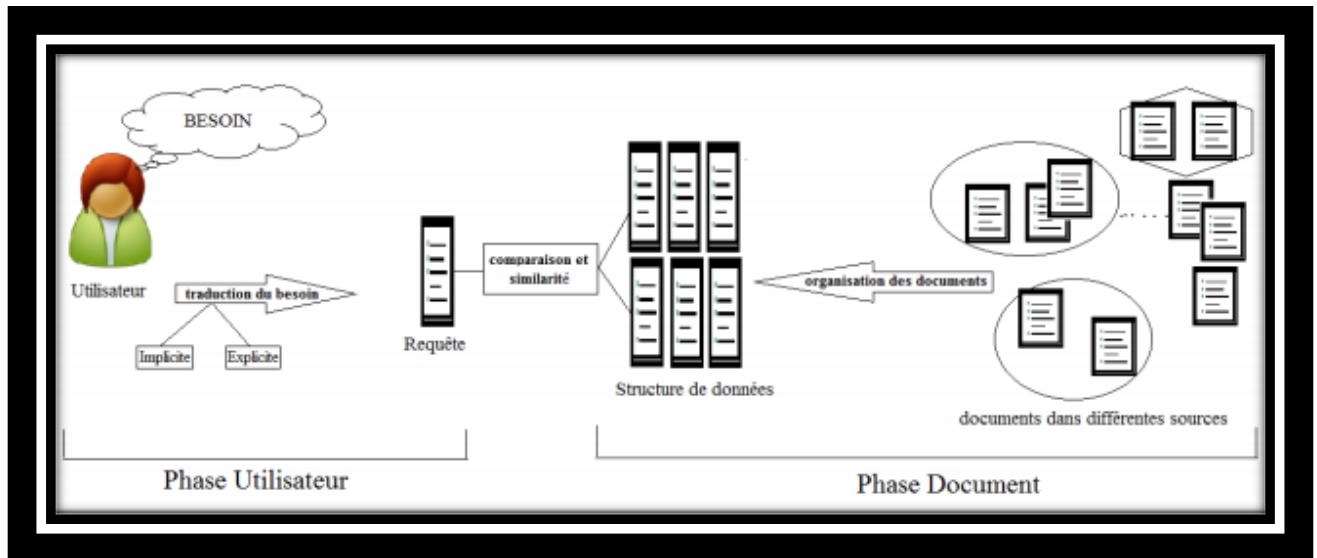


Figure 8: Le processus de recherche d'informations.

2.2.2 Le filtrage d'information :

Le filtrage d'information est l'expression utilisée pour décrire une variété de processus se rapportant à la fourniture de l'information adéquate aux personnes qui en ont besoin. Le filtrage est souvent interprété comme l'élimination de données indésirables sur un flux entrant [33].

2.3 Fonctionnalités des systèmes de recommandations :

Dans sa thèse de doctorat, (Meyer, 2012) [43] fait un découpage des systèmes de recommandation selon les fonctions qu'ils accomplissent. Il en relève 4, énoncées ci-dessous :

- **Aide à décider :** prédire une note pour un utilisateur pour un article.
- **Aide à la comparaison:** classez une liste d'articles de manière personnalisée pour un utilisateur.

- **Aide à la découverte:** fournir à un utilisateur des éléments inconnus qui seront appréciés.
- **Aide à l'exploration:** donnez des éléments similaires à un élément cible donné.

2.4 Les systèmes de recommandation :

2.4.1 Définition de Système de recommandation :

Le Systèmes de recommandation, également appelés moteur de recommandation, est une forme spécifique de filtrage de l'information visant à présenter les éléments d'information et fournissent des conseils aux utilisateurs sur les articles qu'ils pourraient souhaiter acheter ou examiner et même les services. Les produits et les services les plus recommandés sont les voyages, les restaurants, les hôtels, les pages Web, etc. Les recommandations faites par de tels systèmes peuvent aider les utilisateurs à naviguer grands espaces d'informations de descriptions de produits, d'articles de presse ou d'autres éléments. Comme l'information en ligne et le commerce électronique, les systèmes de recommandation sont des outils important. L'objectif de ces systèmes est de filtrer un flux entrant d'informations (documents) de façon personnalisé pour chaque utilisateur, tout en s'adaptant en permanence au besoin d'information de chacun. Pour cela, les moteurs de ces systèmes gèrent des profils d'utilisateurs permettant de choisir quels documents transmettre à chacun, et adaptent ces profils au cours du temps en exploitant au mieux le retour de pertinence que les utilisateurs signalent sur les documents reçus. Les SR les plus connus sont: CiteSeer, Amazon, MovieLens, etc. Actuellement, il existe deux grandes approches de filtrage: le filtrage basé sur le contenu, et le filtrage collaboratif, et de nombreux systèmes, dits « hybrides » les combinent [34].

2.4.2 Les types des systèmes de recommandation :

2.4.2.1 *Recommandation basée sur le Filtrage collaboratif :*

Le filtrage collaboratif est l'approche la plus ancienne et, aujourd'hui encore, la plus couramment utilisée. Elle se base sur les appréciations données par un ensemble d'utilisateurs

sur un ensemble d'items. Il s'agit de recommander à un usager donné les items hautement évalués dans le passé par d'autres usagers qui présentent avec lui des similarités de préférences. Par exemple, pour recommander des livres à un usager x, un système de filtrage collaboratif tentera de trouver, par différentes méthodes de corrélation, des usagers voisins (des pairs) qui ont des préférences similaires, c'est-à-dire qui notent de manière similaire les mêmes livres. Ensuite, seuls les livres les mieux notés par ses pairs seront recommandés à l'utilisateur x [35].

Voici un exemple, une matrice A : [Utilisateur x Produit] [41] :

A	Item 1	Item 2	Item 3	Item 4	Item 5
User 1	9	3	?	7	10
User 2		2	6	7	9
User 3	9	3	9	7	
User 4	6	6		2	1
User 5		7	9	3	4

Tableau 1: Utilisateur x Produit

Dans cette matrice, on trouve qu'User 1 et User 3 sont corrélés (3 sur 5 produits ont le même score). C'est ainsi qu'on recommande à l'un des utilisateurs les produits préférés du deuxième. Mais comment calculer la corrélation ou la similarité entre deux utilisateurs ?

➤ ***Cosine similarity:***

$$\mu = \cos(\text{User 1, User 2}) = \cos(\alpha) = \frac{\text{User1} \cdot \text{User 2}}{\|\text{User 1}\| \|\text{User 2}\|}$$

➤ ***Jaccard similarity:***

$$J(\text{User 1}, \text{User 2}) = \frac{|\text{User 1} \cap \text{User 2}|}{|\text{User 1} \cup \text{User 2}|}$$

Pour déterminer quels sont les voisins les plus pertinents à sélectionner, on utilise généralement l'algorithme du **k-nearest neighbor (k-NN)** qui permet de sélectionner seulement les k meilleurs voisins ayant la plus haute valeur de corrélation.

On classe par ailleurs souvent les algorithmes de filtrage collaboratif selon la technique utilisée :

Ceux à base de [35]:

- ***Mémoire ou heuristique (memory-based / heuristic-based) :***

Les algorithmes à base de mémoire font leurs prédictions en se basant sur la totalité des évaluations des usagers qui sont disponibles. Autrement dit, la valeur de la note inconnue est calculée comme un agrégat des notes de certains autres usagers jugés similaires.

- ***Basés sur les modèles (model-based) :***

Les algorithmes à base de modèle, ils sont eux aussi basés sur les évaluations disponibles des usagers, mais on essaie d'abord de classer les usagers par groupes ou d'apprendre les modèles à partir de leurs données. Une fois les groupes ou les modèles d'usagers trouvés, la prédiction pour un usager donné sera générée automatiquement à partir de son profil. On utilise pour ce faire des méthodes probabilistes (modèles de grappe ou réseaux bayésiens) ou des techniques d'apprentissage machine (réseaux neuronaux artificiels).

De tels systèmes ne tentent pas d'analyser ou de comprendre le contenu des éléments à recommander. La méthode consiste à faire des prévisions automatiques sur les intérêts d'un utilisateur en collectant des avis de nombreux utilisateurs. L'hypothèse sous-jacente de cette approche est que ceux qui ont aimé un élément spécifique dans le passé auront tendance à aimer cet élément spécifique, ou un autre très « proche », à nouveau dans l'avenir[36], le

Le système a besoin de collecter des données sur les utilisateurs et leurs préférences, cette collecte peut se faire de deux façons [37]:

- a) **Collecte Explicite** : Dans ce cas, les utilisateurs sont sollicités pour émettre leurs avis sur des produits/objets. Ils peuvent le faire via un système de notation (ex. une grille de 5 étoiles, un questionnaire de satisfaction), ou bien en publiant leurs avis sur un élément donné (ex. La fonction "J'aime" sur le réseau social Twitter permet aux utilisateurs d'exprimer leur intérêt pour un élément donné).
- b) **Collecte Implicite** : La collecte implicite s'intéresse aux interactions des utilisateurs sur le système. Les exemples de cette collecte incluent la surveillance du nombre de visites sur une page, le nombre de vues sur une vidéo, le temps passé sur une section donnée ou de l'historique des achats sur une plateforme de e-commerce.

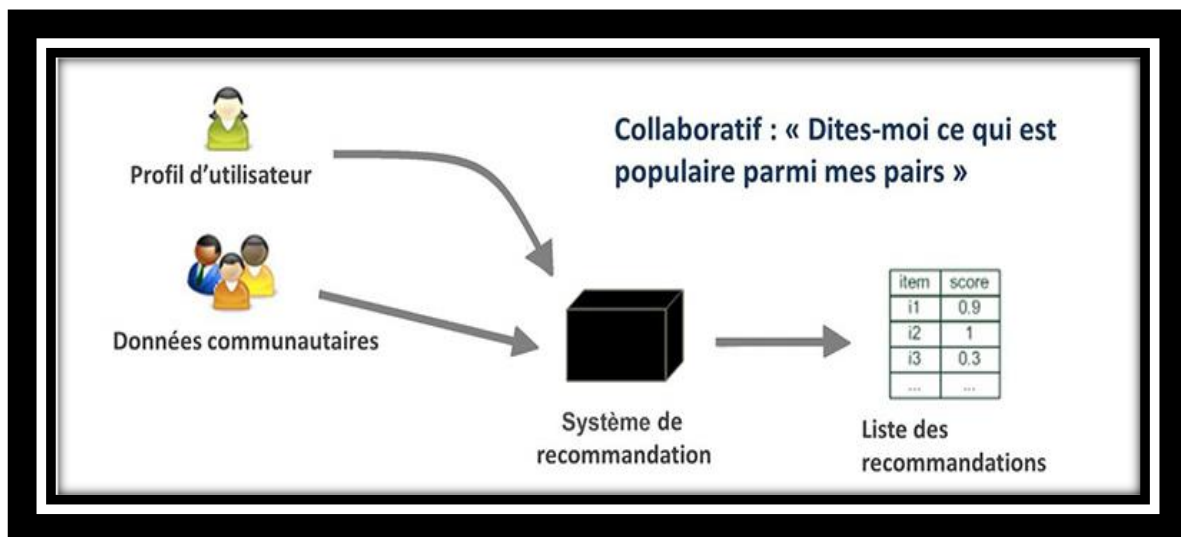


Figure 9: L'architecture d'un système de recommandation [36].

Les systèmes de recommandation collaboratifs ont comme **avantages** [36]:

- ✓ D'utiliser les scores d'autres utilisateurs pour évaluer l'utilité des éléments.
- ✓ De trouver des utilisateurs ou groupes d'utilisateurs dont les intérêts correspondent à l'utilisateur courant.

- ✓ Et plus il y a d'utilisateurs plus il y a de scores : meilleurs sont alors les résultats.

Cependant, de tels systèmes ont aussi des **inconvénients** [36] :

- ✓ Trouver des utilisateurs ou groupes d'utilisateurs similaires est difficile.
- ✓ Le système de recommandation se heurte à la faible densité de la matrice Utilisateurs X Éléments.
- ✓ De plus, il existe aussi le problème du démarrage à froid (cold-start problem) : lorsqu'un nouvel utilisateur utilise le système, ses préférences ne sont pas connues et lorsqu'un nouvel élément est ajouté au catalogue, personne ne lui a attribué de score.
- ✓ Dans les systèmes avec un grand nombre d'éléments et d'utilisateurs, le calcul croît linéairement ; des algorithmes appropriés sont donc nécessaires.
- ✓ La « non-diversité » : il n'est pas utile de recommander tous les films avec l'acteur Antonio Banderas à un utilisateur qui a aimé l'un d'eux dans le passé.

2.4.2.2 Recommandations basées sur le filtrage par contenu :

Les systèmes de filtrage basés sur le contenu [35] recommandent des items en se basant sur la description ou le contenu des items. Plutôt que d'utiliser une corrélation usager-items, on utilise une corrélation item-item. Le processus doit donc commencer par une collecte d'informations sur les items (par exemple, le titre, genre, auteur, mots-clés, etc., d'un livre). La plupart des systèmes utilisent des techniques d'extraction ou d'indexation automatique de l'information pour colliger les données. Ce type de système va ensuite extraire un certain nombre de caractéristiques propres à un item, afin de pouvoir recommander à l'utilisateur de nouveaux items possédant des propriétés similaires. Ainsi, un tel système va tenter de comprendre les similitudes entre les livres auxquels un usager a octroyé de fortes notes dans le passé (auteurs spécifiques, genre, sujet, etc.). Ensuite, seuls les livres ayant un fort degré de similitude avec les livres précédemment bien notés lui seront recommandés. De nombreux systèmes basés sur le contenu sont utilisés pour la

recommandation d'items de type textuel (documents, articles de journaux, pages web, etc.). Mais rapidement, l'utilisation de profils d'utilisateurs a permis d'améliorer les premières approches, les informations permettant de raffiner les préférences des usagers pouvant être fournies explicitement par ceux-ci (par exemple grâce à un questionnaire) ou inférées de leur comportement de navigation ou de leurs activités transactionnelles. Le système de Pandora, par exemple, repose sur un filtrage basé sur le contenu, plus précisément les propriétés d'une chanson (Grossman, 2010 ; Poirier, 2011). Dans un premier temps, les pièces musicales sont analysées par des humains (des musiciens) afin de mettre en évidence leurs attributs. Puis le système de recommandation compare les attributs des différentes pièces musicales et suggère des pièces à écouter en fonction des similitudes. L'utilisateur peut également indiquer s'il aime ou non la pièce suggérée, ce qui permet au système d'affiner son filtre en privilégiant certains attributs en fonction de l'information reçue de l'utilisateur.

Son architecture s'articule autour de 3 modules principaux [37]:

- a) **L'analyseur de contenu** : Selon la nature des données à recommander (texte, éléments multimédia, pages Web, produits commerciaux, etc.) une étape de prétraitement est nécessaire afin de décrire les objets à recommander et d'en extraire les caractéristiques. Le module d'analyse de contenu est responsable de produire une description structurée de ces objets. Cette description va servir d'élément d'entrée aux autres modules.
- b) **Le module d'apprentissage de profils** : Ce module est responsable de l'analyse des interactions passées de l'utilisateur sur les objets du système. En utilisant des méthodes empruntées au monde de l'apprentissage, ce module construit une description des préférences des utilisateurs.
- c) **Le module de filtrage** : A partir des profils utilisateurs et des descriptions des objets à recommander, ce module construit des listes de suggestions à présenter aux utilisateurs.

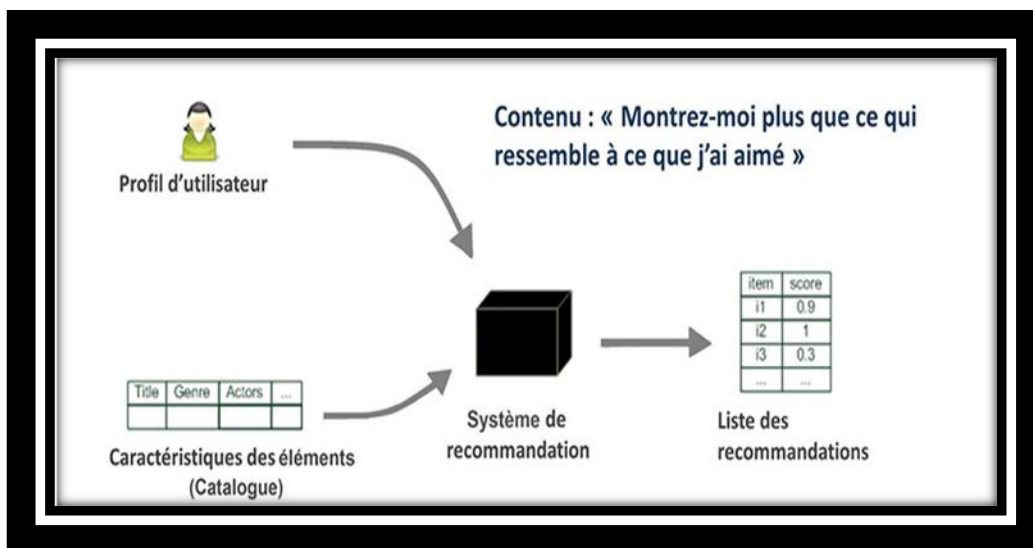


Figure 10: L'architecture d'un système de recommandation basé sur le contenu [36].

Les systèmes de recommandation basés sur le contenu présentent les **avantages** suivants [36] :

- ✓ Ils recommandent des éléments similaires à ceux que les utilisateurs ont aimés dans le passé.
- ✓ Les données relatives aux autres utilisateurs sont inutiles.
- ✓ Il est possible de recommander de nouveaux éléments ou même des éléments qui ne sont pas populaires.

Cependant, de tels systèmes ont aussi des **inconvénients** [36] :

- ✓ Tous les contenus ne peuvent pas être représentés avec des mots-clés (par exemple, les images).
- ✓ Lorsqu'un nouvel utilisateur commence à utiliser le système, il n'existe pas d'historique.

- ✓ Finalement, ces systèmes sont entièrement basés sur les scores d'éléments et les scores d'intérêt : moins il y a de scores, plus l'ensemble de recommandations possibles est limité.

2.4.2.3 *Recommandation basée sur le filtrage hybride*

Comme son nom [35] l'indique, le filtrage hybride consiste à combiner les deux méthodes précédentes, de manière à résoudre certains problèmes rencontrés par les systèmes ayant une approche unique. On peut également faire appel à des sources d'informations complémentaires pour tenter de faire des corrélations personne à personne ou classer les usagers en groupes relativement homogènes, en fonction, par exemple, de données démographiques et socio-économiques.

Différentes méthodes d'hybridation peuvent être envisagées afin de combiner les modèles (Adomavicius et Tuzhilin, 2005). On peut :

1. implanter les deux méthodes séparément.
2. incorporer quelques caractéristiques de la méthode basée sur le contenu dans l'approche collaborative.
3. incorporer quelques caractéristiques de la méthode collaborative dans l'approche basée sur le contenu.
4. construire un modèle général unifié qui incorpore les caractéristiques des deux modèles.

Ainsi le système d'Amazon est de type hybride et très complexe, puisqu'il repose sur la combinaison de trois approches (Linden, Smith et York, 2003 ; Poirier, 2011). Il se base sur le comportement individuel passé de l'utilisateur (historique de navigation et historique d'achat), mais intègre également des recommandations basées sur le contenu, soit les caractéristiques des items eux-mêmes, ainsi que des recommandations basées sur le filtrage collaboratif, soit le comportement des autres utilisateurs.

Le système de Netflix est également hybride, mais repose sur la combinaison de deux approches (Hosanagar et al., 2012) : le filtrage collaboratif (en comparant les habitudes de

visionnement de films d'utilisateurs similaires) et le filtrage basé sur le contenu (en suggérant des films qui partagent des caractéristiques avec des films que l'utilisateur a noté positivement).

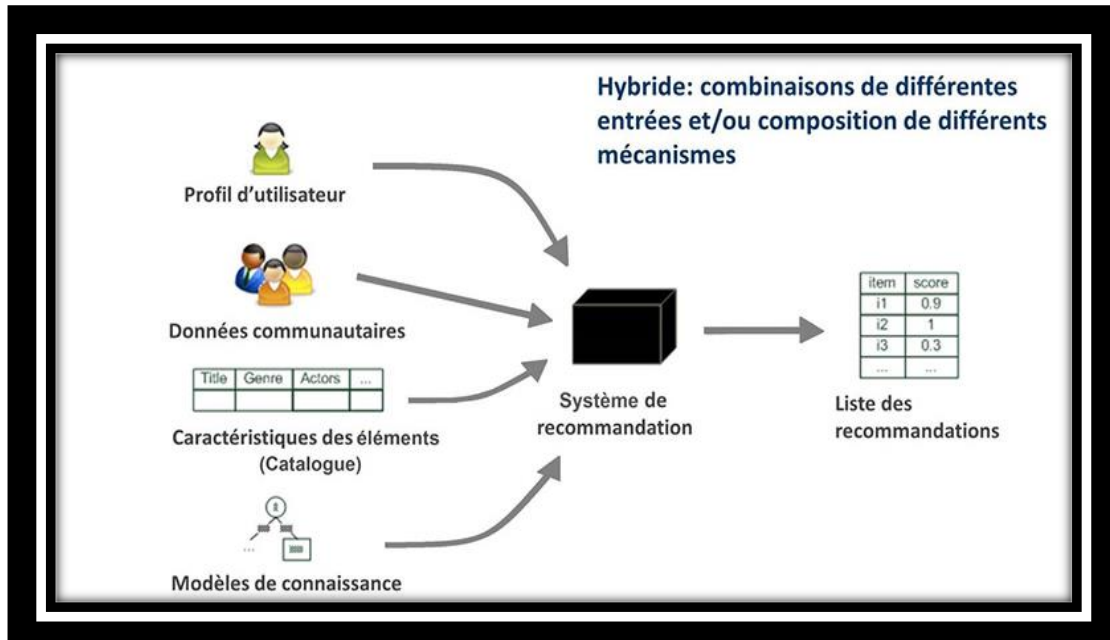


Figure 11:L'architecture d'un système de recommandation hybride [36].

2.5 La recommandation via l'analyse des sentiments :

L'analyse des sentiments porte sur les opinions et les émotions exprimées dans les textes de commentaires. Il se développe rapidement aujourd'hui en raison de la place qu'occupe le web dans notre société et du très grand nombre d'opinions exprimées quotidiennement par les utilisateurs. C'est pourquoi, un système de détection de polarité a été construit afin d'alimenter un système de recommandation en transformant les textes en une évaluation numérique pour mettre en place les filtres. Afin de pouvoir générer des recommandations précises et améliorer la qualité de ces dernières en sélectionnant les articles les plus pertinents des systèmes de recommandation [32], la détection automatique de la positivité et négativité des révisions textuelles des utilisateurs est fournie au moteur de recommandation avec une fonctionnalité implicite complémentaire, en termes de score de polarité, reflétant la satisfaction de l'utilisateur.

Des techniques d'apprentissage automatique et de traitement de texte fournissent un mécanisme automatisé pour détecter les sentiments des avis des utilisateurs et création d'un score de polarité implicite complémentaire. [39].

D'après les travaux déjà faits dans ce domaine de système de recommandation a montré que la méthode de recommandation basée sur l'extraction d'opinions a une meilleure performance que le filtrage collaboratif traditionnel basé sur la factorisation matricielle en situation de petite échelle et précision exigées [38].

Dans [38] Proposer une nouvelle approche du système de recommandation en introduisant l'analyse des sentiments basée sur les aspects dans la modélisation des utilisateurs et des produits. Les résultats expérimentaux montrent qu'il a de meilleures performances que les méthodes de recommandation courantes telles que le filtrage collaboratif. Cette approche de l'extraction d'opinion pourrait miner profondément les intérêts du client et restaurer la réputation de chaque aspect du produit grâce aux critiques qu'il obtient. Considérez les techniques pour le niveau d'aspect de l'analyse des sentiments sont encore en développement et loin d'être idéale, cette approche a beaucoup d'amélioration. Bien que l'analyse des

sentiments rende la modélisation utilisateur plus personnalisée, cette méthode montre une faiblesse dans la recherche de produits candidats à grande échelle. De manière prévisible, d'autres travaux sur hybrides recommandent des systèmes combinés à une analyse des sentiments au niveau de l'aspect et les approches de recommandation sont prometteuses.

Dans [40] ils ont proposées un système d'analyse des sentiments qui peut alimenter les notes extraites des avis des utilisateurs vers un système de recommandation. L'analyse des sentiments réussie pour prédire la satisfaction ou l'insatisfaction des utilisateurs en classant les avis en positif ou en négatif. Cette approche pourrait compenser le manque de notation des utilisateurs sur un élément dans le système de recommandation (rareté des données). Les avis des utilisateurs pourraient être plus appropriés pour évaluer les choses qui ne sont pas appropriées pour être évaluées numériquement, ce qui pourrait conduire à expliquer une recommandation spécifique. Pour les orientations futures, les chercheurs peuvent créer des systèmes de recommandation en fonction des attributs des éléments tirés des critiques.

Dans plus d'une langue comme l'arabe et l'anglais pour réduire le démarrage à froid des nouveaux éléments, cela ouvrirait également la voie à traiter le système de recommandation comme un problème de recherche d'informations en récupérant les recommandations pertinentes des avis d'utilisateurs pertinents.

2.6 Efficacité d'un algorithme de recommandation :

Evaluer l'efficacité des algorithmes de recommandation est loin d'être trivial. En premier lieu, parce que différents algorithmes peuvent être meilleurs ou moins bons en fonction du set de données sélectionné et sur lequel ils sont appliqués. D'autre part, les objectifs fixés par un système de recommandation peuvent être divers et variés. Un système de recommandation peut être mis en place pour estimer avec exactitude la note que donnerait un utilisateur à un élément, alors que d'autres auront comme objectif principal de ne pas proposer des recommandations erronées. On peut donc légitimement se demander jusqu'à quel point ces différentes méthodes de recommandation sont réellement efficaces. Pour déterminer l'efficacité d'un système, des indicateurs comme la précision et le recall sont utilisés.

La précision est un indicateur qui représente la qualité de la recommandation, c'est-à-dire à quel point les suggestions proposées sont conformes aux intérêts de l'utilisateur:

$$\text{Précision} = \text{nbre de suggestions pertinentes} / \text{nbre de suggestions}$$

Le recall indique combien de suggestions pertinentes ont été recommandées à l'utilisateur par rapport au nombre total de suggestions pertinentes disponibles dans le système.

$$\text{Recall} = \text{nbre de suggestions pertinentes proposées à l'utilisateur} / \text{nbre de suggestions pertinentes totale}$$

L'indicateur de précision permet de déterminer la probabilité qu'un élément recommandé soit pertinent.

L'indicateur de recall permet de déterminer la probabilité qu'un élément pertinent soit recommandé.

On utilise aussi des techniques statistiques pour mesurer l'efficacité d'un algorithme de recommandation. L'idée c'est d'évaluer la précision de la prédiction effectuée par le système en comparant les prédictions avec les choix qu'aurait fourni l'utilisateur dans un cas réel. On utilise par exemple l'**Erreur Moyenne Absolue (MAE)** qui mesure la déviation de la recommandation prédite avec les choix réels effectués par les utilisateurs. Plus l'erreur moyenne absolue est faible, meilleure est la prédiction.

Pour conclure, la meilleure mesure de l'efficacité d'un algorithme de recommandation et de la pertinence des suggestions c'est finalement la satisfaction de l'utilisateur, qui n'est pas toujours facile à bien identifier [42].

2.7 Conclusion

Comme présenté dans ce chapitre, les systèmes de recommandation prennent une place importante dans le domaine de la recherche d'information et deviennent rapidement un outil crucial du commerce électronique.

Dans ce chapitre, nous avons présenté la définition du système de recommandation ainsi que les trois types de ce dernier avec quelques travaux déjà fait, en mettant l'accent sur la recommandation basée sur l'analyse des sentiments qui est la plus importante dans notre travail.

3 Chapitre III : Conception

3.1 Introduction

La partie conception dans un projet informatique a une très haute importance, elle permet d'avoir une idée de ce qu'on doit programmer, et déterminer les différentes fonctionnalités de l'application.

Dans ce chapitre nous présentons l'architecture de notre système de recommandation qui se base sur l'analyse des sentiments. Afin de choisir les meilleurs algorithmes pour notre contribution, nous avons d'abord effectué plusieurs expérimentations sur des classificateurs supervisés avec plusieurs configurations, et après l'étude comparative entre les résultats obtenus nous avons pu déterminer les meilleurs entre eux et la configuration qui nous donne de bons résultats.

3.2 Architecture du système

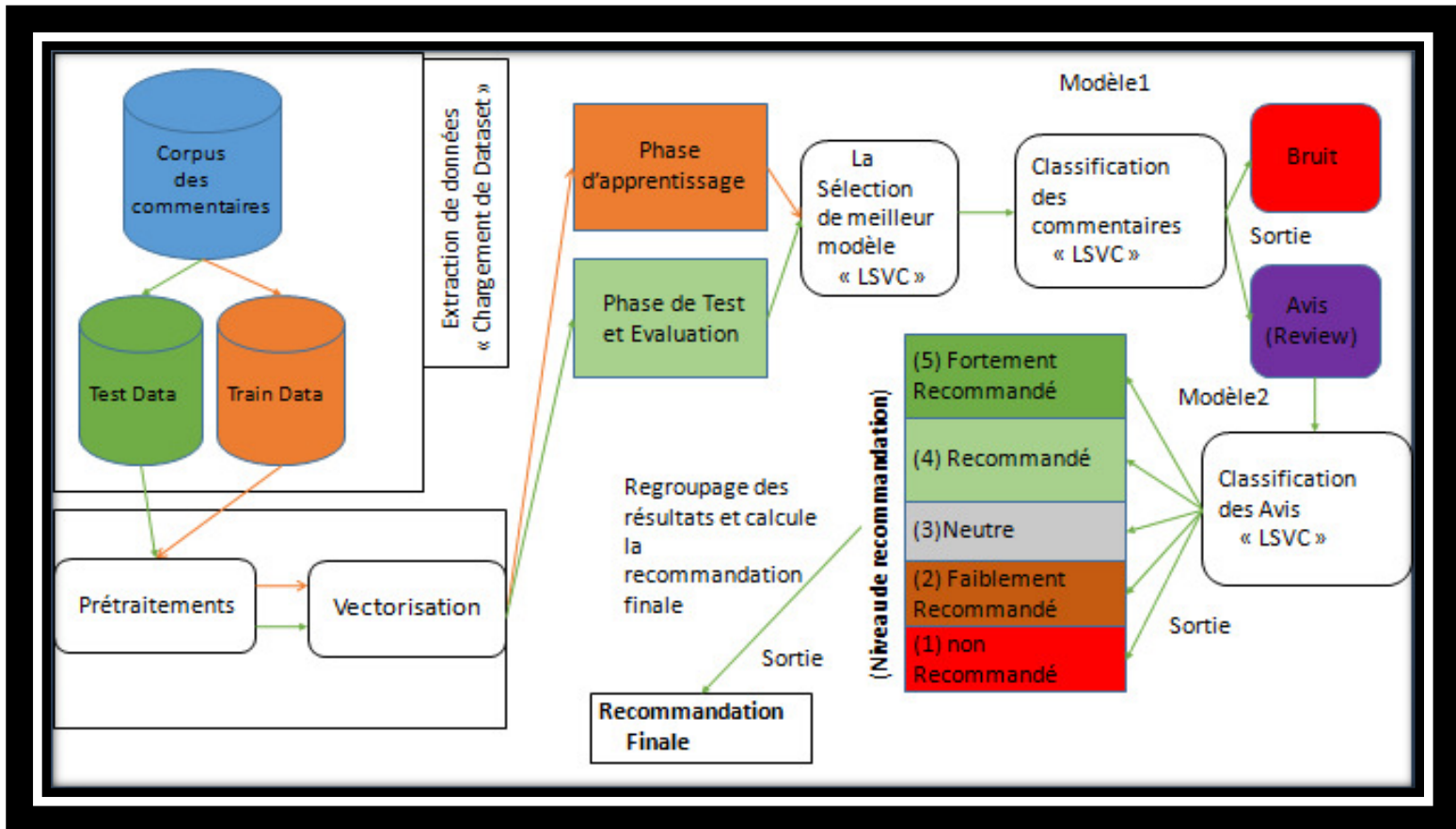


Figure 12:Architecture globale du système.

Module 1: Extraction des données « Chargement de dataset »

Module 2: Prétraitements et Vectorisation.

Module 3: La phase d'apprentissage.

Module 4: La phase de test et évaluation.

Module : La sélection de meilleur modèle « LSVC » et l'utiliser pour classer le commentaire en deux classes (Avis, Bruit), et pour obtenir l'évaluation de chaque avis, aussi pour avoir une recommandation finale d'après les évaluations des avis.

3.2.1 Les Corpus :

Nous avons téléchargé et utilisés pour nos expérimentations des corpus de textes écrits en langue Anglaise et en langue Arabe.

3.2.2 Prétraitements

Le prétraitement des données est le processus de nettoyage et de préparation des données textuelles pour la classification .Avant de procéder à une représentation vectorielle, il est indispensable de réaliser un prétraitement, et ce, pour réduire les informations indésirables. La réduction de bruit dans le texte devrait aider à améliorer les performances du classificateur et à accélérer le processus de classification [44].

Après avoir fait plusieurs expériences, nous sommes arrivés à la conclusion que ces prétraitements que nous avons choisis ont le plus d'impact et d'influence sur l'obtention d'un meilleur résultat :

➤ **Suppression de ponctuation :**

Les commentaires contiennent généralement les ponctuations. Toutes ces ponctuations seront supprimées à l'aide de la syntaxe d'expression régulière.

Exemple :

Excellent location, near many restaurants and sites. Very good service in the hotel, everyone was extremely kind and helpful. We would definitely recommend staying with you!

Figure 14: Exemple d'un commentaire avec la ponctuation.

Excellent location near many restaurants and sites Very good service in the hotel everyone was extremely kind and helpful We would definitely recommend staying with you

Figure 13: Le commentaire de la figure13 sans punctuations

➤ **Suppression des mots vides :**

Les mots vides sont les mots qui n'ajoutent pas beaucoup de sens à une phrase. Ils peuvent être ignorés en toute sécurité sans sacrifier le sens de la phrase. Par exemple, les mots comme the, he, have etc.

Excellent location near many restaurants and sites Very good service in the hotel everyone was extremely kind and helpful We would definitely recommend staying with you

Figure 15:Le commentaire de la figure14 avec les mots vides .

Excellent location near restaurants sites Very good service hotel everyone
extremely kind helpful definitely recommend staying

Figure 16: Le commentaire de la figure15 sans les mots vides.

➤ **La stemmatisation :**

La Stemmatisation est le processus de réduction d'un mot à son format racine ou radical. Prenons un exemple, considérez trois mots, «branching», «branches» et «branched». Ils peuvent être tous réduits au même mot «branch». Après tout, tous les trois transmettent la même idée de quelque chose se séparant en plusieurs chemins ou branches. Le résultat n'est pas forcément un mot existant. Cependant, celle-ci peut parfois réduire deux mots à l'orthographe proche, mais aux sens différents, à une même racine.

➤ **POS tagger (part-of-speech tagger):**

POS tagger [45] est le processus d'attribution d'une balise de partie de discours telle que nom, verbe, pronom, préposition, adverbe, adjectif ou autres balises à chaque mot d'une phrase. Il reflète la catégorie syntaxique du mot en fonction de son contexte aux fins de résolution de l'ambiguïté lexicale.

POS tagger est considéré comme l'un des outils de base nécessaires à l'analyse du langage naturel, à la recherche d'informations et à l'extraction d'opinion. C'est également l'un des principaux outils nécessaires pour développer tout corpus de langage.

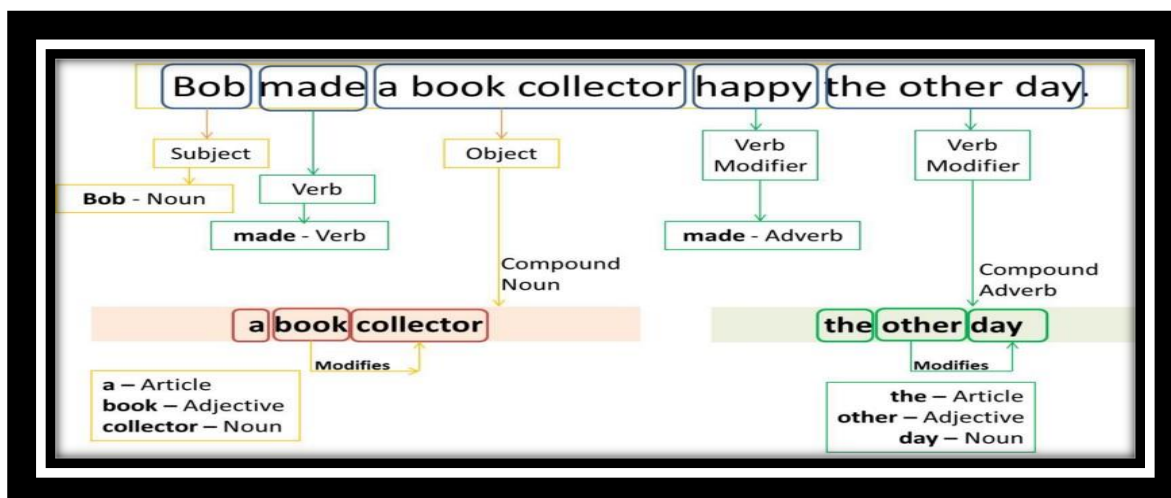


Figure 17:Exemple de Pos tag pour une phrase en anglais.

➤ La lemmatisation :

La lemmatisation [46], qui prend en considération le contexte dans lequel le mot est écrit, a pour but de trouver la forme canonique du mot. Le lemme correspond à l’infinitif des verbes et à la forme au masculin singulier des noms, adjectifs et articles. Par exemple cette méthode est capable de faire la différence entre “nous avons” : verbe avoir et “les avions” : le pluriel d’un avion. La lemmatisation tente de garder au maximum le sens des phrases.

Excellent location near restaurant site Very good service hotel everyone
extremely kind helpful definitely recommend stay

Figure 18: Le commentaire de la figure16 après la lemmatisation.

3.2.3 Vectorisation (TF-IDF) :

TF-IDF [47] est une abréviation de Term Frequency-Inverse Document Frequency et est un algorithme très courant pour transformer le texte en une représentation numérique ou vectorielle. Il est important de convertir un morceau de texte en un vecteur de caractéristiques, pour le traiter de manière efficace parce que c'est l'étape de base de toute approche de classification par apprentissage supervisée LSI, SVM...etc.

- **Fréquence de terme(TF) :**

TF est l'abréviation de l'anglais term frequency (fréquence du terme). Il détermine la fréquence relative d'un mot ou d'une combinaison de mots dans un document [49]. La formule pour déterminer le Term Frequency est la suivante [55]:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

(7)

- **Fréquence de document inverse (IDF) :**

L'**IDF** inclut dans le calcul de la fréquence des documents pour un mot précis, autrement dit l'**IDF** compare le chiffre correspondant à tous les documents connus avec le nombre de textes contenant le mot en question [49].

Pour calculer la valeur de l'Inverse Document Frequency, il faut utiliser la formule suivante [48] :

$$IDF(i) = \log \left(\frac{N_D}{f_i} + 1 \right)$$

(8)

Les différentes composantes de l'équation de l'IDF peuvent être expliquées comme suit :

i: terme dont l'inverse document frequency doit être déterminé

N_D : nombre de tous les documents dans le corpus de documents qui contient les termes pertinents.

f_i : nombre de tous les documents dans lesquels le terme "i" apparaît.

● **Terme Fréquence-Fréquence Document Inverse (TF-IDF) :**

$$W_{i,j} = TF_{i,j} * IDF_i$$

i: terme

(9)

j: document

3.2.4 La phase d'apprentissage

Cette phase est responsable de l'apprentissage des classificateurs. Les résultats de cette phase sont des classificateurs (modèles) entraînés qui sont prêts pour les tests et la Prédiction. Il couvre 80% des données originales.

• Modèles entraînés

K-Nearest Neighbor, Decision Tree, Linear Support Vector Classifier , Logistic Regression, Multinomial Naïve Bayes, Bernoulli Naïve Bayes , Random Forest et Gradient Boosting sont les 8 classificateurs que nous avons utilisé pour faire nos expériences que sont déjà illustrés précédemment dans le premier chapitre.

La résolution des problèmes de la classification à grande échelle est cruciale dans de nombreuses applications telles que la classification de texte, ils existent plusieurs algorithmes ou chaque algorithmes à son efficacité et son rôle donc on ne peut pas comparer entre eux ça dépend les jeux de données ainsi que le domaine et la nature du problème.

I. KNN

K-Nearest Neighbor (KNN) est un algorithme d'apprentissage paresseux qui stocke tous les instances correspondent à des points de données d'entraînement dans un espace à n dimensions. Quand un inconnu des données discrètes sont reçues, il analyse le nombre k le plus proche d'instances sauvegardées (le plus proche voisins) et renvoie la classe la plus courante comme prédiction et pour les données à valeur réelle, il renvoie la moyenne des k voisins les plus proches.

KNN Pseudo Code :

1. *Load the data;*
2. *Initialize the value of k ;*
3. *For getting the predicted class, iterate from 1 to total number of training data points;*
4. *Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it is the most popular method. The other metrics that can be used are Chebyshev, cosine, etc.*
5. *Sort the calculated distances in ascending order based on distance values;*
6. *Get top k rows from the sorted array;*
7. *Get the most frequent class of these rows;*
8. *Return the predicted class.*

II. Arbre de décision

Un arbre de décision (DT) est un algorithme d'apprentissage automatique supervisé qui ressemble à un arbre inversé, où dans chaque nœud représente une variable prédictive (caractéristique), le lien entre les nœuds représente une décision et chaque nœud feuille représente un résultat (variable de réponse)

DT Pseudo Code :

Step 1: Select the feature (predictor variable) that best classifies the data set into the desired classes and assign that feature to the root node.

Step 2: Traverse down from the root node, whilst making relevant decisions at each internal node such that each internal node best classifies the data.

Step 3: Route back to step 1 and repeat until you assign a class to the input data.

III. Classificateur de vecteur de support linéaire :

Dans le classificateur linéaire, un point de données est considéré comme un vecteur p-dimensionnel (liste de nombres de nombres) et nous séparons les points en utilisant un hyperplan dimensionnel (p-1). Il peut y avoir de nombreux hyperplans séparant les données dans un ordre linéaire, mais le meilleur hyperplan est de considérer être celui qui maximise la marge, c'est-à-dire la distance entre l'hyperplan et point de données le plus proche de l'une ou l'autre classe. L'hyperplan à marge maximale est déterminé par les points de données qui se trouvent plus proche de lui. Puisque nous devons maximiser la distance entre l'hyperplan et les données points. Ces points de données qui influencent notre hyperplan sont appelés support vecteurs.

IV. La Régression logistique :

Un algorithme de régression logistique (LR) est un algorithme de régression d'apprentissage automatique qui mesure la manière dont un ensemble de données se conforme à deux variables particulières. L'algorithme dicte les variables, la relation et la manière dont les variables interagissent.

Nous utilisons la fonction / courbe sigmoïde pour prédire la valeur catégorielle. La valeur seuil décide du résultat (gagnant / perdant) . Équation de régression linéaire:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

- y représente la variable dépendante à prédire.
- β_0 est l'ordonnée à l'origine, qui est essentiellement le point sur la ligne qui touche l'axe des y.
- β_1 est la pente de la droite (la pente peut être négative ou positive selon la relation entre la variable dépendante et la variable indépendante.)
- X représente ici la variable indépendante qui est utilisée pour prédire notre résultante valeur dépendante.

Fonction sigmoïde: $p = \frac{1}{1 + e^{-y}}$

V. Multinomial Naïve Bayes :

L'algorithme multinomial Naive Bayes (MNB) est utile pour modéliser les vecteurs de caractéristiques où chaque valeur représente le nombre d'occurrences d'un terme ou de son relatif la fréquence. Par exemple, si un vecteur de caractéristiques a n éléments et que chacun d'eux peut supposer k valeurs différentes avec probabilité p_k , alors:

$$P(X_1 = x_1 \cap X_2 = x_2 \cap \dots \cap X_k = x_k) = n! \prod_i x_i! \prod_i p_i^{x_i}$$

Les probabilités conditionnelles $P(x_i | y)$ sont calculées avec un comptage de fréquence. Le comptage de fréquence correspond à l'application d'une approche du maximum de vraisemblance. Pendant La formule de Bayes multinomiale, le facteur de lissage de Laplace est à garder à l'esprit. Sa valeur par défaut valeur est de 1,0 et empêche le modèle de définir des probabilités nulles lorsque la fréquence vaut zéro.

VI. Bernoulli Naïve Bayes :

Bernoulli Naive Bayes (BNB) fait partie de la famille des Naive Bayes. C'est seulement prend des valeurs binaires. L'exemple le plus général est celui où nous vérifions si chaque valeur être si oui ou non un mot qui apparaît dans un document. C'est un modèle très simplifié.

Dans les cas où le comptage de la fréquence des mots est moins important, Bernoulli peut donner de meilleurs résultats. En termes simples, nous devons compter chaque occurrence de terme binaire de valeur caractéristiques c'est-à-dire qu'un mot apparaît ou non dans un document. Ces fonctionnalités sont utilisées plutôt que trouver la fréquence d'un mot dans le document.

VII. Random Forest :

L'algorithme de forêt aléatoire (RF) est une classification et une régression supervisées algorithme. Comme son nom l'indique, cet algorithme crée aléatoirement une forêt avec plusieurs des arbres.

RF Pseudo Code :

Step 1: Create a Bootstrapped Data Set, Bootstrapping is an estimation method used to make predictions on a data set by re-sampling it. To create a bootstrapped data set, we must randomly select samples from the original data set. A point to note here is that we can select the same sample more than once.

Step 2: Creating Decision Trees, build a Decision Tree by using the bootstrapped data set created in the previous step. Since we are making a Random Forest, we will not consider the entire data set that we created, instead we will only use a random subset of variables at each step.

Step 3: Go back to Step 1 and Repeat, Random Forest is a collection of Decision Trees. Each Decision Tree predicts the output class based on the respective predictor variables used in that tree. Finally, the outcome of all the Decision Trees in a Random Forest is recorded and the class with the majority votes is computed as the output class.

Step 4: Predicting the outcome of a new data point. Bootstrapped the data and used the aggregate from all the trees to make a decision, this process is known as Bagging.

Step 5: Evaluate the Model.

3.2.5 La Phase de Test et Evaluation

Dans cette phase le modèle doit être testé pour s'assurer de sa capacité de généralisation sur les données non utilisées dans la phase d'entraînement .Il couvre 20% des données originales.

Les performances des classificateurs sont évaluées en fonction des résultats de la comparaison des étiquettes de classe prédites avec les étiquettes réelles en utilisant les mesures de performance suivantes précision ,rappel, fl-score.

3.2.6 La détection et classification des avis :

Après avoir sélectionné les deux meilleurs modèles pour chaque tâches dans la phase de teste, nous les avons incorporé dans l'architecture de notre application. Le principe de cette application se résume ainsi:

L'utilisateur va introduire un commentaire qui sera évalué par le premier modèle afin de vérifier s'il consiste d'un avis ou bruit. Pour y arriver, le commentaire sera soumis à des opérations de prétraitements, ensuite il sera transformé en vecteur par la TF-IDF et finalement prédite par le modèle1 (LSVC). Le résultat de cette prédiction sera soit un avis ou non avis (qui ne nous fournissent aucune information sur le produit ou le service). Le commentaire prédit en tant qu'avis sera soumis à une autre tâche de classification par le classificateur LSVC(Modèle 2) pour obtenir son niveau de recommandation: (1) non-recommandé, (3) faiblement-recommandé, (3) Neutre , (4) recommandé et (5) fortement recommandé.

Ceci est pour le cas ou l'utilisateur a entré un seul commentaire. Or, si l'utilisateur a entré plusieurs avis, ils seront soumis un à un à la tache 2 et donc prédis par le modèle2 pour trouver le niveau de recommandation de chaque avis. Et à la fin une moyenne de ces niveaux va nous fournir le niveau de recommandation finale du produit ou service.

• Le classificateur LSVC:

L'objectif d'un SVC linéaire (Support Vector Classifier) est de s'adapter aux données que nous fournissons, en renvoyant un hyperplan «best fit» qui divise ou catégoriser nos données. À partir de là, après avoir obtenu l'hyperplan, nous pouvons ensuite fournir certaines fonctionnalités à notre classificateur pour voir ce qu'est la classe "prédite". Cela rend cet algorithme spécifique plutôt adapté à nos utilisations, bien que nous puissions l'utiliser dans de nombreuses situations.

3.2.7 La recommandation finale

Pour obtenir une recommandation finale d'un produit ou service précis l'utilisateur peut entrer plusieurs commentaires portant d'avis sur ce produit. Le système va attribuer à chaque avis son évaluation et calculer la moyenne des évaluations pour sortir par une recommandation finale sur un produit ou service précise (1 non-recommandé, 2 faiblement recommandé, 3 neutre, 4 recommandé, 5 fortement recommandé).

3.3 Conclusion

Dans ce chapitre, nous avons décrit l'architecture de notre système en expliquant le rôle et les étapes de chaque phase.

Nous avons utilisées les prétraitements qui nous donnent de bons résultats pour réduire le bruit dans le texte avant de passer à la classification qui est le processus de classification des textes par l'utilisation de LSVC qui possèdent des fondements mathématiques solides.

4 Chapitre IV : Implémentation et Expérimentation

4.1 Introduction

Dans ce chapitre Nous décrivons les moyens matériels et logiciels utilisés dans le développement de notre application. Ayant comme objectif d'aboutir à un produit final, exploitable par les utilisateurs. Nous allons présenter des différentes interfaces de notre application sans oublier de mentionner tous les tests majeurs qui ont été faits depuis le début avec leur évaluation et interprétation.

4.2 Environnement de développement

Dans cette partie, nous allons spécifier les outils utilisés pour développer notre application :

4.2.1 Environnement Matériel

Pour développer notre application, nous avons utilisé comme environnement matériel un ordinateur avec les caractéristiques suivantes :

➤ *PC Acer Aspire E1-510:*

- Processor : Intel(R) Pentium(R) quad Core processor N3520(up to 2.42GHz,2M Cache)
- Mémoire vive : 2Go.
- Disque dur : 500 Go.
- Taille d'écran : 15,6 pouces.
- Type de système : Windows 7 avec système d'exploitation 64 bits.

4.2.2 Environnement et langages de programmation utilisés

4.2.2.1 Langages de programmation

4.2.2.1.1 Python

Python est un langage de programmation populaire. Il a été créé par Guido van Rossum et sorti en 1991. Il est utilisé pour:

- ✓ Développement web (côté serveur).
- ✓ Développement de logiciels.
- ✓ Mathématiques.
- ✓ Script système.

Python fonctionne sur différentes plateformes (Windows, Mac, Linux, Raspberry Pi, etc.), a une syntaxe qui permet aux développeurs d'écrire des programmes avec moins de lignes que certains autres langages de programmation [50].

❖ Modules externe utilisé

Nous avons utilisé des différentes packages et bibliothèques comme :

- ***Bibliothèque NLTK***

Le *Natural Language Toolkit* :est une bibliothèque open source pour le langage de programmation Python initialement écrit par Steven Bird, Edward Loper et Ewan Klein pour une utilisation dans le développement et l'éducation. Il est utilisé aussi pour créer des programmes Python qui fonctionne avec des données de langage humain pour des applications dans le traitement automatique du langage naturel (NLP) [51].

• *Scikit-learn* :

Scikit-learn est une bibliothèque en Python qui fournit de nombreux algorithmes d'apprentissage non supervisés et supervisés. Il repose sur certaines des technologies, comme NumPy, pandas et Matplotlib. Les fonctionnalités fournies par scikit-learn incluent [52]:

- Régression, y compris la régression linéaire et logistique.
- Classification, y compris K-Nearest Neighbours.
- Clustering, y compris K-Means et K-Means ++.
- Sélection de modèle.
- Prétraitement, y compris la normalisation Min-Max.

• *Joblib* :

Joblib est un ensemble d'outils pour fournir un pipelining léger en Python. En particulier: mise en cache transparente sur disque des fonctions et réévaluation paresseuse (modèle de mémorisation), et calcul parallèle simple et facile [53].

• *Numpy* :

NumPy est une bibliothèque open source python a été créé en 2005 par Travis Oliphant utilisée pour travailler avec des tableaux. Il a également des fonctions pour travailler dans le domaine de l'algèbre linéaire notamment le calcul matriciel.

• *Pandas* :

pandas est un package Python qui fournit des structures de données rapides, flexibles et expressives conçues pour rendre le travail avec des données structurées et chronologiques à

la fois faciles et intuitives. Il vise à être le bloc de construction fondamental de haut niveau pour effectuer une analyse de données pratique et réelle en Python.

4.2.2.1.2 Flask

Flask est un Framework d'application Web. Il est conçu pour rendre la mise en route rapide et facile, avec la possibilité de s'adapter à des applications complexes. Il a commencé comme un simple wrapper autour de Werkzeug et Jinja et est devenu l'un des frameworks d'application Web Python les plus populaires [54]. Il est utilisé pour faire un lien entre le code python et l'interface de notre application.

4.2.2.2 Environnement

- **Spyder**

Spyder est un environnement scientifique puissant écrit en Python, pour Python, Il présente une combinaison unique des fonctionnalités avancées d'édition, d'analyse, de débogage et de profilage d'un outil de développement complet avec l'exploration de données, l'exécution interactive, l'inspection approfondie et les superbes capacités de visualisation d'un package scientifique.

- **Editeur de texte**

Sublime Text est un éditeur de texte générique (en anglais) codé en C ++ et Python, disponible sur Windows, Mac et Linux. Pour rappel. Sublime Texte prend en charge plus de 44 langages de programmation majeurs.

- **Google Colab**

Google Colaboratory ou Colab, un outil Google simple et gratuit qui permet d'améliorer vos compétences de codage en langage de programmation Python, et d'utiliser un environnement de développement (Jupyter Notebook) qui ne nécessite aucune configuration. Il a été utilisé pour exécuter les commandes d'entraînement.

4.3 Réalisation du projet

Dans ce qui suit, nous allons présenter quelques interfaces de l'application :

4.3.1 Description de page d'accueil

Toute personne souhaitant l'accès au système doit passer par cette page qui comprend :

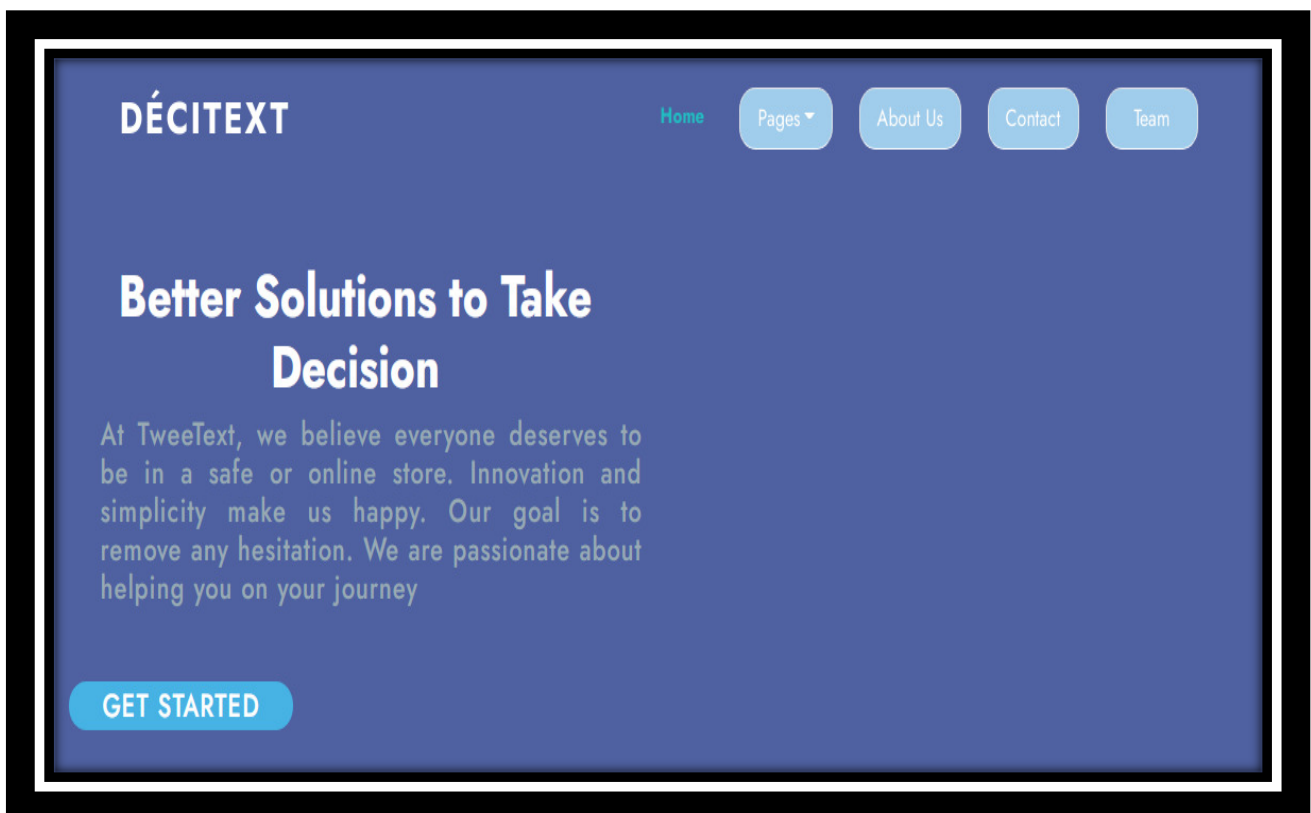


Figure 19:Page d'accueil partie 1.

Cette figure permet aux utilisateurs d'accéder aux autres pages de l'application soit la page de la langue Arabe ou la page de la langue Anglaise en cliquant sur le bouton Pages et choisir la langue.

4.3.2 Description de page de la langue anglaise

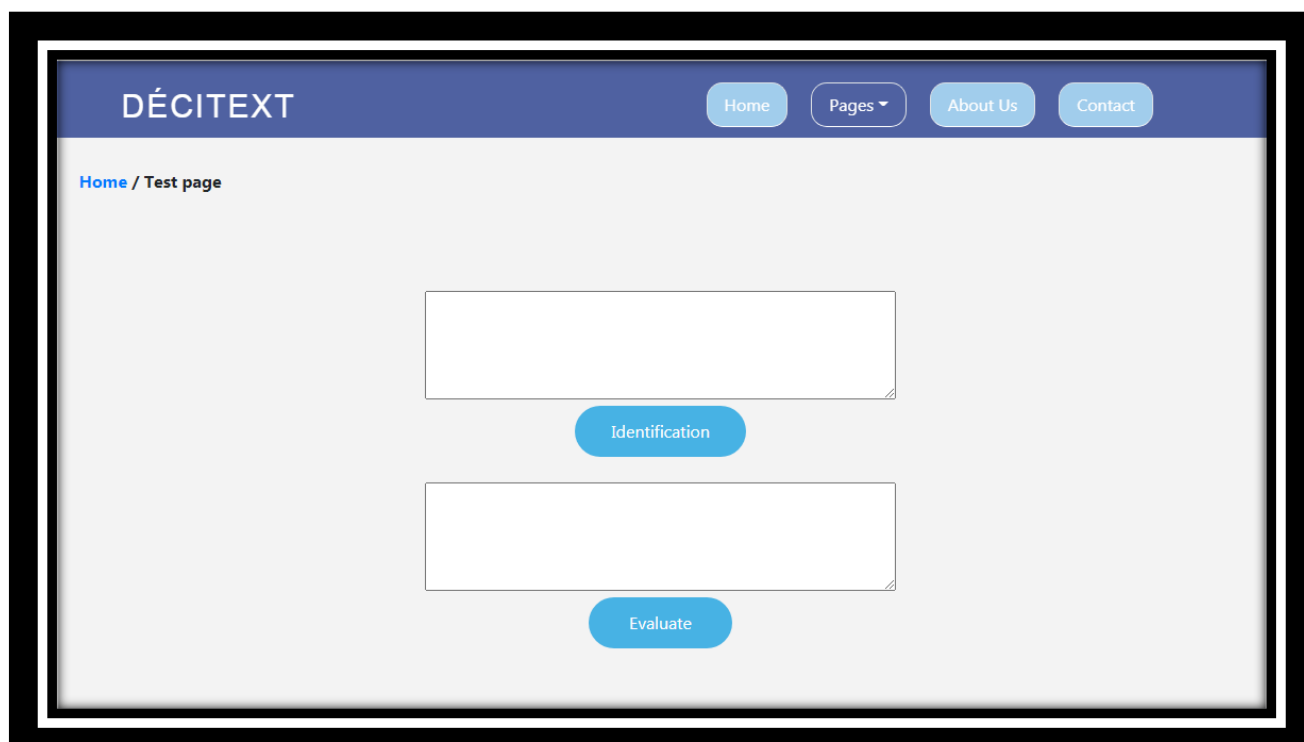


Figure 20:Interface de la langue anglaise prtie1.

Cette figure permet aux utilisateurs d'identifier les commentaires portant d'avis de ceux non (bruit) et de savoir l'évaluation d'un seul avis (un par un) des produits.

Nous avons pris des exemples des commentaires sur des hôtels (Figure 28), et nous avons utilisé des boutons qui nous permettent d'exécuter diverses fonctions :

Identification : qui identifier les commentaires de chaque hôtel soit c'est des avis ou c'est des textes bruit

Evaluate : pour avoir l'évaluation de chaque commentaire portant d'avis.

Hotels	COMMENTS
	Wonderful staff!
	I have stayed at many Ramada's over the years. This location was very run down. Carpet was soiled and dated. Had to call front desk for a TV remote and then get it ourselves.
Hotel Phillips	Place needs work. Staff was not friendly other then head housekeeping and Paul the bartender . Bed was comfortable. For having not much in the area the prices were high .
	Made trip in late December. Very satisfied with overall experience. Room and service were pretty good. Easy to connect Wifi. Parking is nearby. Oceanfront room with nice .
	If you want a new vine hop over there and like my most recent post :)
	Staff was rude, hotel is dirty, no hot water, questionable neighborhood. Was one of the worst hotels I have visited. The low rate is not low enough. Would not recommend .
W Boston	Looking for something fun with your family and friends in Philadelphia? Click below! #visitphilly https://t.co/yVIDUynFJ6
	Nice staff, room almost too small to turn around in. Kind of dirty feeling, bed should've been replaced five years ago. Can hear every noise next door and in the hall.
	While black people live here and make no effort to go into a field that will NEVER subside.
	I'm not wearing makeup and everyone in my class thought I was a new student _Ù°Ä_ÙÓÇ

Figure 21: Interface de la langue anglaise prtie2.

Cette partie contient des commentaires sur des hôtels nous les avons pris comme exemple pour les essayer.

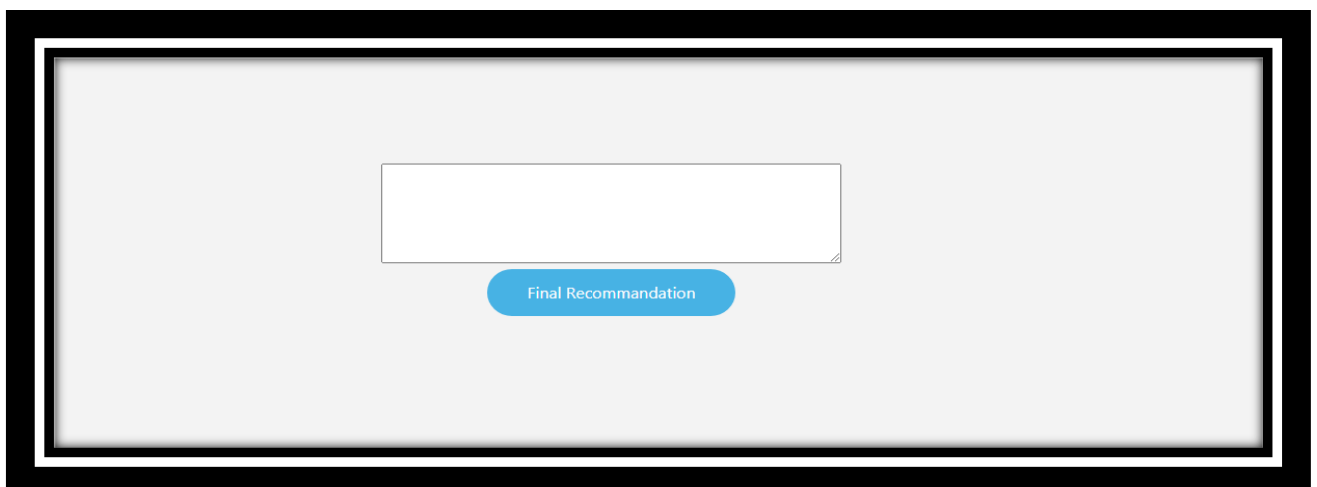


Figure 22:Interface de la langue anglaise prtie3.

Final Recommendation : pour avoir l'évaluation de chaque commentaire portant d'avis avec le calcul de la moyenne des évaluations de tous les avis d'un hôtel (niveau de recommandation finale).

4.3.3 Description de page de la langue arabe



Figure 23:Interface de la langue arabe partie 1.

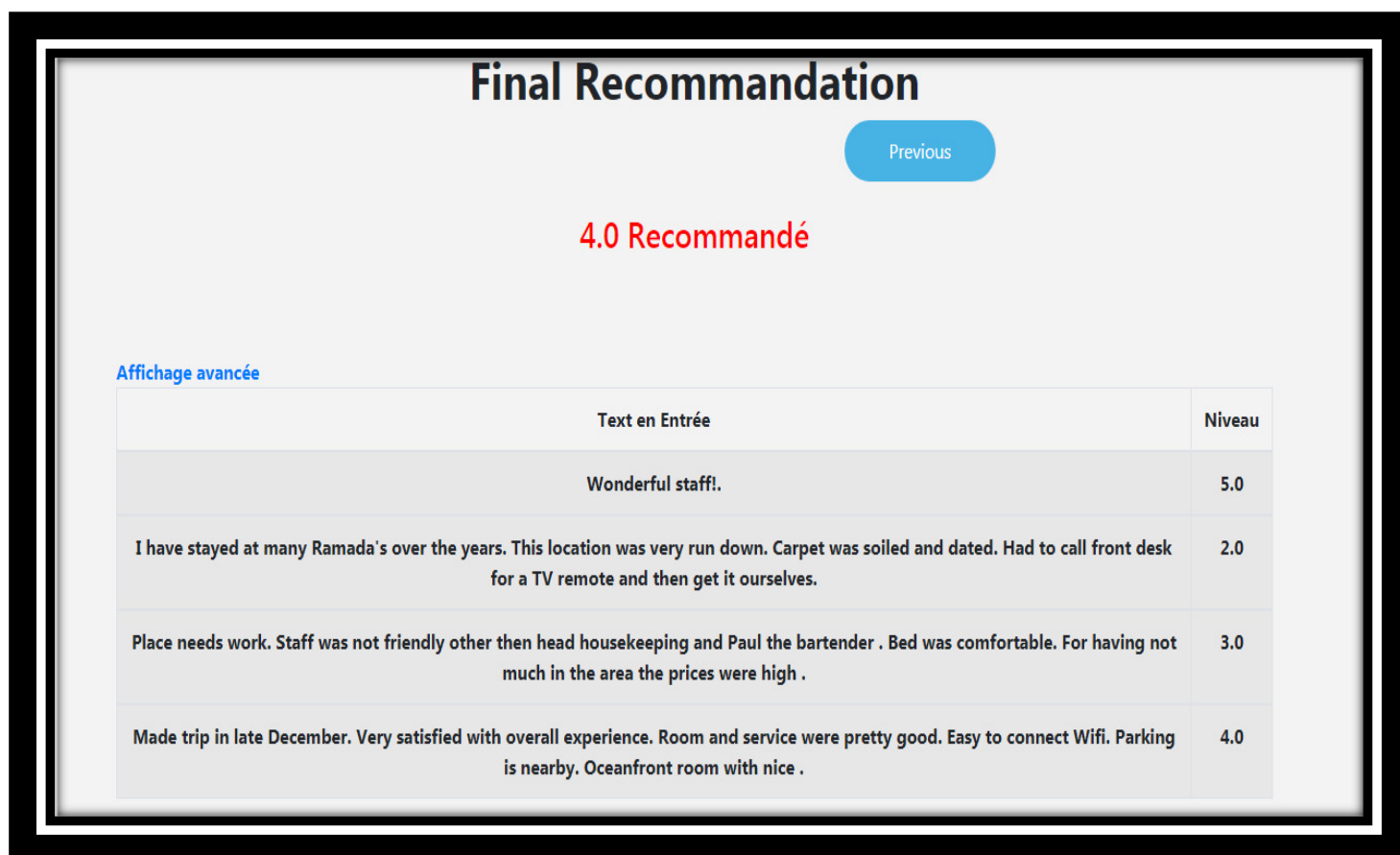
فنادق	لا تعليقات
Hotel Phillips	ضعيف جداً . من اسوأ الفنادق تماماً حجزت الفندق قبل وصولي بأسبوع غرقينمغ والذتي كبيرة بالنس غير قادره على المشياالمجموع ٥ اشخاص ... وصلت الفندق بعد العمره والانهاك بلغ منا مبلغهكونا حضرنا من مسافة ٩٠٠ كيلو دخولي الفندق السادسه فجر الجمعه وحجزني من الخميس الى السبت ، قالي من بالاستقباللا يوجد غرف اخبرته انني حاجر بوكنج فأخبرني ان الحضور يكون الساعه ١٣ ظهرأفأخبرته ان حجزني من الخميس الى السبتفرد بشكل سريع ومفاجئ ان حجزني كُنسل لأنني لم أتي بالامس علماً لم يعرف اسميولم يأخذ رقم حجري ..اخبرته ان يأخذ قيمة الامس ويسلم لي غرقتي اليومفأجابني باعتذار بارد ان الغرفه كُنسلت !!!سأنته ولو لم احضر هل ستأخذون المبلغلأن حجز غير قابل للتعديل اخبرنيانهم سيأخذونه !!!!لا اطيلهذا الفندق يسيء لحجوزات بوكنج ويفقد المصداقيه بشريحة الجمهور
	لن اكرر الحجز في الفندق . بعض العاملين في الفندق كان تعاملهم جيد. صغر الغرفه والخدمة معدومه اخدم نفسك بنفسك والحمام من نفس الغرفه"
	جيد. النظافه الهدو. المبلغ فيه شوي ارتفاع
W Boston	جيد جداً . مريح و نظيف ويوجد ألعاب أطفال للعوائل ومقهى ممتاز و غرفه واسمه. السعر يعتبر عالي"
	فندق رائع . المكان نظيف والمواقف متاحة طول الوقت. لا شيء"
	استثنائي. ممتاز الى حد كبير والمطعم رائع جداً. لا شيء
	اللي عنده مرض نفسي يقدر يحفظ القرآن حسب المرض وسل طبييا نفسيا عن المرض وأثاره د
	ممتاز . جناح النافورة. السعر"
	جيد جداً. موقعه قريب من محطة المترو، مناسب والموظفين جيدين. لا شي
	جيد. هدوء والاطلاله الجزئيه. نوعا ما قديم

Figure 24:Interface de la langue arabe partie 2.

Figure 25:Interface de la langue arabe partie 3.

La page de la langue arabe porte le même principe que la langue anglaise.

4.3.4 Description de page d'affichage des résultats



The screenshot displays a web interface titled "Final Recommendation". At the top right, there is a blue button labeled "Previous". Below the title, the text "4.0 Recommandé" is shown in red. A link labeled "Affichage avancée" is positioned above a table. The table has two columns: "Text en Entrée" and "Niveau". It contains five rows of review data.

Text en Entrée	Niveau
Wonderful staff!.	5.0
I have stayed at many Ramada's over the years. This location was very run down. Carpet was soiled and dated. Had to call front desk for a TV remote and then get it ourselves.	2.0
Place needs work. Staff was not friendly other then head housekeeping and Paul the bartender . Bed was comfortable. For having not much in the area the prices were high .	3.0
Made trip in late December. Very satisfied with overall experience. Room and service were pretty good. Easy to connect Wifi. Parking is nearby. Oceanfront room with nice .	4.0

Figure 26:La page de l'affichage des résultats.

Lorsque l'utilisateur a entré plusieurs avis d'un hôtel le système affiche un tableau des évaluations de chaque avis et la moyenne des ses évaluation qui est la recommandation finale.

4.4 Expérimentations

Cette partie représente toutes les phases des tests qui ont été effectués. Nous avons testés des différents classificateurs supervisés qui sont LSVC, LR, RF, DT, GB, MNB, BNB, KNN que sont déjà mentionnés.

4.4.1 Jeux de données

Afin de tester les classificateurs nous avons utilisées des corpus de texte en deux langues pour chaque langue nous avons combinés deux corpus l'un qui contient des avis et l'autre contient des textes bruit.

Les corpus anglais :

- (Hotel_Reviews)¹ contient 10,000 avis avec leurs évaluations, et l'autre contient des bruit (twitter_gender)².

Les corpus arabe :

- (HARDArabic-Dataset)³ contient 93700 avis d'hôtels, et l'autre contient des bruit(Arabic Conversational Dataset)⁴.

Tant que nous utilisons la méthode d'apprentissage supervisé, nous avons divisés les corpus en deux parties, 80% pour l'entraînement et 20% pour le test.

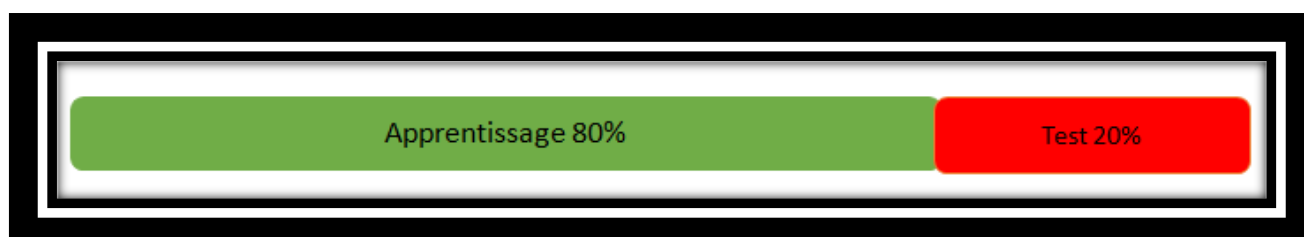


Figure 27: Découpage d'un ensemble de données en un ensemble d'apprentissage et un ensemble de test.

¹ https://www.kaggle.com/datafiniti/hotel-reviews?select=Datafiniti_Hotel_Reviews.csv.

² https://github.com/mdeff/ntds_2016/tree/master/project/reports/twitter_gender.

³ <https://github.com/elagara/HARD-Arabic-Dataset#description>.

⁴ <https://www.kaggle.com/ahmedkaramdev/arabic-conversational-dataset>.

Pour faire nos expériences, nous avons créé un script de commande, chaque commande contient deux opérations :

Opération 1 : Les prétraitements

```
['-s', '-p True', '-st True', '-lem True', '-pos True']
```

Figure 28:Opérateur de prétraitement.

Opération 2 : Analyseurs

```
['-vec word', '-vec char', '-vec char-wb', '-vec all']
```

Figure 29:Opérateur d'analyseur.

Nous avons proposés cinq stages et chacune à un nombre différent d'expériences pour nous puissions arriver au meilleur des classificateurs pour identifier et évaluer les commentaires et quelles sont les configurations que nous donneront de bons résultats. Dans chaque expérience, nous avons classé les classificateurs selon leurs f1-score, de plus grand au plus petit.

➤ Nous allons expliquer brièvement chaque stage dans le suivant :

- **Stage 1** : (Dans ce stage nous avons testés une fonctionnalité par vectoriseur):stopword, stemmatisation, lemmatisation, postagger, ...etc,avec trois analyseurs qui nous donnent 20 expériences avec 8classifiers, entre les 8 classificateurs, les deux derniers seront éliminés par f1-score.
- **Stage2** : Dans ce stage nous avons testés deux fonctionnalités par vectoriseur : stop-stemm, stop- lemm, stop-postagger,...etc, avec trois analyseurs qui nous donnent

40 expériences. Entre les 6, classificateurs, les deux derniers seront éliminés par fl-score.

- **Stage3** : Dans ce stage nous avons testés trois fonctionnalités par vectoriseur : stop-stemm-lem, stop-stemm-postagger,...etc, qui nous donnent 40 expériences .Entre les 4 classificateurs, les deux derniers seront éliminés par fl-score.
- **Stage 4** : Dans ce stage nous avons testés quatre fonctionnalités par vectoriseur : stop-stemm-lemm-postagger,...etc,avec trois analyseurs qui nous donnent 20 expériences avec 2classifiers. entre les 2 classificateurs, le dernier sera éliminé par fl-score.
- **Stage5** : Dans ce stage nous avons testés cinq fonctionnalités par vectoriseur : stop-ponct-stemm-lemm_pos,...etc, qui nous donnent 4 expériences avec le meilleur classificateur qui est le plus performant.

➤ ***Pendant nos expériences nous avons passé par plusieurs étapes :***

Étape 1 : Dans cette étape nous avons entraînés les classificateurs sur le corpus combiné (Avis+Bruit) pour étudier l'impacte d'identification des commentaires, nous avons passé par les cinq stages.

Étape 2 : Dans cette étape nous avons entraînés les classificateurs sur le corpus des avis pour étudier l'impacte d'évaluation des Avis. Nous avons passées par les mêmes stages que nous avons faits déjà avec le premier corpus.

Étape 3 : Nous avons choisir les 16 meilleures commandes, ensuite nous l'avons utilisé pour entraîné notre modèle «LSVC» sur tous le corpus (commentaires+label+évaluation) pour étudier l'impacte d'identification et d'évaluation des commentaires.

Étape 4 : Après avoir passé par presque 300 expériences nous avons pu savoir le meilleur classificateur et les meilleures configurations qui nous donnaient de bons résultats.

Nous avons donc les testés sur la langue Arabe pour voir si cela nous donneront des bons résultats et entraîné notre modèle sur la langue arabe.

4.4.2 Résultats obtenus :

4.4.2.1 Résultats obtenus de corpus anglais :

Étape 1 :

- *Stage 1 :*

Les résultats obtenus sont montrés dans le tableau 2 :

	0	1	2	3	4	5	6	7	PreProcessing	TFIDF Features	Morpho Features
Exp1	LSVC (99.600%)	RF (99.217%)	LR (99.168%)	MNB (99.001%)	GB (98.835%)	DT (98.103%)	BNB (95.274%)	KNN (57.537%)	RemovePonct	Union 5grams	-
Exp2	MNB (99.084%)	RF (98.818%)	GB (98.169%)	LSVC (96.888%)	LR (96.356%)	DT (94.692%)	KNN (93.710%)	BNB (78.552%)	RemovePonct	Union 5grams	-
Exp3	LSVC (99.450%)	RF (99.151%)	LR (98.868%)	GB (98.851%)	MNB (98.153%)	DT (97.886%)	BNB (95.757%)	KNN (56.638%)	RemovePonct	Union 5grams	-
Exp4	LSVC (99.600%)	RF (99.184%)	LR (99.168%)	MNB (99.001%)	GB (98.835%)	DT (97.737%)	BNB (95.274%)	KNN (57.537%)	-	Union 5grams	Stemmer
Exp5	MNB (99.084%)	RF (99.001%)	GB (98.186%)	LSVC (96.888%)	LR (96.356%)	DT (94.675%)	KNN (93.710%)	BNB (78.552%)	-	Union 5grams	Stemmer
Exp6	LSVC (99.450%)	RF (99.051%)	LR (98.868%)	GB (98.851%)	MNB (98.153%)	DT (98.02%)	BNB (95.757%)	KNN (56.638%)	-	Union 5grams	Stemmer
Exp7	LSVC (99.600%)	LR (99.168%)	RF (99.134%)	MNB (99.001%)	GB (98.835%)	DT (98.02%)	BNB (95.274%)	KNN (57.537%)	-	Union 5grams	PosTagger
Exp8	LSVC (99.600%)	LR (99.168%)	RF (99.118%)	MNB (99.001%)	GB (98.835%)	DT (97.720%)	BNB (95.274%)	KNN (57.537%)	-	Union 5grams	Lemmatizer
Exp9	MNB (99.084%)	RF (98.935%)	GB (98.186%)	LSVC (96.888%)	LR (96.356%)	DT (94.376%)	KNN (93.710%)	BNB (78.552%)	-	Union 5grams	PosTagger
Exp10	MNB (99.084%)	RF (98.868%)	GB (98.169%)	LSVC (96.888%)	LR (96.356%)	DT (94.625%)	KNN (93.710%)	BNB (78.552%)	-	Union 5grams	Lemmatizer
Exp11	LSVC (99.450%)	RF (99.201%)	LR (98.868%)	GB (98.851%)	MNB (98.153%)	DT (97.97%)	BNB (95.757%)	KNN (56.638%)	-	Union 5grams	PosTagger
Exp12	LSVC (99.450%)	RF (99.168%)	LR (98.868%)	GB (98.835%)	MNB (98.153%)	DT (97.953%)	BNB (95.757%)	KNN (56.638%)	-	Union 5grams	Lemmatizer
Exp13	LSVC (99.534%)	RF (99.168%)	LR (98.968%)	MNB (98.935%)	GB (98.652%)	DT (97.787%)	BNB (96.638%)	KNN (52.312%)	RemovePonct	Union 5grams	-
Exp14	LSVC (99.534%)	RF (99.234%)	LR (98.968%)	MNB (98.935%)	GB (98.685%)	DT (97.620%)	BNB (96.638%)	KNN (52.312%)	-	Union 5grams	Stemmer
Exp15	LSVC (99.534%)	LR (99.118%)	RF (99.034%)	MNB (98.619%)	GB (98.386%)	DT (97.354%)	BNB (95.673%)	KNN (70.665%)	RemoveStopWords	Union 5grams	-
Exp16	MNB (99.168%)	LR (98.086%)	GB (97.97%)	LSVC (97.903%)	RF (97.753%)	KNN (93.943%)	DT (91.098%)	BNB (72.429%)	RemoveStopWords	Union 5grams	-
Exp17	LSVC (99.384%)	RF (99.018%)	LR (98.785%)	GB (98.369%)	DT (97.986%)	MNB (97.787%)	BNB (96.322%)	KNN (68.768%)	RemovePonct	Union 5grams	-
Exp18	LSVC (99.534%)	RF (99.201%)	LR (98.968%)	MNB (98.935%)	GB (98.652%)	DT (97.853%)	BNB (96.638%)	KNN (52.312%)	-	Union 5grams	PosTagger
Exp19	LSVC (99.534%)	RF (99.134%)	LR (98.968%)	MNB (98.935%)	GB (98.685%)	DT (97.870%)	BNB (96.638%)	KNN (52.312%)	-	Union 5grams	Lemmatizer
Exp20	LSVC (99.500%)	RF (99.084%)	LR (98.951%)	MNB (98.469%)	GB (98.452%)	DT (97.304%)	BNB (96.822%)	KNN (60.815%)	RemovePonct	Union 5grams	-

Tableau 2: Les résultats obtenus de stage 1 de l'expérience 1.

D'après les résultats montrés dans le tableau 2 les deux classificateurs qui sont classés en dernier entre tous les configurations testés sont KNN et BNB, donc nous avons éliminés les classifieur KNN et BNB.

- **Stage 2 :** Les résultats obtenus sont montrés dans la Figure 30 :

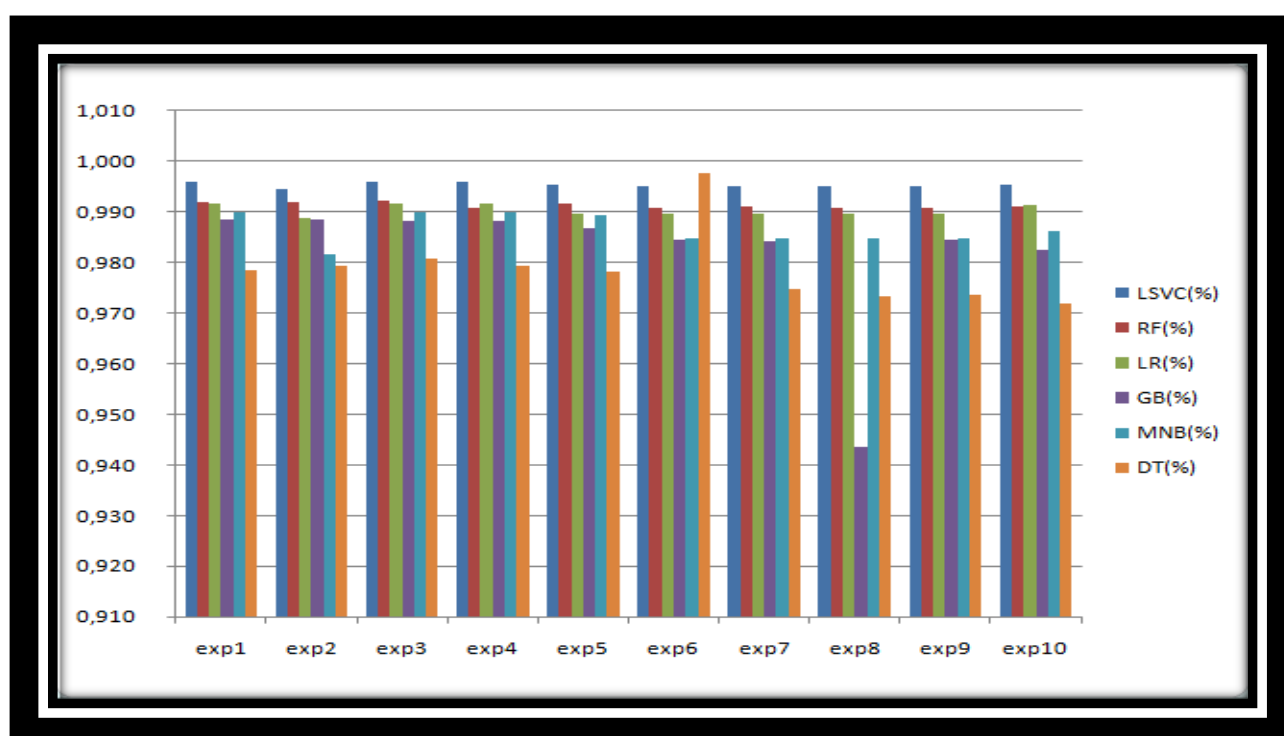


Figure 30: Les résultats obtenus de stage 2 de l'expérience 1.

D'après les résultats montrés dans le tableau les deux classifieur qui sont classés en dernier entre tous les configurations testés sont DT et GB, donc nous avons éliminés les classifieur DT et GB.

- **Stage 3** : Les résultats obtenus sont montrés dans la Figure 31 :

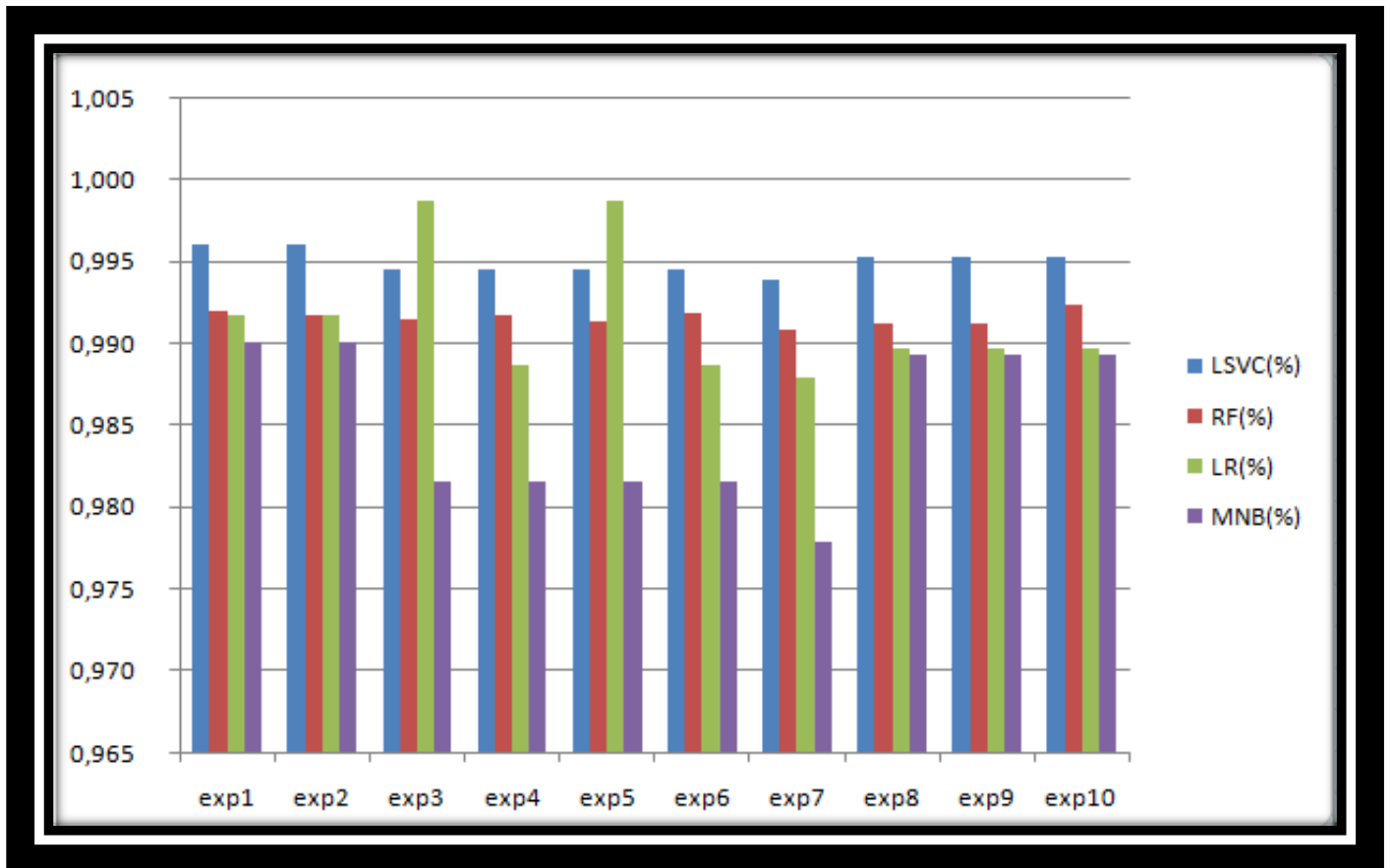


Figure 31: Les résultats obtenus de stage3 de l'expérience 1.

D'après les résultats montrés dans le tableau, les deux classificateurs qui sont classés en dernier parmi tous les configurations testées sont MNB et LR, donc nous avons éliminés les classificateurs MNB et LR.

- **Stage 4** : Les résultats obtenus sont montrés dans le tableau 3 :

	0	1	PreProcessing	TF-IDF features	Morpho Features
Exp1	LSVC (99.600%)	RF (99.184%)	RemoveStopWord	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp2	RF (98.935%)	LSVC (96.888%)	RemovePonct	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp3	LSVC (99.450%)	RF (99.184%)	RemovePonct	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp4	LSVC (99.534%)	RF (99.118%)	RemovePonct	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp5	LSVC (99.534%)	RF (99.084%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+PosTagger
Exp6	LSVC (97.903%)	RF (97.470%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+Lemmatizer
Exp7	LSVC (97.903%)	RF (97.670%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+PosTagger
Exp8	LSVC (99.384%)	RF (99.051%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+Lemmatizer
Exp9	LSVC (99.384%)	RF (99.018%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+PosTagger
Exp10	LSVC (99.534%)	RF (99.284%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+Lemmatizer
Exp11	LSVC (99.534%)	RF (99.101%)	RemovePonct	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp12	LSVC (97.903%)	RF (97.520%)	RemoveStopWords+RemovePonct	Union 5grams	PosTagger+Lemmatizer
Exp13	LSVC (99.384%)	RF (98.985%)	RemoveStopWords+RemovePonct	Union 5grams	PosTagger+Lemmatizer
Exp14	LSVC (99.534%)	RF (99.118%)	RemoveStopWords+RemovePonct	Union 5grams	PosTagger+Lemmatizer
Exp15	LSVC (97.903%)	RF (97.321%)	RemoveStopWords	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp16	LSVC (99.384%)	RF (99.034%)	RemoveStopWords	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp17	LSVC (99.500%)	RF (99.068%)	RemoveStopWords	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp18	LSVC (99.500%)	RF (99.101%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+PosTagger
Exp19	LSVC (99.500%)	RF (99.184%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+Lemmatizer
Exp20	LSVC (99.500%)	RF (99.168%)	RemoveStopWords+RemovePonct	Union 5grams	PosTagger+Lemmatizer

Tableau 3: Les résultats obtenus de stage 4 del'expérience 1.

D'après les résultats obtenus nous avons éliminé le classificateur RF.

- **Stage 5** : Les résultats obtenus sont montrés dans le tableau 4 :

	0	PreProcessing	TF-IDF features	Morpho Features
Exp1	LSVC (99.534%)	RemoveStopWords+RemovePonct	Union5grams	Stemmer+PosTagger+Lemmatizer
Exp2	LSVC (97.903%)	RemoveStopWords+RemovePonct	Union5grams	Stemmer+PosTagger+Lemmatizer
Exp3	LSVC (99.384%)	RemoveStopWords+RemovePonct	Union5grams	Stemmer+PosTagger+Lemmatizer
Exp4	LSVC (99.500%)	RemoveStopWords+RemovePonct	Union5grams	Stemmer+PosTagger+Lemmatizer

Tableau 4: Les résultats obtenus de stage5 de l'expérience 1.

Étape 2 :

- **Stage1** : Les résultats obtenus sont montrés dans le tableau 5 :

	0	1	2	3	4	5	6	7	pre-processing	TFIDF Features	Morpho Features
Exp1	LR (55.35%)	LSVC (55.25%)	GB (52.15%)	RF (48.85%)	BNB (45.65%)	MNB (44.1%)	KNN (41.6%)	DT (38.800%)	RemovePonct	Union 5grams	-
Exp2	LSVC (55.800%)	LR (52.5%)	GB (51.7%)	RF (47.9%)	KNN (47.85%)	MNB (44.2%)	BNB (44.1%)	DT (38.6%)	RemovePonct	Union 5grams	-
Exp3	LSVC (54.7%)	LR (54.35%)	GB (51.249%)	BNB (51.05%)	RF (48.199%)	MNB (44.15%)	DT (40.75%)	KNN (40.400%)	RemovePonct	Union 5grams	-
Exp4	LR (55.35%)	LSVC (55.25%)	GB (52.849%)	RF (49.1%)	BNB (45.65%)	MNB (44.1%)	KNN (41.6%)	DT (38.7%)	-	Union 5grams	Stemmer
Exp5	LSVC (55.800%)	LR (52.5%)	GB (51.800%)	KNN (47.85%)	RF (46.9%)	MNB (44.2%)	BNB (44.1%)	DT (39.45%)	-	Union 5grams	Stemmer
Exp6	LSVC (54.7%)	LR (54.35%)	GB (51.749%)	BNB (51.05%)	RF (48.75%)	MNB (44.15%)	KNN (40.400%)	DT (40.150%)	-	Union 5grams	Stemmer
Exp7	LR (55.35%)	LSVC (55.25%)	GB (52.25%)	RF (49.35%)	BNB (45.65%)	MNB (44.1%)	KNN (41.6%)	DT (38.550%)	-	Union 5grams	PosTagger
Exp8	LR (55.35%)	LSVC (55.25%)	GB (52.349%)	RF (49.2%)	BNB (45.65%)	MNB (44.1%)	KNN (41.6%)	DT (38.3%)	-	Union 5grams	Lemmatizer
Exp9	LSVC (55.800%)	LR (52.5%)	GB (52.1%)	KNN (47.85%)	RF (47.099%)	MNB (44.2%)	BNB (44.1%)	DT (38.550%)	-	Union 5grams	PosTagger
Exp10	LSVC (55.800%)	LR (52.5%)	GB (51.949%)	KNN (47.85%)	RF (46.300%)	MNB (44.2%)	BNB (44.1%)	DT (39.550%)	-	Union 5grams	Lemmatizer
Exp11	LSVC (54.7%)	LR (54.35%)	GB (51.249%)	BNB (51.05%)	RF (49.1%)	MNB (44.15%)	DT (40.949%)	KNN (40.400%)	-	Union 5grams	PosTagger
Exp12	LSVC (54.7%)	LR (54.35%)	GB (51.4%)	BNB (51.05%)	RF (49.7%)	MNB (44.15%)	KNN (40.400%)	DT (39.15%)	-	Union 5grams	Lemmatizer
Exp13	LSVC (54.1%)	LR (53.449%)	GB (52.75%)	BNB (52.15%)	RF (49.35%)	MNB (44.1%)	DT (39.050%)	KNN (38.95%)	RemovePonct	Union 5grams	-
Exp14	LSVC (54.1%)	LR (53.449%)	GB (52.800%)	BNB (52.15%)	RF (49.1%)	MNB (44.1%)	KNN (38.95%)	DT (38.0%)	-	Union 5grams	Stemmer
Exp15	LR (54.400%)	LSVC (53.849%)	GB (51.2%)	RF (48.949%)	BNB (46.2%)	MNB (44.05%)	KNN (41.6%)	DT (38.800%)	RemoveStopWords	Union 5grams	-
Exp16	LSVC (54.1%)	LR (51.800%)	GB (51.5%)	RF (48.1%)	KNN (45.550%)	MNB (44.3%)	BNB (44.05%)	DT (41.949%)	RemoveStopWords	Union 5grams	-
Exp17	LSVC (54.05%)	LR (52.7%)	GB (51.1%)	BNB (50.749%)	RF (49.1%)	MNB (44.1%)	KNN (40.400%)	DT (38.15%)	RemovePonct	Union 5grams	-
Exp18	LSVC (54.1%)	LR (53.449%)	GB (52.7%)	BNB (52.15%)	RF (49.2%)	MNB (44.1%)	KNN (38.95%)	DT (38.25%)	-	Union 5grams	PosTagger
Exp19	LSVC (54.1%)	LR (53.449%)	GB (52.849%)	BNB (52.15%)	RF (49.7%)	MNB (44.1%)	KNN (38.95%)	DT (38.35%)	-	Union 5grams	Lemmatizer
Exp20	LSVC (53.900%)	LR (53.5%)	GB (52.2%)	BNB (52.1%)	RF (50.7%)	MNB (44.1%)	KNN (40.949%)	DT (40.0%)	RemovePonct	Union 5grams	-

Tableau 5: Les résultats obtenus de stage2 de l'expérience 2.

D'après les résultats montrés dans le tableau les deux classifieur qui sont classés en dernier entre tous les configurations testés sont DT et KNN, donc nous avons éliminés les classifieur DT et KNN.

- **Stage 2** : Les résultats obtenus sont montrés dans la Figure 32 :

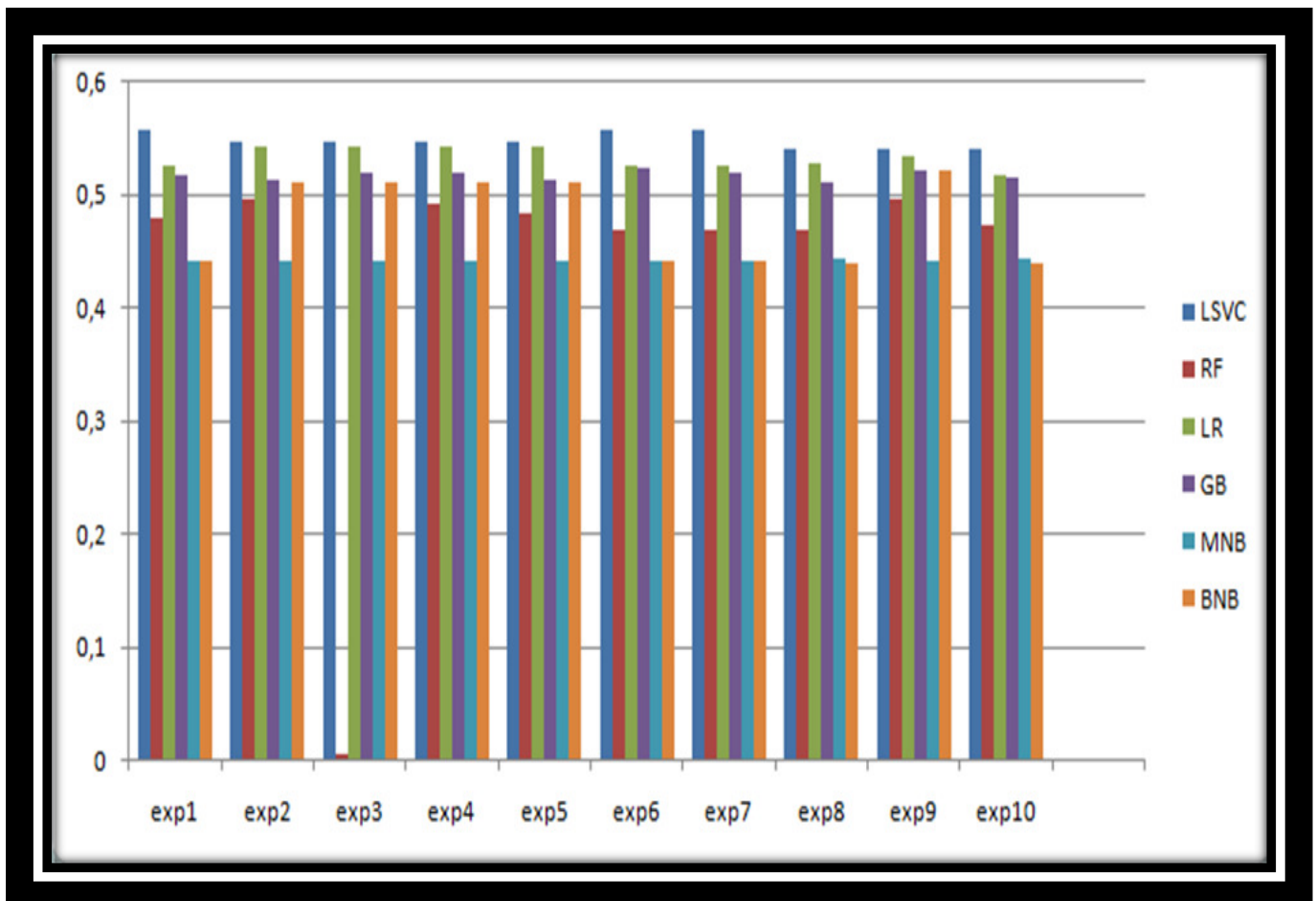


Figure 32: Les résultats obtenus de stage 2 de l'expérience 2.

D'après les résultats montrés dans le tableau les deux classificateurs qui sont classés en dernier entre tous les configurations testés sont GB et BNB, donc nous avons éliminés les classificateurs GB et BNB.

- **Stage 3** : Les résultats obtenus sont montrés dans la Figure 33 :

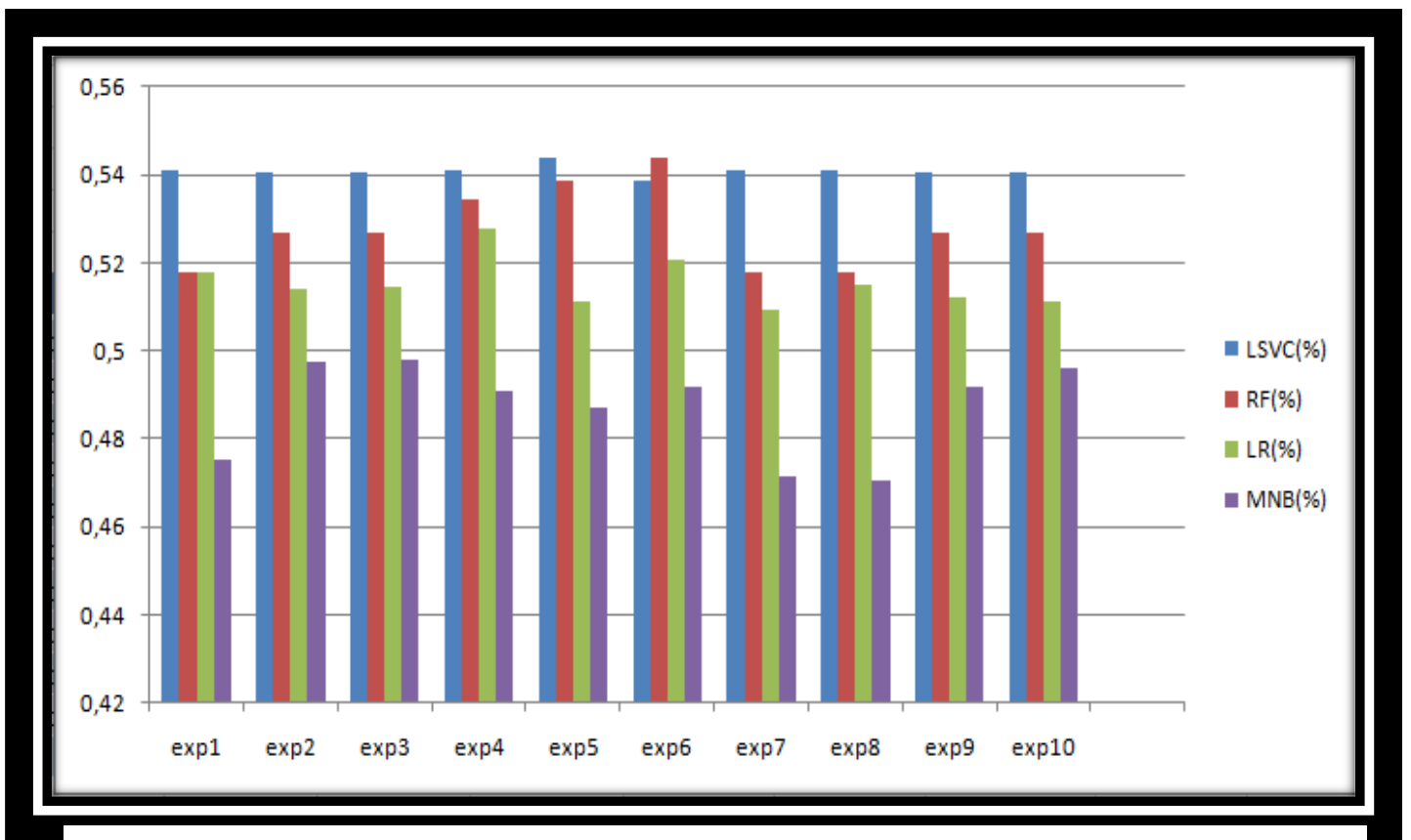


Figure 33: Les résultats obtenus de stage 3 de l'expérience 2.

D'après les résultats montrés dans le tableau les deux classificateurs qui sont classés en dernier entre tous les configurations testés sont MNB et RF, donc nous avons éliminés les classificateurs MNB et RF.

- **Stage 4** : Les résultats obtenus sont montrés dans le tableau 6 :

	0	1	PreProcessing	TF-IDF features	Morpho Features
Exp1	LR (55.35%)	LSVC (55.25%)	RemoveStopWords	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp2	LSVC (55.800%)	LR (52.5%)	RemovePonct	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp3	LSVC (54.7%)	LR (54.35%)	RemovePonct	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp4	LR (54.400%)	LSVC (53.849%)	RemovePonct	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp5	LR (54.400%)	LSVC (53.849%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+PosTagger
Exp6	LSVC (54.1%)	LR (51.800%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+Lemmatizer
Exp7	LSVC (54.1%)	LR (51.800%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+PosTagger
Exp8	LSVC (54.05%)	LR (52.7%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+Lemmatizer
Exp9	LSVC (54.05%)	LR (52.7%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+PosTagger
Exp10	LSVC (54.1%)	LR (53.449%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+Lemmatizer
Exp11	LR (54.400%)	LSVC (53.849%)	RemovePonct	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp12	LSVC (54.1%)	LR (51.800%)	RemoveStopWords+RemovePonct	Union 5grams	PosTagger+Lemmatizer
Exp13	LSVC (54.05%)	LR (52.7%)	RemoveStopWords+RemovePonct	Union 5grams	PosTagger+Lemmatizer
Exp14	LR (54.400%)	LSVC (53.849%)	RemoveStopWords+RemovePonct	Union 5grams	PosTagger+Lemmatizer
Exp15	LSVC (54.1%)	LR (51.800%)	RemoveStopWords	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp16	LSVC (54.05%)	LR (52.7%)	RemoveStopWords	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp17	LSVC (53.900%)	LR (53.5%)	RemoveStopWords	Union 5grams	Stemmer+PosTagger+Lemmatizer
Exp18	LSVC (53.900%)	LR (53.5%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+PosTagger
Exp19	LSVC (53.900%)	LR (53.5%)	RemoveStopWords+RemovePonct	Union 5grams	Stemmer+Lemmatizer
Exp20	LSVC (53.900%)	LR (53.5%)	RemoveStopWords+RemovePonct	Union 5grams	PosTagger+Lemmatizer

Tableau 6: Les résultats obtenus de stage4 de l'expérience2.

D'après les résultats montré dans le tableau, on va éliminer le classifieur LR.

- **Stage 5 :**

Les résultats que nous avons obtenus sont montrés dans le Tableau 7:

	0	PreProcessing	TF-IDF features	Morpho Features
Exp1	LSVC (53.849%)	RemoveStopWords+RemovePonct	Union5grams	Stemmer+PosTagger+Lemmatizer
Exp2	LSVC (54.1%)	RemoveStopWords+RemovePonct	Union5grams	Stemmer+PosTagger+Lemmatizer
Exp3	LSVC (54.05%)	RemoveStopWords+RemovePonct	Union5grams	Stemmer+PosTagger+Lemmatizer
Exp4	LSVC (53.900%)	RemoveStopWords+RemovePonct	Union5grams	Stemmer+PosTagger+Lemmatizer

Tableau 7: Les résultats obtenus de stage5 de l'expérience 2.

Étape 3 : Les résultats obtenus sont montrés dans le tableau 8 :

Model	Text PreProcessing	TFIDF Features	Morpho Features	Exps	Texte-Label(Exp1)	Review-Rating(Exp2)	Texte-Label-Rating(Exp4)
					F1-score	F1-score	F1-score = F1-scoreExp1*F1-scoreExp2
LSVC	-	Union 5grams	-	exp115	98 %	81 %	79 %
LSVC	RemovePunct	Union 5grams	-	exp117	98 %	83 %	81 %
LSVC	-	Union 5grams	Stemmer	exp417	99 %	84 %	83 %
LSVC	-	Union 5grams	PosTagger	exp717	99 %	84 %	83 %
LSVC	-	Union 5grams	Lemmatize	exp817	99 %	84 %	83 %
LSVC	RemovePunct	Union 5grams	Stemmer	exp119	98 %	83 %	81 %
LSVC	RemovePunct	Union 5grams	PosTagger	exp419	95 %	65 %	62 %
LSVC	RemovePunct	Union 5grams	Lemmatize	exp519	98 %	83 %	81 %
LSVC	-	Union 5grams	PosTagger+Stemmer	exp1019	99 %	84 %	83 %
LSVC	-	Union 5grams	Stemmer + Lemmatize	exp1119	99 %	84 %	83 %
LSVC	-	Union 5grams	PosTagger+Lemmatize	exp1619	98 %	82 %	80 %
LSVC	RemovePunct	Union 5grams	Stemmer + PosTagger	exp121	98 %	81 %	79 %
LSVC	RemovePunct	Union 5grams	Lemmatize+Stemmer	exp221	98 %	83 %	81 %
LSVC	RemovePunct	Union 5grams	PosTagger+Lemmatize	exp721	99 %	80 %	79 %
LSVC	-	Union 5grams	Stemmer +PosTagger+ Lemmatize	exp1021	99 %	83 %	82 %
LSVC	RemovePunct	Union 5grams	Stemmer +PosTagger+ Lemmatize	exp123	98 %	82 %	80 %

Tableau 8:Les résultats obtenus de l'expérience 3.

4.4.2.2 *Résultats obtenus de corpus Arabe*

Étape 4 :

Voici le résultat que nous avons obtenu :

Model	PreProcess	FIDF Features	erpho Features	Exps	Texte-Label(Exp1)	Review-Rating(Exp2)	Texte-Label-Rating(Exp4)
					F1-score	F1-score	F1-score = F1-scoreExp1 * F1-scoreExp2
LSVC	-	vec all	Stemmer	exp417	97 %	72 %	70 %
LSVC	-	vec all	Lemmatize	exp817	85 %	80 %	68 %
LSVC	-	vec all	Tagger+Stemmer	exp1019	92 %	68 %	63 %
LSVC	-	vec all	mer + Lemmer	exp1119	90 %	90 %	81 %

Tableau 9: Les résultats obtenus de la langue arabe.

4.4.3 Evaluation :

Nous avons utilisé le rappel, la précision et le f1 score pour évaluer la performance de nos classificateurs. Ensuite, nous avons pris le score en considération pour éliminer les pires classificateurs.

4.4.4 Discussion :

Notre approche a consisté en premier lieu à faire des comparaisons avec différents configurations entre les huit classificateurs mentionnés ci-dessus en termes de f-score.

Après avoir testé plusieurs classifieurs avec plusieurs configurations nous remarquons que LSVC est le meilleur qui donne des bons résultats pour détecter les avis dans les commentaires et les évaluer en deux langues (arabe, anglais).

Lors des tests avec le corpus anglais pour identifier les commentaires et évaluer les avis, notre modèle a également donné d'excellents résultats (99.6%) dans les tests d'identification des commentaires et (55.8%) pour dans les tests d'évaluation des avis.

Lors des tests avec le corpus arabe pour identifier les commentaires et évaluer les avis, notre modèle a également donné d'excellents résultats (97%) dans les tests d'identification des commentaires et (90%) pour dans les tests d'évaluation des avis.

Selon les tests, nous avons remarqué que les fonctionnalités (stemming, postagger, lemmatisation) ont amélioré les résultats de classificateur LSVC.

4.5 Conclusion

Dans ce dernier chapitre, nous avons entamé le processus de l'implémentation afin de réaliser notre système en déterminant l'ensemble des facteurs de développement, tout en respectant la conception élaborée. Puis nous avons présenté les interfaces de notre application.

Dans ce chapitre nous avons aussi présenté toute les tests expérimentaux qui ont été effectués sur deux dataset, pour atteindre au plus performant des classificateur pour identifier et évaluer les commentaires, cette évaluation, qui nous aide à prendre une décision envers un produit ou un service.

Conclusion générale

Notre projet intitulé « Conception d'un système de recommandation multilingue basé sur l'analyse des sentiments des opinions dans les commentaires en ligne » consiste à la conception et la réalisation d'une application web destiné aux gens qui souhaitent acheter un produit ou utiliser un service et veulent savoir s'il est bon selon ce que les autres en pensent d'après leurs expériences avec ces produits et services ce qui aide à prendre des décisions et, aussi pour prendre de bonnes décisions concernant le produit à acheter parmi le vaste choix de produits proposés.

Notre travail vise à mettre en place un système de recommandation qui se base sur l'analyse des sentiments destiné spécifiquement au couple de langues Anglaise et Arabe. La tâche de l'analyse des sentiments est encore en développement et loin d'être complète, en fait il s'agit d'un important domaine de recherche actuel qui a connu une évolution majeure depuis les années 2000 et a réalisé une évolution majeure et un grand intérêt depuis la naissance de l'apprentissage machine.

En ce qui concerne la démarche, nous avons en premier lieu effectué une phase d'étude des différents algorithmes de l'apprentissage automatique utilisées pour la classification des données textuelles comme les tweets .et afin d'améliorer le travail de notre système nous avons testés plusieurs modèle pour extraire le meilleur pour le travail que nous faisons. . En deuxième lieu nous avons spécifié notre application pour discerner les fonctionnalités .En troisième lieu, nous avons procédé à sa conception ainsi qu'aux choix technologiques pour sa réalisation. Enfin, nous l'avons mise en œuvre.

Toutes les fonctionnalités décrites ont été développées et validées. Néanmoins, notre projet pourra être amélioré par l'ajout d'autre fonctionnalités comme :

- Héberger l'application sur un serveur.
- Développer le modèle pour traiter le problème du sarcasme.

- Il devrait être possible d'utiliser notre approche dans d'autres langues.

Il est important à noter que la réalisation de ce projet nous a été bénéfique sur tous les plans. Sur le plan technique, ce projet nous a été une bonne occasion pour découvrir et maîtriser les différents outils de programmation notamment python, d'approfondir nos connaissances sur la démarche de travail chez les centres de recherche qui exige la ponctualité et l'intégration dans un groupe de travail.

REFERENCES BIBLIOGRAPHIQUE

- [1] Bo, Pang, Lillian, Lee, Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, Vol. 2, Nos. 1–2 (2008).
- [2] Bing, Liu, Sentiment analysis : mining opinions sentiments, and emotions. First edition USA 2015.
- [3]-Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of COLING, pp. 36–44 (2010).
- [4] Bing, Liu. Sentiment Analysis and Opinion mining. University of Toronto: Graeme Hirst, 2012, p 1-5.
- [5] cours, «Traitement Automatique des Langues», 2011/2012. [En ligne]. Available: <http://lalic.paris-sorbonne.fr/PAGESPERSO/atanassova/lfa/Cours10.pdf> .[Accès le 20 mai 2020].
- [6] Wikipédia. [En ligne]. Available: <https://fr.m.wikipedia.org/wiki/Sentiment>. [Accès le 20 mai 2020].
- [7] dictionnaire de francais larousse. (s.d.). [En ligne]. Available: <https://www.larousse.fr/dictionnaires/francais/opinion/56197> . [Accès le 21 mai 2020].
- [8] KOLKUR, Seema, DANTAL, Gayatri, et MAHE, Reena. Study of different levels for sentiment analysis. *International Journal of Current Engineering and Technology*, 2015, vol. 5, no 2, p. 768-770.
- [9] Federico Pascual, Twitter Sentiment Analysis with Machine Learning, June 7, 2019. [En ligne]. Available at : <https://monkeylearn.com/blog/sentiment-analysis-of-twitter/>. [Accès le 25 mai 2020].
- [10] Kharde, V., & Sonawane, S.S. Sentiment analysis of twitter data: a survey of techniques. *International Journal of Computer Applications* (0975 – 8887), April 2016, Vol.139 – No.11 .
- [11]-Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: The good the bad and the omg! In: Proceedings of the ICWSM (2011).
- [12] [En ligne]. Available at : <http://help.sentiment140.com/for-students>. [Accès le 30 mai 2020].
- [13] Walaa, Medhat, Ahmed, Hassan, Hoda. Korashy, Sentiment analysis algorithms and applications: A survey, *Aln Shams Engineering Journal* (2014) 5, p1095–1104.
- [14] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proc. ACL 2011 Workshop on Languages in Social Media, pp. 30–38 (2011) .
- [15] Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 132-149).
- [16] Patil, L. H., & Atique, M. (2015). An Improved feature selection based on neighborhood positive approximation rough set in document classification. *International Journal of Soft Computing and Software Engineering*, 5(1), 13-30.
- [17] Rohith Gandhi, Naive Bayes Classifier, May 5, 2018
- [18] Rakotomalala, R. (2005). Arbres de décision. *Revue Modulad*, 33, 163-187.
- [19] Rivals, I., Personnaz, L., Dreyfus, G., & Ploix, J. L. (1995). Modélisation, classification et commande par réseaux de neurones: principes fondamentaux, méthodologie de

conception et illustrations industrielles. *Les réseaux de neurones pour la modélisation et la commande de procédés*, JP Corriou, ed.(Lavoisier Tec & Doc, 1995).

[20] Avinash Navlani, Understanding Random Forests Classifiers in Python.[En ligne]:available at: <https://www.datacamp.com/community/tutorials/random-forests-classifier-python>. [Accès le 2juillet 2020].

[21] Avinash Navlani , Understanding Logistic Regression in Python, [En ligne]:available at: <https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python>. [Accès le 5 juillet 2020].

[22] Magnan, C. (2005). Apprentissage semi-supervisé asymétrique et estimations d'affinités locales dans les protéines. In *CAP* (pp. 297-312).

[23] Mehdi Hadji, Analyse des sentiments : Généralités, [En ligne]:available at: <https://medium.com/@mehdihadji/analyse-des-sentiments-g%C3%A9n%C3%A9ralit%C3%A9s-99ab87503a5e>. [Accès le 28 juillet 2020].

[24] Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in human behavior*, 31, 527-541.

[25] Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment analysis of review datasets using naive bayes and k-nn classifier. *arXiv preprint arXiv:1610.09982*.

[26] Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

[27] Karoui, J., Gilles, N., Zitoune, F. B., & Belguith, L. (2014, April). Le langage figuratif dans le web social: cas de l'ironie et du sarcasme.

[28] Zhang Wenhao, Hua Xu, Wan Wei. Weakness finder: find product weakness from Chinese reviews by using aspects based sentiment analysis. *Expert Syst Appl* 2012;39:10283–91.

[29] Moraes Rodrigo, Valiati Joao Francisco, Gaviao Neto Wilson P.Document-level sentiment classification: an empirical comparison between SVM and ANN. *Expert Syst Appl* 2013;40: 621–33.

[30] Li Yung-Ming, Li Tsung-Ying. Deriving market intelligence from microblogs. *Decis Support Syst* 2013.

[31] Kang Han Hoon, Yoo Seong Joon, Han Dongil. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Syst Appl* 2012;39:6000–10.

[32] ZIANI Amel, « La recommandation via l'analyse d'opinions », thèse de doctorat, Université de Badji Mokhtar Annaba, 2017/2018.

[33] Nouali, O., Kirat, S., & Meziani, H. (2011). Mécanisme de prédiction pour une plateforme de filtrage collaboratif. *Revue RIST*, 19(2), 7-10.

[34] Nguyen, A. T., Denos, N., Berrut, C., & Thi, B. T. D. (2006). Modèle de formation multiple de communautés dans un système de recommandation hybride. *INFORSID*.

[35] Ménard, M. (2014). Systèmes de recommandation de biens culturels. *Les Cahiers du numérique*, 10(1), 69-94.

[36] Elsa Negre, (20/09/2018), Les systèmes de recommandation : une catégorisation , [En ligne].available at : <https://interstices.info/les-systemes-de-recommandation-categorisation/>..[Accès le 1août 2020].

[37] M. R. Dahimene, « Filtrage et Recommandation sur les Réseaux Sociaux », thèse de doctorat, École Doctorale Informatique, Télécommunication et Électronique (Paris) Laboratoire CEDRIC – Équipes ISID/VERTIGO , le 8 Décembre 2014.

- [38] Zhang, Y., Liu, R., & Li, A. (2015, December). A novel approach to recommender system based on aspect-level sentiment analysis. In *2015 4th National Conference on Electrical, Electronics and Computer Engineering*. Atlantis Press.
- [39] Ebadi, A., & Krzyzak, A. (2016). A hybrid multi-criteria hotel recommender system using explicit and implicit feedbacks. *International Journal of Computer and Information Engineering*, 10(8), 1377-1385.
- [40] Hassan, A. K. A., & Abdulwahhab, A. B. A. (2017). Reviews Sentiment analysis for collaborative recommender system. *Kurdistan journal of applied research*, 2(3), 87-91.
- [41] Publié par Ahmad Darwich | Mai 13, 2019 | Machine Learning, A LA DÉCOUVERTE DES SYSTÈMES DE RECOMMANDATION, <https://blog.invivoo.com/systeme-de-recommandation/>. [Accès le 1 août 2020].
- [42] Dossier de Mathieu dans l'épisode #83.en ligne at : <https://www.podcastscience.fm/dossiers/2012/04/25/les-algorithmes-de-recommandation/>. [Accès le 4 août 2020].
- [43] Frank Meyer, "Recommender systems in industrial contexts", 25 janvier 2012, préparée au sein du Laboratoire LIG dans l'École Doctorale MSTII.
- [44] Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17, 26-32.
- [45] Al Shamsi, F., & Guessoum, A. (2006, April). A hidden Markov model-based POS tagger for Arabic. In *Proceeding of the 8th International Conference on the Statistical Analysis of Textual Data, France* (pp. 31-42).
- [46] Introduction au NLP : le traitement de texte automatisé. [En ligne].available at : <https://blog.coddity.com/article/natural-language-processing/>. [Accès le 15 août 2020].
- [47] [En ligne].available at : <https://www.commonlounge.com/discussion/99e86c9c15bb4d23a30b111b23e7b7b1>. [Accès le 17 août 2020].
- [48] [En ligne].available at : <https://www.ionos.fr/digitalguide/web-marketing/analyse-web/analyse-tf-idf/>. [Accès le 20 août 2020].
- [49] TF*IDF. [En ligne].available at : https://fr.ryte.com/wiki/TF*IDF#TF. [Accès le 23 août 2020].
- [50] [En ligne].available at : https://www.w3schools.com/python/python_intro.asp. [Accès le 27 août 2020].
- [51] Natural Language Toolkit (NLTK). [En ligne].available at : <https://www.techopedia.com/definition/30343/natural-language-toolkit-nltk>. [Accès le 1 septembre 2020].
- [52] [En ligne].available at : codecademy.com/articles/scikit-learn. [Accès le 1 septembre 2020].
- [53] [En ligne].available at : <https://joblib.readthedocs.io/en/latest/>. [Accès le 3 septembre 2020].
- [54] [En ligne].available at : <https://pypi.org/project/Flask/>. [Accès le 27 août 2020].
- [55] C. Maklin, «TF IDF | Exemple Python TFIDF,» 5 mai 2019. [En ligne]. Available: <https://towardsdatascience.com/natural-language-processing-feature-engineering-using-tfidf8b9d00e7e76>. [Accès le 20 septembre 2020].
- [56] Mehdi Hadji, « Analyse des sentiments : Généralités ». [En ligne]. Available: <https://medium.com/@mehdihadji/analyse-des-sentimentsg%C3%A9n%C3%A9ralit%C3%A9s-99ab87503a5e>. [Accès le 25 septembre 2020].
- [57] [En ligne]. Available: <https://books.openedition.org/oep/204>. [Accès le 25 septembre 2020].

[58] [En ligne]. Available: <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>. [Accès le 25 septembre 2020].

[59] [En ligne]. Available : <http://cedric.cnam.fr/vertigo/Cours/ml2/coursArbresDecision.html>. [Accès le 25 septembre 2020].

[60] Flora Even, l'influence de Facebook sur les idées politiques. [En ligne].available at : https://www.rtb.be/culture/dossier/chroniques-culture/detail_l-influence-de-facebook-surlesidees-politiques-flora-eveno?id=9458372, Visité le 16/02/2019.
<https://joblib.readthedocs.io/en/latest/>. [Accès le 26 septembre 2020].