

République Algérienne Démocratique et Populaire  
Ministère de l'enseignement Supérieur et de la Recherche Scientifique  
Université de Mohamed El Bachir El Ibrahimi de Bordj Bou Arréridj  
Faculté des Mathématiques et d'Informatique  
Département d'informatique



## **MEMOIRE**

Présenté en vue de l'obtention du diplôme

### **Master en informatique**

Spécialité : Réseaux et multimédia

## **THEME**

Manipulation des données multilingues dans l'analyse des sentiments

***Présenté par :***

DEBAB Wafa

DJERBOA Zohra

***Soutenu publiquement le :*** 20/06/2024

***Devant le jury composé de:***

Dr.BENAOUDA Nadjib

MCB- Université de BBA

Président

Dr.MOUHDEB Djamila

MCA- Université de BBA

Examineur

Dr.SAIFI Lynda

MCB- Université de BBA

Encadreur

**2023/2024**

# *Dédicace*

La route n'était pas courte et elle ne devrait pas l'être. Le rêve n'était pas proche, mais le chemin était bordé de facilités. Mais je l'ai fait, louange à Dieu qui facilite les débuts et nous fait atteindre la fin par Sa grâce et Sa générosité.

À la lumière qui a éclairé ma route et à la lampe dont l'éclat ne s'éteint jamais dans mon cœur, à celui qui a couronné son front de sueur et m'a enseigné que le succès ne vient qu'avec la patience et la détermination, à celui qui a offert le cher et le précieux, et dont j'ai puisé ma force et ma fierté en moi-même.

« Mon cher père (*Nourredine*) »

À celle qui incarne la vie dans la vie, à toi devant qui la lettre s'incline avec amour et gratitude, à celle que Dieu a élevée au-dessus des cieux et dont le cœur m'a embrassé devant sa main, à celle qui a illuminé mes peines de ses prières, à ce cœur compatissant et à la bougie qui était ma force et mon guide dans les nuits sombres et la lumière de mon chemin, à Dia ma vie.

« Ma Cher mère (*Hayet*) »

À ceux qui ont tendu leurs mains dans ma faiblesse et ont cru en ma capacité, à mon pilier solide, à ceux dont la présence me donne une force et un amour sans limite, à ceux avec qui j'ai appris la signification de la vie.

« Mes frères (*Najib, Abdelhak, Ilyas*) »

« Mes sœurs (*Ibtissam, Hanane*) »

À ceux qui font partie de moi, à ceux qui m'insufflent de l'optimisme et de la vie, à mes petits et à mes chéris.

« Les enfants de mes sœurs (*Rassim, Sidra, Taim*) »

À celle qui s'est distinguée par la fraternité, la loyauté et le don, à ma compagne, A celle qui a partagé mes plus grandes joies, à celle dont la présence est un baume à mon âme.

« Mon binôme (*Zahra*) »

À ceux avec qui j'ai partagé les plus beaux moments de ma vie, à mes bougies, à mes amis les plus fidèles et compagnons des années

« Mes amis (*Aya, Dounia, Nadine, Fatima, Hiba*) »

Je vous dédie cette réalisation et le fruit de mon succès que j'ai toujours souhaité. Enfin, je voudrais souligner que ce n'est pas la fin, mais plutôt le début d'un nouveau voyage plein de défis riche en opportunités et en expériences à venir.

*Wafa*

# *Dédicace*

Aujourd'hui, je me tiens devant vous, mon cœur rempli d'un mélange de sentiments contradictoires, la joie de l'accomplissement et la tristesse de la séparation, des émotions mêlées de souvenirs d'un long parcours éducatif, rempli de défis et de réussites.

Ce parcours n'a pas été court, et il ne devrait pas l'être. Le rêve n'était pas proche, et le chemin n'était pas pavé de facilités. Mais je l'ai fait, je l'ai réalisé.

Je dédie ce modeste travail et ma profonde gratitude:

À celui qui fut le premier et éternel soutien, dont la présence m'a procuré des efforts inlassables et dont les prières incluaient toujours mon nom, mon premier professeur, maman *Ahlam*, je vous dédie cette réalisation qui n'aurait pas été possible sans vous. Je vous dédie toutes mes étapes et réalisations. Le mérite et la louange reviennent au Dieu, puis à votre lutte pour moi ici. Aujourd'hui, je vous dédie la connaissance et le témoignage que je les ai abandonnés pour mes soins et mon éducation. Reconnaisante que Dieu t'ait choisie parmi l'humanité comme ma mère.

À celui qui m'a soutenu dans mon parcours sans limites et m'a donné sans attendre en retour, *Benhamadi Lahcene* mon exemple éternel, mon soutien moral et source de joie et de bonheur, celui Tu m'as appris à être forte, à ne jamais abandonner et à pour suivre mes rêves avec passion. Tu es une source d'inspiration pour moi et pour tant d'autres. Merci d'être le meilleur oncle du monde,

À *Ma grand-mère* et *Mon grand-père*, qui ont été d'une aide et d'un soutien constants depuis mon enfance.

Aux amis des années, *Fatima* et *Riham*, aux personnes traversant des épreuves et à ceux qui inspirent ma réussite. À ceux qui m'ont fait sourire dans les moments difficiles, à ceux qui ont révélé l'étendue de ma force et de mes capacités, à ceux qui ne cessent de m'encourager et croient en mon courage. Aux bougies qui éclairent toujours mon chemin

A mon binôme mon bras droit *Wafa* pour son soutien moral, sa patience et sa compréhension tout au long de ce projet et au long des 5 années d'université

Mes merveilleuses filles : *Aya, Amina, Khadidja, Fatima, Nadine, Dounia, Hiba, Nada* avec elles j'ai passé l'une des meilleures années de mon cursus universitaire.

Dieu merci, pour Sa grâce et gratitude, grâce à laquelle je regarde aujourd'hui. Dans un rêve tant attendu qui est devenu une réalité fière.

*Zohra*

# *Remerciements*

Ce mémoire a été réalisé dans le cadre de nos études en vue d'obtenir le diplôme d'ingénieur en informatique. Tout d'abord, nous souhaitons exprimer notre gratitude envers Dieu pour nous avoir accordé la force et la santé nécessaires pour mener à bien ce travail. Nous tenons à remercier chaleureusement nos parents pour leur soutien, leurs sacrifices et leur compréhension tout au long de nos années d'études.

Nos sincères remerciements vont à *Mme Lynda Saïfi* pour son encadrement exceptionnel, ses précieux conseils, ses remarques pertinentes et sa disponibilité lors de l'élaboration de cette étude. Nous tenons également à exprimer notre profonde reconnaissance envers notre collègue *Omar Embarki* pour sa collaboration précieuse et son soutien continu tout au long de notre projet.

Nous souhaitons également remercier tous les membres du jury pour avoir accepté de participer à l'évaluation de ce travail, démontrant ainsi leur intérêt pour cette recherche. Nos remerciements s'adressent également à l'ensemble du personnel enseignant, technique et administratif du département d'informatique de l'université de Borj Bou Arréridj pour leur disponibilité et leur bienveillance.

Enfin, nous tenons à exprimer notre gratitude à toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce modeste travail. Merci à tous pour votre soutien et votre collaboration précieuse.

# Résumé

Ce projet de fin d'études se concentre sur la gestion efficace du volume important de commentaires et d'avis multilingues de consommateurs pour les entreprises et les porteurs de projets. En se basant sur l'analyse des sentiments et le text mining, l'étude explore différentes approches, telles que les réseaux neuronaux, les SVM, la régression logistique, le Naïve de Bayes, les arbres de décision et les forêts aléatoires, pour traiter les données en français et en anglais. Une comparaison détaillée de ces méthodes est réalisée pour déterminer la plus adaptée à l'analyse des sentiments et au text mining multilingue. De plus, deux méthodes de numérisation distinctes, tf-idf et le codage one-hot vecteur, sont expérimentées pour évaluer leur efficacité dans l'analyse des données multilingues

**Mots-clés :** Fouille de texte, analyse de sentiments, réseaux neuronaux, classification, comparaison, prédiction, données multilingue.

# Abstract

This end-of-studies project focuses on the effective management of the large volume of multilingual consumer comments and reviews for companies and project leaders. Based on sentiment analysis and text mining, the study explores different approaches, such as neural networks, SVMs, logistic regression, Bayes Naive, decision trees and random forests, to process data in French and English. A detailed comparison of these methods is made to determine the most suitable for sentiment analysis and multilingual text mining. In addition, two distinct scanning methods, tf-idf and one-hot vector coding, are being tested to assess their effectiveness in analyzing multilingual data

**Keywords:** Text mining, sentiment analysis, neural networks, classification, comparison, prediction, multilingual data

## ملخص

يركز مشروع نهاية الدراسات هذا حول الإدارة الفعالة لحجم كبير من تعليقات المستهلكين واستعراضاتهم بمختلف اللغات للشركات وقادة المشاريع. من خلال تحليل المشاعر وتعدين النصوص، تبحث الدراسة في أساليب مختلفة مثل الشبكات العصبية، و SVMs، والانحدار اللوجستي، Naive Bayes، وأشجار القرار والغابات العشوائية لمعالجة البيانات باللغتين الفرنسية والإنجليزية. يُجرى مقارنة دقيقة بين هذه الأساليب لتحديد الأنسب لتحليل المشاعر وتعدين النصوص بلغات متعددة. بالإضافة إلى ذلك، يتم اختبار طريقتين مميزتين للمسح، وهما ترميز ناقل واحد ساخن و tf-idf، لتقييم فعالتهما في تحليل البيانات بلغات متعددة.

**الكلمات الرئيسية:** تعدين النصوص، تحليل المشاعر، الشبكات العصبية، مقارنة، متعدد اللغات

# Table des matières

**Liste des abréviations**

**Liste des figures**

**Liste des tableaux**

**Liste des algorithmes**

**Introduction Générale .....16**

**Chapitre 01 : Text Mining .....18**

1.1. Introduction .....18

1.2. Définition .....18

1.3. Le processus du text mining .....18

1.4. Les Taches du text mining .....20

1.5. Représentation du texte .....21

1.6. Domaines d'utilisation du text mining .....22

1.7. Les Avantages du text mining .....22

1.8. La Relation entre text mining et l'analyse des sentiments .....23

1.9. Conclusion .....23

**Chapitre 02 : Analyse des sentiments.....24**

2.1. Introduction .....24

2.2. Définition de l'analyse des sentiments .....24

2.3. Les niveaux de l'analyse des sentiments .....24

2.4. Différentes approches et techniques.....26



2.5. Domaines d'application de l'analyse des sentiments.....	27
2.6. Défis dans les analyses des sentiments .....	27
2.7. Conclusion .....	28
<b>Chapitre 03: Méthodologie.....</b>	<b>29</b>
3.1. Introduction .....	29
3.2. Description du projet.....	29
3.3. Description de la méthodologie de conception.....	29
3.3.1. Collection de données.....	30
3.3.2. Prétraitement de données .....	32
3.3.3. Extraction des caractéristiques.....	34
3.4. Méthodes utilisées pour la classification du texte.....	35
3.4.1. Réseau neuronal (RNA).....	35
3.4.2. Méthodes de base .....	35
1. Machine à vecteurs de support (SVM).....	35
2. Régression logistique .....	36
3. Naïve de Bayes .....	37
3.4.3. Apprentissage par ensemble (Ensemble Learning) .....	38
1. Arbre de décision .....	38
2. Les forêts aléatoires.....	39
3.5. Mesures des performances .....	40
3.6. Conclusion .....	41
<b>Chapitre 04 : Implémentation.....</b>	<b>42</b>
4.1. Introduction .....	42

4.2. Environnement et outils d'implémentation .....	42
4.2.1. Matériel .....	42
4.2.2. Langage de programmation .....	42
4.2.3. Environnement de programmation.....	43
4.2.4. Les principaux package Python utilisés.....	43
4.3. Analyse exploratoire des données multilingues.....	44
4.4. Génération du nuage de mots .....	44
4.5. Diagrammes à bandes.....	46
4.6. Diagramme de n gramme .....	46
4.7. Application des algorithmes de classification de données .....	48
4.8. Génération d'une interface des résultats .....	49
4.9. Conclusion.....	50
<b>Chapitre 05 : Résultats et discussions.....</b>	<b>51</b>
5 .1. Introduction.....	51
5.2. Les Résultats obtenus .....	51
5.2.1. Réseaux neuronal (RNA).....	51
5.2.2. Les Cinq Méthodes de base de classification.....	53
5.3. Résultat d'exécution .....	55
5.4. Discussion des résultats.....	56
5.5. Conclusion.....	56
<b>Conclusion générale .....</b>	<b>57</b>
<b>Les références .....</b>	<b>59</b>
<b>Annexe.....</b>	<b>61</b>

## Liste des abréviations

**BOW:** Bag Of Words

**EI :** L'extraction d'informations

**FPR :** False Positive Rate

**NLP :** Traitement du langage naturel

**NLTK :** Natural Language Toolkit

**PRC :** courbe de la Précision –Rappel

**PHP:** Hyper text Preprocessor

**RAM:** Random Access Memory

**RBF :** Radial Basis Function

**RI :** La Recherche d'Information

**RNA:** Réseau neuronal

**ROC:** Receiver Operating Characteristic

**SVM :** Machine à vecteurs de support

**TF-IDF:** Term Frequency-Inverse Document Frequency

**TPR:** True Positive Rate

**URL:** Uniform Resource Locator

**Vs:** Visual Studio

# Liste des figures

<b>Figure 1:</b> Le processus de text mining .....	18
<b>Figure 2:</b> Les taches du Text Mining .....	20
<b>Figure 3:</b> L'analyse des sentiments au niveau de la phrase .....	25
<b>Figure 4:</b> Architecture de system proposé.....	30
<b>Figure 5:</b> Fichier train.csv .....	31
<b>Figure 6:</b> Diagramme à barres représentant les étoiles.....	32
<b>Figure 7:</b> Un cercle relatif représentant les langues.....	32
<b>Figure 8:</b> Répartition des avis selon les catégories de produits....	32
<b>Figure 9:</b> Architecture générale d'un réseau de neurones artificiels .....	35
<b>Figure 10:</b> Le principe de SVM.....	35
<b>Figure 11:</b> Classification de Naïve Bayes .....	37
<b>Figure 12:</b> Nuage de mots anglais (100 meilleurs mots) .....	45
<b>Figure 13:</b> Nuage de mots français (100 meilleurs mots) .....	45
<b>Figure 14:</b> Diagramme de bandes de l'ensemble de donne.....	46
<b>Figure 15:</b> Diagramme anglais de 2 gramme (top 20) .....	46
<b>Figure 16:</b> Diagramme anglais de 3 gramme (top 20) .....	47
<b>Figure 17:</b> Diagramme français 2gramme (top 20) .....	47
<b>Figure 18:</b> Diagramme anglais 3 gramme (top 20).....	48
<b>Figure 19:</b> Découpage d'une ensemble de données en 80% pour l'entrainement et 20% pour le test.....	48
<b>Figure 20 :</b> Une fenêtre d'analyse des sentiments.....	49

<b>Figure 21:</b> Courbe ROC (gauche) et courbe PRC (droite) et courbe de Training loss de Réseau neuronal avec tf_idf et one-hot vecteur .....	52
<b>Figure 22:</b> La matrice de confusion du Réseau neuronal avec tf_idf(gauche) et avec one-hot vecteur (droite) .....	53
<b>Figure 23:</b> Courbe ROC (gauche) et courbe PRC (droite) des méthodes de base avec tf_idf et one-hot vecteur .....	55
<b>Figure 24:</b> Résultat d'exécution des commentaires.....	55

# Liste des tableaux

<b>Tableau 1:</b> Le principe de vecteur One Hot .....	34
<b>Tableau 2:</b> Matrice de confusion .....	40
<b>Tableau 3:</b> Caractéristiques des matériels utilisés .....	42
<b>Tableau 4:</b> Nombre du commentaire avant et après le prétraitement .....	44
<b>Tableau 5:</b> Exemples des commentaires après l'application des fonctions de prétraitement .....	49
<b>Tableau 6:</b> Le résultat de performance de Réseau neuronal avec tf-idf et one-hot vecteur	51
<b>Tableau 7:</b> Comparaison des résultats de performance des méthodes de bases avec tf-idf. .....	54
<b>Tableau 8:</b> Comparaison des résultats de performance des méthodes de bases avec one-hot vecteur.....	54

# Liste des algorithmes

Algorithme 1 : les étapes de la méthode SVM.....	36
Algorithme 2: les étapes de la méthode régression logistique .....	37
Algorithme 3: les étapes de la méthode Naïve de bayes.....	38
Algorithme 4: les étapes de l'algorithme Arbre de décision .....	38
Algorithme 5: les étapes de la méthode foret aléatoire.....	39

# Introduction Générale

## 1. Contexte

La manipulation de données multilingues, conjuguée à l'analyse des sentiments et au text mining, constitue un domaine de recherche en expansion constante dans le contexte actuel de mondialisation et de digitalisation croissante des échanges d'informations. Les progrès technologiques ont facilité la collecte et l'analyse de vastes ensembles de données provenant de sources variées et rédigées dans différentes langues, ouvrant ainsi la voie à des opportunités uniques pour appréhender les tendances, les opinions et les émotions des utilisateurs à l'échelle mondiale.

## 2. Problématique :

Cependant, avec l'avancée actuelle de la science et de la technologie, les porteurs de projets et les entreprises se retrouvent confrontés à la difficulté de lire des milliers de commentaires pour comprendre l'avis des consommateurs sur leur produit, évaluer sa réussite et déterminer s'il est apprécié par les clients. Ce volume important de commentaires représente un véritable défi en termes de gestion du temps, surtout lorsque ces avis sont écrits dans différentes langues.

Quel est le meilleur moyen d'interpréter efficacement ce volume important de commentaires et d'avis multilingues de consommateurs pour les entreprises et les porteurs de projets ?

## 3. Objectif :

Dans le cadre de ce projet de fin d'études, nous aborderons les défis et les opportunités associés à la manipulation de données multilingues, en mettant particulièrement l'accent sur l'analyse des sentiments et le text mining pour extraire des informations pertinentes et exploitables à partir de ces données diversifiées. Nous étudierons diverses approches pour traiter efficacement ces données multilingues (français et anglais), en utilisant des techniques telles que les réseaux neuronaux, les SVM, la régression logistique, le Naïve de Bayes, les arbres de décision et les forêts aléatoires. Une comparaison approfondie de ces méthodes sera réalisée afin de déterminer celle qui convient le mieux à l'analyse des sentiments et au text mining dans un contexte multilingue. De plus, nous expérimenterons deux méthodes



distinctes de numérisation de text mining, à savoir TF-IDF (Term Frequency-Inverse Document Frequency) et le codage one-hot vecteur, pour traiter les données multilingues et évaluer leur efficacité dans notre analyse.

#### **4. Structure du rapport :**

La structure de ce rapport comprend cinq chapitres qui détaillent la méthodologie utilisée dans ce projet :

**Chapitre 01:** présente Le processus du text mining, ses taches, quelques méthodes de présentation du texte, ses domaines d'utilisation et quelques avantages

**Chapitre 02 :** présente l'analyse des sentiments, ses différents niveaux, ses approches, ses domaines d'application et quelques défis.

**Chapitre 03 :** concerne la conception et l'architecture de notre système.

**Chapitre 04 :** présente la partie implémentation de notre système, où nous présenterons l'environnement de développement, Les principaux packagent python utilisés, Ainsi que la préparation de données pour la classification

**Chapitre 05 :** concerne les résultats obtenus, nous avons exposé tous les tests expérimentaux réalisés sur deux ensembles de données

# Chapitre 01 : Text Mining

## 1.1. Introduction :

L'analyse d'une grande quantité de données est difficile et demande beaucoup de temps et d'effort pour comprendre ces données. Par conséquent, la recherche de données est devenue une nécessité. Le développement technologique rapide, cela a conduit à l'accumulation rapide de la quantité de données stockées dans les bases de données. Les outils, méthodes et techniques d'exploration de données nous permettent d'analyser ces données et de trouver des modèles et des informations cachées. Les techniques de Text Mining ont également été utilisées pour faciliter l'extraction de données en cas de problèmes avec les textes.

## 1.2. Définition :

Le Text Mining communément appelé Fouille de texte ou découverte de connaissances à partir de données textuelles fait généralement référence au processus d'extraction de modèles d'intérêt ou de connaissances auparavant inconnues. Le Text Mining fait partie intégrante en science des données, donc en intelligence artificielle. C'est un ensemble de méthodes d'analyse linguistique, de techniques et d'outils utilisés pour analyser et traiter des données textuelles, qui sont généralement des données non structurées et non référencées dans une base de données. Il existe différents types de données textuelles : textes dactylographiés, mots de passes , e-mails, PowerPoint, etc. [1]

## 1.3. Le processus du text mining :

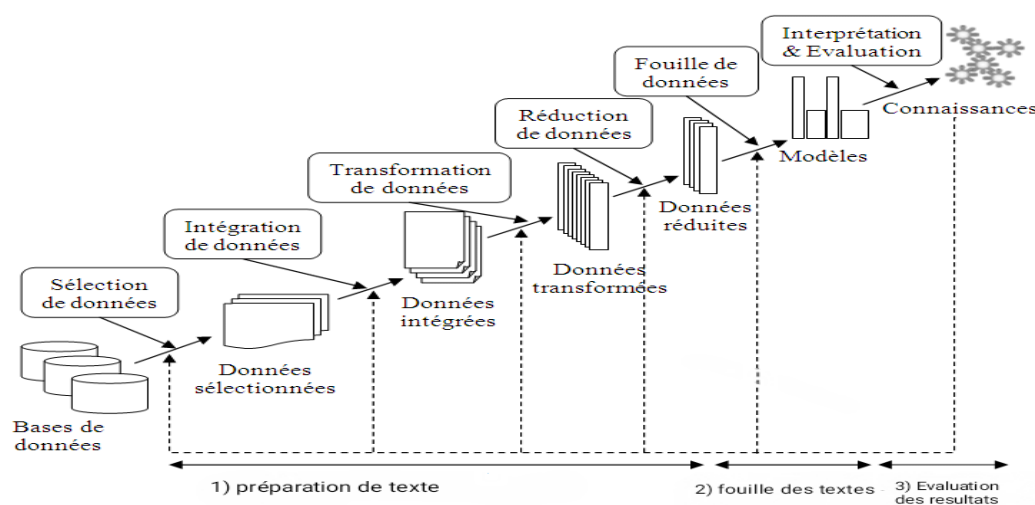


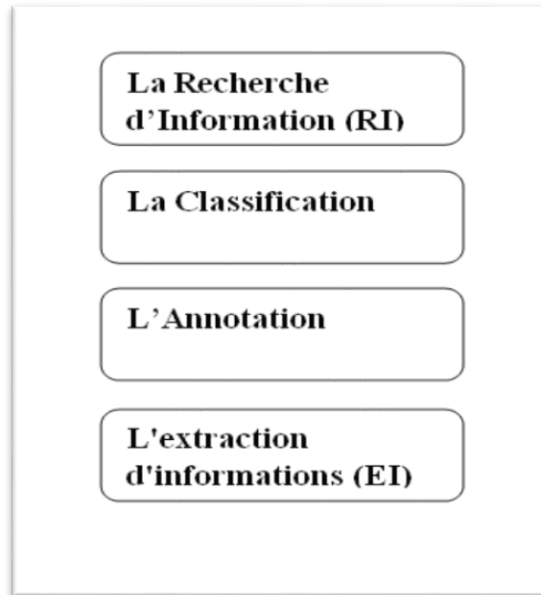
Figure 1:Le processus de text mining [26]

- 1- Collecte des données** : rassembler un grand corpus de documents textuels à analyser, tels que des articles de presse, des livres, des messages sur les réseaux sociaux, etc
- 2- Prétraitement des données** : Le prétraitement de texte est une étape essentielle dans le processus d'analyse de données textuelles. Il vise à nettoyer les données en supprimant la ponctuation, les chiffres, les mots vides, et en normalisant le texte en le convertissant en minuscules, Voici quelques étapes courantes du prétraitement de texte :
  - **Tokenisation** : découper le texte en tokens, qui peuvent être des mots individuels, des phrases, ou d'autres éléments textuels.
  - **Analyse lexicale** : repérer les mots clés et les entités nommées dans le texte afin d'en extraire des informations pertinentes.
  - **Analyse syntaxique** : étudier la structure grammaticale du texte pour comprendre les relations entre les mots et les phrases.
- 3- Extraction d'informations** : extraire des informations spécifiques du texte, telles que des tendances, des opinions, ou des relations entre les entités.
- 4- Modélisation et visualisation** : recourir à des techniques de modélisation statistique ou d'apprentissage automatique pour analyser les données textuelles et présenter les résultats sous forme graphique ou sous forme de tableaux.
- 5- Interprétation des résultats** : interpréter les informations extraites du texte pour en déduire des conclusions éclairées et prendre des décisions importantes.

Ce processus synthétise les principes fondamentaux du text mining tels qu'ils sont exposés dans la littérature académique. Il peut varier en fonction du contexte spécifique et des techniques employées dans chaque application. [3]

## 1.4. Les taches du text mining :

La fouille de texte se divise en quatre tâches élémentaires qui sont :



**Figure 2:**Les taches du Text Mining

- 1. La Recherche d'Information (RI) :** La recherche d'information consiste à trouver des documents non structurés dans une grande collection en fonction d'un besoin d'information, sans requêtes structurées ni réponses exactes. Elle est utilisée dans divers contextes comme la recherche sur le Web, dans les boîtes mail ou sur les ordinateurs. [4]
- 2. La Classification :** La classification en text mining consiste à attribuer des étiquettes ou des catégories à des documents textuels en fonction de leur contenu, en utilisant des algorithmes d'apprentissage automatique pour structurer efficacement de grandes quantités de données textuelles. [5]
- 3. L'Annotation :** L'annotation dans le text mining consiste à ajouter des étiquettes ou des métadonnées à des documents textuels pour les organiser et les analyser efficacement. Elle peut être faite manuellement ou automatiquement grâce à des algorithmes. Cette étape est importante pour structurer les données textuelles en vue d'une analyse ultérieure et peut être utilisée pour entraîner des modèles d'apprentissage automatique.[6]

- 4. L'extraction d'informations (EI) :** La restructuration consiste à extraire des données structurées à partir de données non structurées pour les rendre exploitables. Dans le contexte de la classification de texte, l'extraction d'informations identifie des entités telles que des mots-clés ou des dates dans les documents, améliorant ainsi la précision et l'efficacité des techniques de classification en extrayant des informations pertinentes des textes. [7]

## **1.5. Représentation du texte :**

La représentation du texte revêt une importance cruciale dans le processus de classification du texte, car les algorithmes d'apprentissage ne peuvent pas traiter directement le texte, qui est sous forme de données non structurées, tout comme les images, les sons et les vidéos. [1]

### **- Représentation par n-grammes :**

La notion de n-grammes, en mettant l'accent sur les bi-grammes et les trigrammes (avec respectivement  $n=2$  et  $n=3$ ), a été introduite par Claude Shannon en 1948 dans le contexte de systèmes de prédiction de caractères en fonction des caractères précédemment saisis. Un n-gramme de X est défini comme une séquence de n éléments X consécutifs, pouvant être des caractères ou des mots.

La construction des n-grammes de caractères et de mots s'effectue en utilisant la méthode du déplacement de fenêtre, où chaque déplacement correspond à une étape, représentant un caractère ou un mot. Les éléments contenus dans la fenêtre ainsi définie servent de descripteurs pour un corpus donné.[1]

### **- Représentation en « sac de mots » « bag of words » :**

Le modèle Bag Of Words (BOW) est une approche simplifiée largement utilisée dans le traitement du langage naturel et dans la recherche d'informations. Dans ce modèle, un texte est représenté comme une collection non ordonnée de mots, sans considération de la grammaire ou de l'ordre des mots. En classification de texte, chaque mot reçoit un poids basé sur sa fréquence dans le document et entre les différents documents, formant ainsi le BOW. Bien que largement utilisé dans la classification de documents, le BOW présente des limitations, notamment en raison de l'exclusion de l'analyse grammaticale et de la structure des phrases. Cette approche peut rencontrer des difficultés avec les mots composés, les abréviations et les mots outils, ainsi que dans la distinction des mots d'une même famille. En conséquence, la représentation en "sac de mots" peut entraîner une perte de sémantique du texte. [1]

- **Représentation des textes par des phrases :**

Certains chercheurs suggèrent d'utiliser des phrases comme unité de représentation plutôt que des mots, malgré la simplicité de ces derniers. Les phrases conservent des informations sur la position des mots dans la phrase, réduisant ainsi l'ambiguïté par rapport aux mots individuels. Elles sont plus utiles que les mots seuls. Cependant, le traitement de toutes les combinaisons possibles de phrases peut poser un problème de taille. Pour remédier à cela, une approche consiste à sélectionner des phrases pertinentes et riches en sens. [1]

- **Représentation des textes par TF-IDF (Term Frequency-Inverse Document Frequency) :**

Consiste à assigner un poids à chaque terme en fonction de sa fréquence dans un document donné par rapport à sa fréquence générale dans l'ensemble du corpus. Les termes qui sont fréquents dans un document mais rares dans l'ensemble des documents se verront attribuer un poids plus important. [1] Nous avons expliqué cette représentation brièvement voir chapitre 03 section 3.3

## **1.6. Domaines d'utilisation du text mining :**

Le Text Mining offre de nombreuses applications, notamment dans le service client, où il trie les requêtes, évalue l'efficacité et la satisfaction des clients grâce à l'analyse des retours. Dans la gestion des risques, il surveille les tendances financières pour une meilleure prise de décision. En maintenance industrielle, il permet une maintenance proactive en identifiant les problèmes émergents. Dans le domaine de la santé, il extrait des informations médicales, et en cybersécurité, il aide à filtrer les spams pour renforcer la protection contre les cyberattaques. [3]

## **1.7. Les avantages du text mining :**

Le text mining permet d'extraire des informations utiles à partir de grandes quantités de texte, ce qui aide à comprendre les tendances, les besoins des clients et les opportunités du marché. Il automatise ce processus, ce qui économise du temps et des ressources. De plus, il aide à prendre des décisions éclairées, à personnaliser les services, à détecter les fraudes, à améliorer la qualité des produits et à rester compétitif en surveillant les concurrents [2]

## **1.8. La relation entre text mining et l'analyse des sentiments :**

L'exploration de texte dans l'analyse des sentiments implique l'utilisation de techniques de traitement du langage naturel pour extraire et analyser des informations subjectives à partir de données textuelles. Ce processus vise à identifier et classifier les opinions, émotions et sentiments exprimés dans le texte, tels que positifs, négatifs ou neutres. Les techniques d'exploration de texte comme le codage, le marquage partiel de la parole et l'analyse du lexique des sentiments sont fréquemment employées dans l'analyse émotionnelle pour extraire et analyser les émotions des données textuelles. En appliquant ces techniques à l'analyse des sentiments, les entreprises et institutions peuvent obtenir des informations précieuses sur les opinions des clients, les tendances du marché et les perceptions du public.

## **1.9. Conclusion :**

Dans cette section, notre travail a été consacré à la fouille de texte, nous avons présenté : le processus et les tâches de text mining , Nous avons également abordé la représentation de texte, Les avantages, et les domaines de text mining

# **Chapitre 02 : Analyse des sentiments**

## **2.1. Introduction :**

Dans 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. 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 concept et définitions liée à l'analyse des sentiments.

## **2.2. Définition de l'analyse des sentiments :**

L'analyse de sentiments est une méthode qui permet d'identifier les émotions, les opinions et les attitudes exprimées dans un texte, un discours ou tout autre type de communication. Elle fait appel à des technologies telles que l'intelligence artificielle, le traitement du langage naturel et d'autres techniques d'analyse de données pour détecter et évaluer le ton, le sentiment et l'émotion présents dans un contenu donné. Cette analyse trouve des applications dans divers domaines tels que le marketing, la surveillance de la concurrence, le service client, la politique et d'autres secteurs où la compréhension des sentiments et des opinions revêt une grande importance. [9]

## **2.3. Les niveaux de l'analyse des sentiments :**

L'analyse des sentiments peut être réalisée à divers niveaux, en fonction de l'objectif et du contexte. Ces niveaux d'analyse peuvent être combinés et adaptés en fonction des besoins spécifiques du projet ou de l'application. La plupart des outils d'analyse des sentiments offrent des fonctionnalités pour travailler à différents niveaux d'analyse afin de répondre aux besoins variés des utilisateurs .Voici quelques niveaux d'analyse des sentiments :

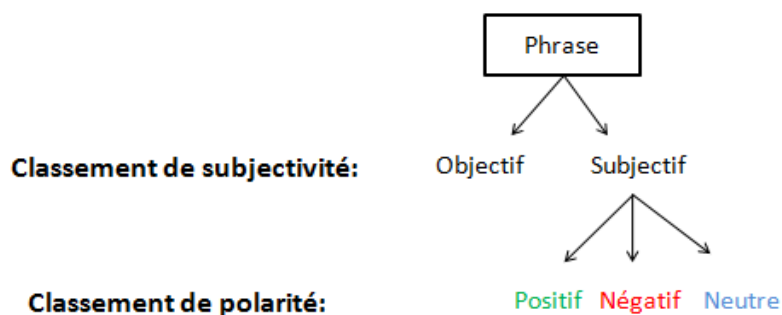


- **Analyse au niveau du document** : Cette approche se concentre sur l'ensemble d'un document, tel qu'un article de presse, un billet de blog ou un rapport, afin de déterminer le ton général, les émotions et les opinions exprimées. [10]

- **Analyse au niveau de la phrase ou du paragraphe** : Cette méthode s'intéresse à des unités plus petites de texte, comme des phrases ou des paragraphes, pour repérer les nuances et les variations de ton à l'intérieur d'un document. 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. 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 [11]:

**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 3:**L'analyse des sentiments au niveau de la phrase

- **Analyse au niveau de l'entité ou du mot-clé** : À ce niveau, l'analyse des sentiments se concentre sur des entités spécifiques ou des mots-clés dans le texte, tels que des marques, des produits ou des sujets spécifiques. [10]

- **Analyse au niveau des modelés** : À ce niveau d'analyse, l'accent est mis sur la détermination du sentiment envers des caractéristiques ou aspects spécifiques commentés dans un document. Il implique l'identification et l'extraction des caractéristiques, la détermination de la polarité du sentiment (positive, négative, neutre) pour chaque caractéristique, ainsi que le regroupement des caractéristiques synonymes. [11]

## 2.4. Différentes approches et techniques :

Il existe diverses méthodes et techniques pour évaluer les sentiments, chacune présente ses avantages et ses inconvénients. Voici un aperçu des principales approches :

- **Analyse lexicale** : Cette méthode repose sur l'utilisation de dictionnaires de mots préalablement étiquetés avec des polarités positives, négatives ou neutres pour évaluer le sentiment d'un texte. Les mots se voient attribuer des scores en fonction de leur polarité, qui sont ensuite agrégés pour évaluer le sentiment global du texte. [13]

- **Classification de texte** : Cette technique fait appel à des algorithmes d'apprentissage automatique pour classer les textes en fonction de leur sentiment. Les méthodes couramment utilisées incluent les machines à vecteurs de support (SVM), les arbres de décision, les réseaux de neurones et les méthodes basées sur l'apprentissage profond. [13]

- **Analyse des émotions** : Cette approche vise à identifier les émotions spécifiques exprimées dans un texte, telles que la joie, la tristesse, la colère, la peur, etc. Les techniques d'analyse des émotions peuvent inclure l'utilisation de modèles linguistiques spécifiques pour détecter ces émotions.

- **Traitement du langage naturel (NLP)** : Les techniques de NLP sont largement utilisées pour l'analyse des sentiments, notamment pour la détection des sentiments dans les textes en utilisant des outils tels que la tokenisation, la lemmatisation, la détection de phrases-clés et l'analyse syntaxique.

- **Méthodes hybrides** : Certaines approches combinent plusieurs techniques d'analyse des sentiments pour obtenir des résultats plus précis. Par exemple, une méthode hybride pourrait

combiner l'analyse lexicale avec la classification de texte pour améliorer la précision de l'analyse des sentiments. [14]

Chaque méthode présente ses forces et ses faiblesses, et le choix dépend souvent du type de données disponibles, de la précision requise et du contexte d'application.

## **2.5. Domaines d'application de l'analyse des sentiments :**

L'analyse des sentiments est largement utilisée dans de nombreux domaines pour différentes applications. Voici quelques exemples [11] :

- **Veille médiatique** : Les entreprises surveillent les médias sociaux, les actualités en ligne et d'autres sources d'informations pour évaluer la perception de leur marque, de leurs produits ou de leurs services par le public.

- **Service client** : Les entreprises analysent les commentaires des clients sur les réseaux sociaux, les forums en ligne et les plateformes de commentaires afin de repérer d'éventuels problèmes et d'améliorer l'expérience client.

- **Analyse de marché** : Les entreprises cherchent à comprendre les tendances du marché, les opinions des consommateurs et les préférences des clients en analysant les conversations en ligne et les avis sur les produits.

- **Recherche académique** : Les chercheurs étudient les opinions publiques, les tendances politiques, les réactions aux événements mondiaux et d'autres sujets dans le cadre de la recherche académique en utilisant l'analyse des sentiments.

- **Recommandations personnalisées** : Les plateformes de streaming, de commerce électronique et d'autres services recommandent des produits, des films, des émissions de télévision ou de la musique en fonction des préférences et des émotions des utilisateurs grâce à l'analyse des sentiments.

## **2.6. Défis dans les analyses des sentiments :**

Les analyses de sentiment présentent plusieurs défis, notamment :

- **Ambiguïté du langage** : Les mots et les expressions peuvent avoir des significations multiples et être interprétés différemment en fonction du contexte, ce qui rend difficile la détermination précise du sentiment exprimé.

- **Ironie et sarcasme** : Les analyses de sentiment doivent être capables de reconnaître l'ironie et le sarcasme, qui peuvent inverser le sens littéral des mots pour exprimer une émotion opposée à celle qui est explicitement indiquée.

- **Variabilité culturelle et linguistique** : Les normes culturelles et les particularités linguistiques peuvent influencer la manière dont les émotions sont exprimées, ce qui rend difficile l'application de modèles d'analyse de sentiment génériques à des contextes culturels et linguistiques différents.

- **Données non structurées** : Les sources de données textuelles utilisées pour l'analyse de sentiment, telles que les médias sociaux ou les commentaires en ligne, peuvent contenir des erreurs grammaticales, des abréviations, des acronymes, des émoticônes, etc., ce qui complique l'analyse automatique.

- **Prise en compte du contexte** : Comprendre le contexte dans lequel les opinions sont exprimées est crucial pour interpréter correctement les sentiments. Par exemple, un même mot peut avoir des connotations positives ou négatives en fonction du sujet abordé.

- **Gestion des biais** : Les analyses de sentiment peuvent être influencées par des biais liés aux données d'entraînement des modèles, aux choix de catégorisation des émotions, ou encore à la présence de discours haineux ou polarisés... [12]

## 2.7. Conclusion :

Dans ce chapitre, nous avons présenté ce qu'est l'analyse du sentiment, avec ses quatre différents niveaux, ses utilisations et énuméré toutes les approches de classification des sentiments existantes, et nous avons aussi parlé des problèmes qui existaient dans ce type de tâche.

L'analyse des sentiments des fichiers textuels algériens représente un défi supplémentaire en raison de plusieurs facteurs. Tout d'abord, ces textes sont rédigés dans un dialecte arabe distinct, ce qui rend la tâche plus complexe. De plus, ces textes présentent fréquemment des problèmes d'orthographe et de grammaire. En outre, les ensembles de données disponibles sont souvent limités en termes d'échantillons de formation et de test, en raison de la complexité linguistique et du manque de normalisation. Les prochains chapitres aborderont ces défis et proposeront des solutions pour les surmonter.

# Chapitre 03: Méthodologie

## 3.1. Introduction :

Au cours de ce chapitre, nous présenterons une vue d'ensemble de notre système en mettant en avant ses aspects conceptuels et méthodologiques. Ensuite, nous examinerons en détail chaque étape du projet en mentionnant les principaux algorithmes et techniques employés, ainsi que les différentes mesures d'évaluation utilisées pour évaluer leurs performances.

## 3.2. Description du projet :

Dans ce projet, nous visons à comparer l'efficacité des performances des méthodes de classification traditionnelles (par exemple, SVM, arbres de décision, régression logistique ...) avec les réseaux neuronaux sur des ensembles de données.

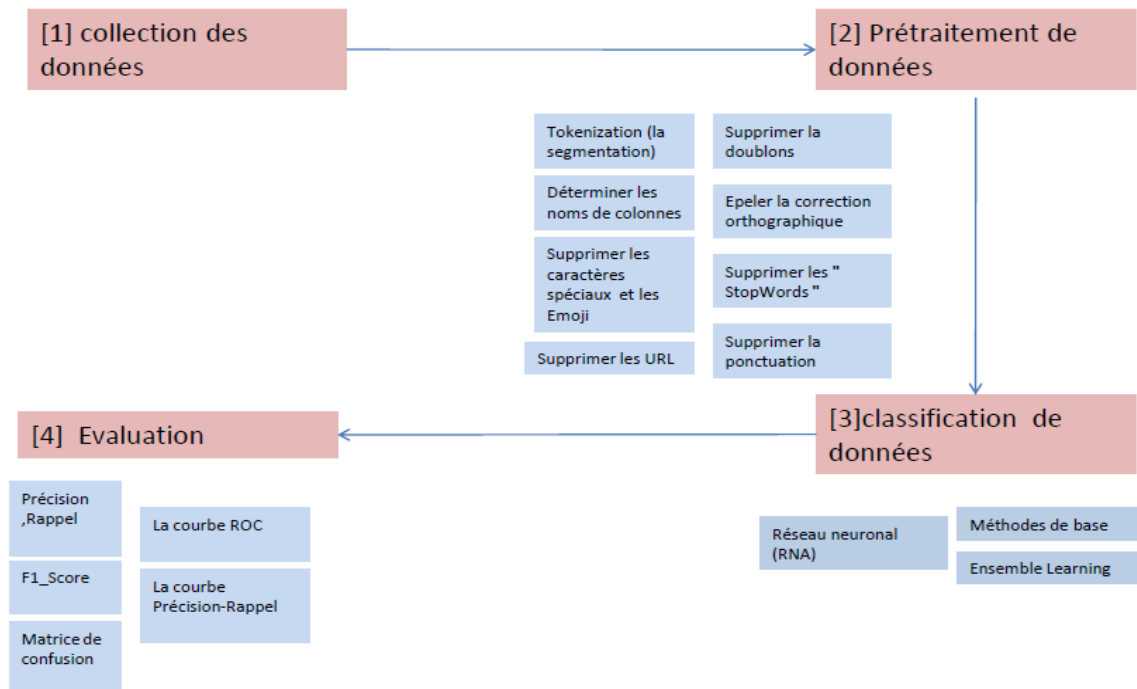
En menant une analyse comparative approfondie, ce projet vise à fournir des informations précieuses sur les performances des méthodes de classification traditionnelles et des réseaux neuronaux sur des ensembles de données, aidant ainsi les praticiens à prendre des décisions éclairées lors de la conception de systèmes de classification.

L'ensemble de données "Amazon Reviews Multi" comprend plus de 1200000 commentaires collectés depuis la plateforme Kaggle, Ces commentaires contiennent des avis d'utilisateurs sur des produits vendus sur Amazon.

## 3.3. Description de la méthodologie de conception :

Les méthodologies d'analyse de sentiments peuvent varier en fonction des données et des objectifs visés. En général, ce processus se décompose en plusieurs étapes.

1. Collection de données.
2. Prétraitement et préparation de données pour l'étape suivante.
3. Classification de texte avec différent méthode d'apprentissage
4. Evaluation des résultats obtenus.



**Figure 4:** Architecture de system proposé

### 3.3.1. Collection de données :

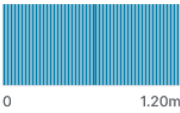

L'ensemble de données "Amazon Reviews Multi" comprend plus de 1200000 commentaires collectés depuis la plateforme Kaggle selon url suivant :

[//www.kaggle.com/datasets/mexwell/amazon-reviews-multi?select=train.csv](https://www.kaggle.com/datasets/mexwell/amazon-reviews-multi?select=train.csv)

Nous avons utilisé une quantité de 400 00 commentaires parmi un total de 12 000 00, dont 200 00 en français (100 00 positifs et 100 00 négatifs) et 200 00 en anglais (100 00 positifs et 100 00 négatifs) Puis nettoyé et préparé pour la classification de texte avec différent méthode d'apprentissage.

**train.csv** (346.24 MB) ↓ ↗ >

Detail Compact Column 9 of 9 columns ▾

#	Δ review_id	Δ product_id	Δ reviewer_id	# stars	Δ review_boc
id	review id	product id	reviewer id	starts	review body
	<b>1200000</b> unique values	<b>963043</b> unique values	<b>1098290</b> unique values		<b>1191</b> unique
0				1	
1.20m				5	
344	de_0240944	product_de_0341669	reviewer_de_0820307	1	Beim einleg MicroSD-Kar haben sich verhakt. Le Tage zu spä eine Rek...
345	de_0766742	product_de_0468064	reviewer_de_0013878	1	wurde kaput geliefert , instabil, n empfehlensw Probleme we Rücksendung
346	de_0184105	product_de_0936185	reviewer_de_0611022	1	Das Zelt is dünn vom Ma her. Die Nä reißen leic die Reißver

**Figure 5:**Fichier train.csv

L'ensemble de données est conçu pour une distribution équilibrée des cotes en étoiles, ce qui présente des avantages à des fins de classification. Cependant, cet équilibre peut entraîner un contournement ou une sous-représentations de certains types de langage par rapport à la distribution originale des avis. Les champs de données comprennent l'identifiant de l'avis, l'identifiant du produit, l'identifiant du critique, les étoiles allant de 1 à 5[voire la **figure 6**] (nous avons pris étoile 1 pour être le négatif (0) et étoile 5 pour être le positif (1)), le corps de l'avis, le titre de l'avis, six langues différentes[voire la **figure 7**] Nous avons pris que la langue français et la langue anglaise ,la catégorie de produit[voire la **figure 8**] et un fichier d'entraînement (train.csv) ,Cet ensemble de données reconnaît la photo prise par Towfiq barbhuiya sur Unsplash.

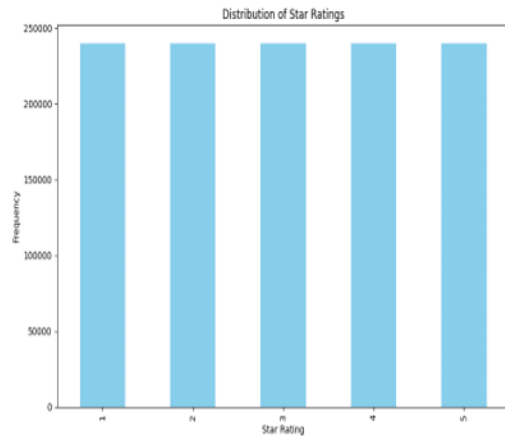


Figure 6: Diagramme à barres représentant les étoiles

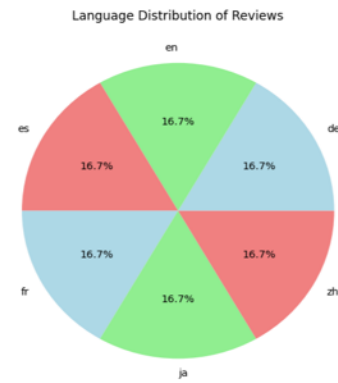


Figure 7: Un cercle relatif représentant les langues

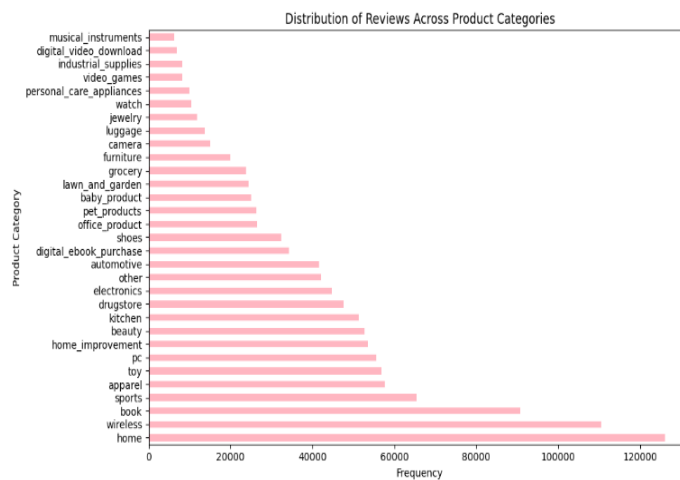


Figure 8: Répartition des avis selon les catégories de produits

### 3.3.2. Prétraitement de données :

Le prétraitement des données revêt une importance capitale dans la classification de texte, car il a pour objectif de supprimer tout bruit présent dans le texte analysé, comme les mots superflus et redondants, les chiffres, les préfixes et suffixes, etc. Un prétraitement efficace des données peut améliorer significativement les performances de classification, quel que soit le classificateur utilisé. Les étapes clés du prétraitement comprennent :

**-Déterminer les noms de colonnes :** est une fonction utilisée qui permet de modifier dynamiquement les noms de colonnes en fonction de la langue choisie, ce qui peut être utile pour rendre l'application plus conviviale pour les utilisateurs de différentes langues.



- **Tokenization (la segmentation)** : c'est une fonction qui découpe une chaîne de caractères en mots individuels, ou "tokens", afin de faciliter leur traitement et leur analyse. Cette fonction peut prendre en compte différents critères pour déterminer les limites entre les mots, tels que les espaces, les signes de ponctuation, les caractères spéciaux, etc. nous avons utilisé **word\_tokenize** (). De nombreuses bibliothèques, notamment **NLTK** et **TextBlob**, effectuent la tokenisation.

-**Supprimer les doublons** : une fonction qui nettoie les données en supprimant les doublons, en éliminant les valeurs manquantes et en supprimant les espaces vides inutiles autour des avis dans la colonne spécifiée. Cela permet d'avoir des données plus propres et prêtes à être utilisées pour une analyse ultérieure .nous avons utilisé **drop\_duplicates** () de bibliothèques **pandas**

-**Supprimer les caractères spéciaux et les Emojis** : une fonction qui permet de supprimer tous les emojis présents dans le texte, et les caractères tels que ~, %, \*, !, +, ", {}@

-**Epeler la correction orthographique** : une fonction qui permet de ajuster une correction orthographique à la langue choisie pour améliorer la qualité de la trame de données et la fonction linguistique des utilisateurs .Nous avons utilisé **correct\_spelling** de bibliothèques **pandas**.

-**Supprimer la ponctuation** : c'est une fonction qui supprime tous les signes de ponctuation d'une chaîne de caractères, tels que les virgules, les points, les points-virgules, les points d'exclamation, etc

-**Supprimer les URLs** : une fonction qui permet de supprimer tous les liens URL présents dans le texte.

-**Supprimer les "stopwords"** : une fonction qui permet d'éliminer les mots courants qui ne contribuent pas au sens du texte, comme "le", "la", "de", "un", "une", etc. Toutefois, il est important de noter que la liste des stopwords peut varier selon la langue utilisée. En Anglais , les stopwords courants incluent des mots tels que "a", "an", "the", "and", "but", "or", "of", "with", "in", "on", "at", "to".

Les stopwords en français sont utilisés à partir d'un fichier téléchargé qui contient les mots qu'il faut supprimer.

### 3.3.3. Extraction des caractéristiques :

Après avoir effectué le prétraitement des données, nous avons procédé à la vectorisation du texte en utilisant la technique TF-IDF, et le vecteur One Hot

#### 1. Term Frequency-Inverse Document Frequency (TF-IDF):

TF-IDF est une technique couramment utilisée en traitement de texte pour la vectorisation du texte. Elle permet de quantifier l'importance d'un terme dans un document en fonction de sa fréquence d'apparition et de sa rareté dans l'ensemble des documents, La formule de TF-IDF est donnée par [15]:

$$\text{TF-IDF} = \text{TF} * \text{IDF}$$

**TF (Term Frequency)** représente la fréquence du terme dans le document donné. Elle est calculée en divisant le nombre d'occurrences du terme dans le document par le nombre total de termes dans le document.

**IDF (Inverse Document Frequency)** mesure l'importance du terme dans l'ensemble de la collection de documents. Elle est calculée en prenant le logarithme inverse de la fraction du nombre total de documents sur le nombre de documents contenant le terme.[15]

#### 2. Le vecteur One Hot :

Les vecteurs "One Hot" sont souvent utilisés en analyse de données et en machine learning pour encoder des variables catégorielles. Chaque catégorie est représentée par un vecteur binaire où un seul bit est activé à 1 pour indiquer la présence de la catégorie, et les autres bits sont mis à 0. Cette approche permet de traiter efficacement les données catégorielles dans les algorithmes d'apprentissage automatique en les convertissant en une forme adaptée à l'entrée des modèles. [16]

id	Couleur
1	Rouge
2	bleu
3	Vert
4	bleu

id	Couleur_rouge	Couleur_bleu	Couleur_vert
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

**Tableau 1:**Le principe de vecteur One Hot

### 3.4. Méthodes utilisées pour la classification du texte :

#### 3.4.1. Réseau neuronal (RNA):

Un réseau de neurones formels à temps discret est constitué de deux types d'éléments, à savoir les entrées du réseau et les neurones. Chaque neurone (qui est déterministe) agit comme un processeur non linéaire, généralement simulé sur ordinateur (parfois implémenté sous forme de circuit électronique). À chaque instant discret  $k$ , le neurone calcule son potentiel  $\mathbf{v}_i(\mathbf{k})$  et son activité  $\mathbf{z}_i(\mathbf{k})$  selon la méthode suivante [17]:

$$\mathbf{Z}_i(\mathbf{k}) = \mathbf{f}_i(\mathbf{v}_i(\mathbf{k})) \text{ ou } \mathbf{v}_i(\mathbf{k}) = \sum_{j \in P} \sum_{\tau=0}^{q_{ij}} \mathbf{C}_{ij} \tau \mathbf{Z}_j(\mathbf{k} - \tau)$$

Un réseau de neurones est conçu pour effectuer une tâche définie par le concepteur à l'aide d'un ensemble de valeurs d'entrée et d'un ensemble correspondant de valeurs désirées pour les activités de certains neurones du réseau, appelés neurones de sortie. Ces ensembles sont appelés "exemples d'apprentissage". Les neurones qui ne sont pas des neurones de sortie sont qualifiés de cachés.

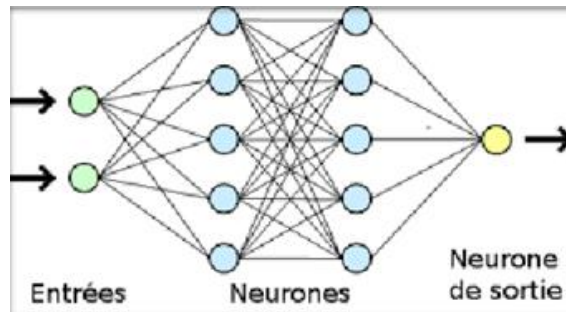
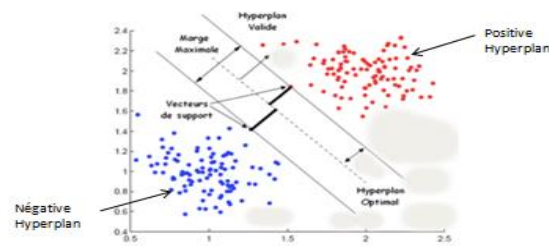


Figure 9: Architecture générale d'un réseau de neurones artificiels [17]

#### 3.4.2. Méthodes de base :

##### 1. Machine à vecteurs de support (SVM) :



**Figure 10:**Le principe de SVM [27]

Les SVM fonctionnent en cherchant une frontière de décision qui sépare les exemples d'une classe de ceux d'une autre classe. Cette frontière est choisie pour maximiser la marge, qui est la distance entre la frontière de décision et les exemples les plus proches de chaque classe. Voici les étapes de l'algorithme des SVM :

### **Algorithme 1 : les étapes de la méthode SVM**

1. représenter chaque point de données par un vecteur caractéristique dans un espace n-dimensionnel.
2. ajuster les caractéristiques pour qu'elles aient des plages similaires.
3. choisir un hyperplan aléatoire pour commencer l'optimisation.
4. identifier les points de données les plus proches de l'hyperplan, appelés vecteurs de support, et calculer la marge.
5. effectuer un processus d'optimisation pour trouver l'hyperplan optimal en minimisant l'erreur de classification et maximisant la marge.
6. utiliser le truc du noyau pour mapper les données dans un espace de dimensions supérieures si elles ne sont pas linéairement séparables.
7. choisir une fonction de noyau appropriée, telle que linéaire, polynomiale, RBF ou sigmoïde.
8. évaluer les performances du modèle en utilisant des mesures telles que l'exactitude, la précision, le rappel et le score F1

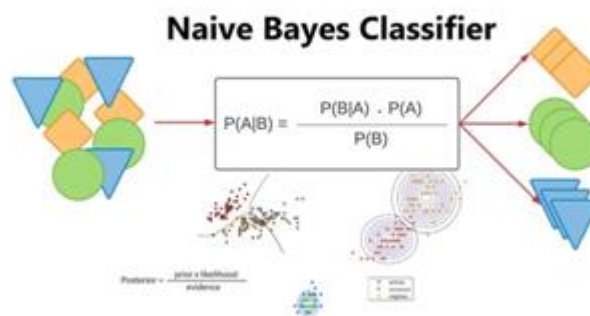
### **2. Régression logistique :**

La régression logistique est un algorithme de classification qui utilise une fonction logistique pour calculer la probabilité d'appartenance d'un échantillon à une classe donnée. La fonction logistique est une fonction en S qui prend une valeur d'entrée  $x$  et produit une sortie comprise entre 0 et 1. Voici les étapes de l'algorithme de la régression logistique:

## Algorithme 2: les étapes de la méthode régression logistique

1. Préparer l'ensemble de données étiqueté avec les caractéristiques et les étiquettes de classe binaire.
2. Mettre les caractéristiques à l'échelle pour s'assurer qu'elles ont des plages similaires.
3. Utiliser la fonction sigmoïde pour transformer la sortie en une valeur de probabilité entre 0 et 1.
4. Modéliser la probabilité du résultat binaire à l'aide de la fonction d'hypothèse qui applique la fonction sigmoïde à la combinaison linéaire de valeurs et de poids des caractéristiques.
5. Choisir une fonction de perte appropriée, comme la perte logistique, pour quantifier la différence entre les probabilités prévues et les étiquettes de classe réelles.
6. Estimer les valeurs optimales pour le vecteur de poids et le terme de biais en minimisant la fonction de perte à l'aide d'algorithmes d'optimisation comme la descente en gradient ou la méthode de Newton.
7. Établir une limite décisionnelle qui sépare les deux classes de l'espace de présentation.
8. Prédire la probabilité du résultat binaire pour les nouvelles données en appliquant les poids et le biais formés à la fonction de l'hypothèse et en classant en fonction d'un seuil choisi.

## 3. Naïve de Bayes :



**Figure 11:**Classification de Naïve Bayes [28]

Le Naïve Bayes est un algorithme de classification qui se base sur la théorie de Bayes. Il utilise les probabilités conditionnelles des caractéristiques (ou attributs) d'un échantillon pour déterminer la classe à laquelle il appartient. Le terme "naïf" fait référence à l'hypothèse simplificatrice selon laquelle les caractéristiques sont indépendantes les unes des autres, ce qui peut ne pas être vrai dans la pratique, mais qui permet des calculs plus simples et plus rapides. Voici les étapes de l'algorithme de Naïve Bayes :

### **Algorithme 3: les étapes de la méthode Naïve de Bayes**

1. Recueillir un ensemble de données étiqueté avec les caractéristiques et les étiquettes de classe correspondantes.
2. Calculer la probabilité antérieure de chaque classe dans l'ensemble de données.
3. Pour chaque caractéristique, calculer la probabilité d'observer cette caractéristique, compte tenu de chaque classe de l'ensemble de données.
4. Calculer la probabilité conditionnelle de chaque classe compte tenu des caractéristiques observées à l'aide du théorème de Bayes.
5. Faire l'hypothèse naïve que les caractéristiques sont conditionnellement indépendantes compte tenu de la classe. Bien que cette hypothèse soit souvent violée dans la réalité, Bayes naïve peut encore bien fonctionner dans la pratique.
6. Calculer la probabilité conjointe des caractéristiques observées pour chaque classe en multipliant les probabilités conditionnelles.
7. Calculer la probabilité postérieure de chaque classe en multipliant la probabilité conjointe par la probabilité antérieure de la classe.
8. Attribuer l'étiquette de classe ayant la probabilité postérieure la plus élevée comme classe prédite pour les nouvelles instances.

### **3.4.3. Apprentissage par ensemble (Ensemble Learning) :**

L'apprentissage par ensemble repose sur l'idée principale que la combinaison de plusieurs modèles peut réduire les biais, la variance et les erreurs de chaque modèle individuel. Voici les pseudo-codes des méthodes d'apprentissage par ensemble que nous avons implémentées dans notre application :

#### **1. Arbre de décision**

Est un modèle d'apprentissage automatique employé pour la classification et la régression, qui prédit la valeur cible en apprenant des règles de décision simples basées sur les caractéristiques des données. Voici les étapes de l'algorithme d'Arbre de décision :

#### **Algorithme 4: les étapes de l'algorithme Arbre de décision**

1. Préparation des données étiquetées avec les caractéristiques et les étiquettes de classe ou les valeurs cibles correspondantes.

2. Sélection de la meilleure fonction (caractéristique) dans l'ensemble de données pour servir de nœud racine de l'arbre de décision, basée sur des critères tels que le gain d'information ou l'impureté de Gini.

3. Division de l'ensemble de données en sous-ensembles en fonction de la fonction sélectionnée, chaque sous-ensemble représentant une valeur ou une plage unique de cette fonction.

4. Répétition récursive des étapes 2 et 3 pour chaque sous-ensemble, en choisissant la meilleure fonction à chaque niveau et en créant des sous-ensembles plus petits jusqu'à ce qu'un critère d'arrêt soit satisfait (par exemple, une limite de profondeur ou un nombre minimum d'échantillons dans un nœud).

5. Attribution d'une étiquette de classe ou d'une valeur prédite à chaque nœud de feuille en fonction de la classe majoritaire ou de la valeur moyenne des échantillons dans ce nœud.

6. Construction d'un arbre de décision complet qui peut être utilisé pour faire des prédictions sur de nouvelles données en parcourant l'arbre à partir de la racine en fonction des valeurs des caractéristiques, en suivant les branches appropriées jusqu'à atteindre un nœud de feuille et en retournant l'étiquette de classe ou la valeur prédite associée.

## **2. Forêt aléatoire:**

Est une méthode d'apprentissage automatique utilisée pour des tâches comme la classification et la régression. Elle repose sur la construction de plusieurs arbres de décision lors de l'entraînement.

### **Algorithme 5: les étapes de la méthode forêt aléatoire**

1. Préparer un ensemble de données étiqueté avec les caractéristiques et les étiquettes de classe ou les valeurs cibles correspondantes.

2. Sélectionner au hasard des sous-ensembles de l'ensemble de données (avec remplacement) pour créer plusieurs ensembles de données de formation, appelés échantillons bootstrap.

3. Construire un arbre de décision sur chaque échantillon bootstrap en utilisant un sous-ensemble aléatoire de caractéristiques à chaque nœud.

4. Pour les tâches de classification, chaque arbre prédit indépendamment l'étiquette de classe, et la classe avec la majorité des votes devient la prédiction finale. Pour les tâches de régression, les arbres prédisent une valeur continue, et la prédiction finale est souvent la moyenne ou la médiane des prédictions de chaque arbre.

5. Combiner les prédictions de tous les arbres pour faire une prédiction finale.

6. Calculer l'importance de chaque caractéristique en fonction de la mesure dans laquelle la précision diminue lorsque la caractéristique est permutée au hasard.

7. Optimiser le rendement du modèle de forêt aléatoire en réglant des hyper paramètres comme le nombre d'arbres, la profondeur des arbres et le nombre de caractéristiques considérées à chaque split.

8. Utiliser le modèle de forêt aléatoire pour faire des prédictions pour de nouvelles données invisibles en agrégeant les prédictions de tous les arbres.

### 3.5. Mesures des performances :

Les principales métriques d'évaluation couramment utilisées pour évaluer la performance d'un modèle de classification de texte sont les suivantes :

- **La précision** : représente la proportion de prédictions positives correctes parmi toutes les prédictions positives. Elle évalue la capacité du modèle à identifier correctement les exemples positifs et peut être calculée à l'aide de la formule [18]:

✓ **Précision**=True Positives / (True Positives+False Positives) ou

$$✓ \text{Précision} = \frac{TP}{TP+FP}$$

- **Le rappel** : également appelé "recall" en anglais, est la proportion de prédictions positives correctes par rapport à toutes les instances positives réelles. Il évalue la capacité du modèle à trouver tous les exemples positifs et peut être représenté par la formule [18] :

✓ **Rappel** =True Positives / (True Positives+ False Negatives) ou

$$✓ \text{Rappel} = \frac{TP}{TP+FN}$$

- **Le F1\_score** : Il s'agit d'une mesure harmonique de la précision et du rappel, qui combine les deux métriques en une seule valeur. Cette mesure peut être représentée par la formule : [16]

✓ **F1-Score**=2 \* (Precision\*Recall) / (Precision+Recall) ou

$$✓ \text{F1-Score} = \frac{2*P*R}{P+R}$$

- **La matrice de confusion** :utilisée pour visualiser la performance d'un modèle de classification, affiche le nombre de vrais positifs, de vrais négatifs, de faux positifs et de faux négatifs, et peut être représentée sous forme de tableau comme suit [18]:



	Class positive (Actual )	Class Negative (Actual )
Class Positive (Predicted )	True Positive (TP)	False Positive(FP)
Class negative (Predicted )	False Negative (FN)	True Negative (TN)

**Tableau 2:**Matrice de confusion

- **La courbe ROC** : qui évalue la performance d'un modèle de classification binaire en traçant le taux de vrais positifs par rapport au taux de faux positifs pour divers seuils de classification, peut être représentée par la formule suivante [18] :

$$\checkmark \text{ TPR( True Positive Rate)} = \frac{TP}{TP+FN}$$

$$\checkmark \text{ FPR( False Positive Rate)} = \frac{FP}{TN+FP}$$

- **La courbe Précision-Rappel** : est employée pour évaluer l'efficacité d'un modèle de classification binaire en représentant la précision par rapport au rappel pour divers seuils de classification. Elle s'avère utile lorsque la répartition des classes n'est pas équilibrée [18].

### 3.6. Conclusion :

Au cours de ce chapitre, nous avons exposé les objectifs de notre projet ainsi que la conception de notre système. Nous avons également présenté les méthodes implémentées dans ce projet, ainsi que les métriques d'évaluation de la qualité de classification.

# Chapitre 04 : Implémentation

## 4.1. Introduction :

Ce chapitre nous détaillerons l'environnement de développement utilisé ainsi que les outils et bibliothèques intégrées pour la création du système de classification.

## 4.2. Environnement et outils d'implémentation :

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

### 4.2.1. Matériel :

Caractéristiques	Poste de travail N°01	Poste de travail N°02
PC	Fujitsu	ASUS
Système d'exploitation	Windows 10 Professionnel	Windows 7 Professionnel
Processeur	Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz 2.70 GHz	Intel(R) Core (TM) i5-2430M CPU @ 2,40 GHz 2,40 GHz
RAM	8,00 Go	4,00 Go
Type de système	SE 64 bits	SE 64 bits

Tableau 3:Caractéristiques des matériels utilisés

### 4.2.2. Langage de programmation :

Nous avons utilisé le langage de programmation Python, version 3.7.3 , pour atteindre notre objectif, ce dernier est un langage de programmation adaptable, polyvalent , gratuit et c'est un langage raisonnablement simple à apprendre et très efficace qui permet aux développeurs de logiciels de fournir des solutions informatiques. Le langage python a été créé par Guido van Rossum et il est rendu public en 1991. C'est un langage open source qui ne cesse d'évoluer depuis sa création.

### 4.2.3. Environnement de programmation :

**Visual Studio Code** : est un éditeur de code source léger mais puissant qui fonctionne sur votre ordinateur et est compatible avec Windows, macOS et Linux. Il offre une prise en charge intégrée de JavaScript, TypeScript et Node.js, ainsi qu'un large éventail d'extensions pour d'autres langages et environnements d'exécution tels que C++, C#, Java, Python, PHP, Go et NET. [19]

**Google Colaboratory** : ou Colab, un outil Google simple et gratuit qui permet d'améliorer 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. [20]

### 4.2.4. Les principaux package Python utilisés :

Voici quelques packages que nous avons utilisés pour mener à bien ce projet :

**Bibliothèque NLTK** : Natural Language Toolkit est une bibliothèque open source pour le langage de programmation Python, initialement développée par Steven Bird, Edward Loper et Ewan Klein dans le but d'être utilisée dans le domaine du développement logiciel et de l'éducation. Elle est largement utilisée pour la création de programmes Python qui traitent des données en langage humain, notamment dans le domaine du traitement automatique du langage naturel (NLP). [21].

**Pandas** : est une bibliothèque open-source pour Python qui propose des outils et des structures de données pour l'analyse de données, spécifiquement pour la manipulation de données tabulaires et des séries temporelles. Cette bibliothèque fournit des fonctionnalités pour le traitement et l'analyse des données, comme la sélection, le filtrage, le tri, la fusion, la transformation et l'agrégation des données. [22]

**Numpy** : est une bibliothèque open source Python créée en 2005 par Travis Oliphant, qui est utilisée pour manipuler des tableaux. Elle propose également des fonctions pour effectuer des opérations dans le domaine de l'algèbre linéaire, notamment le calcul matriciel. [23]

**Matplotlib**: est une bibliothèque Python largement adoptée pour la représentation visuelle des données. Elle est couramment utilisée pour générer des graphiques, des diagrammes et des visualisations interactives. En outre, Matplotlib propose la possibilité de créer des nuages de

mots en s'appuyant sur la bibliothèque wordcloud, ce qui permet de produire des représentations visuelles attrayantes à partir de texte. [24]

**Wordcloud:** le nuage de mots, également appelé Word Cloud, est une illustration graphique des termes les plus courants dans un corpus de texte. Les mots les plus fréquents sont généralement affichés en caractères plus grands et en avant-plan, tandis que les mots moins fréquents sont présentés en caractères plus petits est en arrière-plan. [25]

### **4.3. Analyse exploratoire des données multilingues :**

Après l'application des étapes de prétraitement sur les deux bases de données anglaise et française, le nombre de commentaire a été réduit comme le montre le tableau suivant :

<b>Evénements multilingues</b>	<b>Nombre du commentaire avant le prétraitement</b>	<b>Nombre du commentaire après le prétraitement</b>
Anglais	20000	19975
Français	20000	19933

**Tableau 4:**Nombre du commentaire avant et après le prétraitement

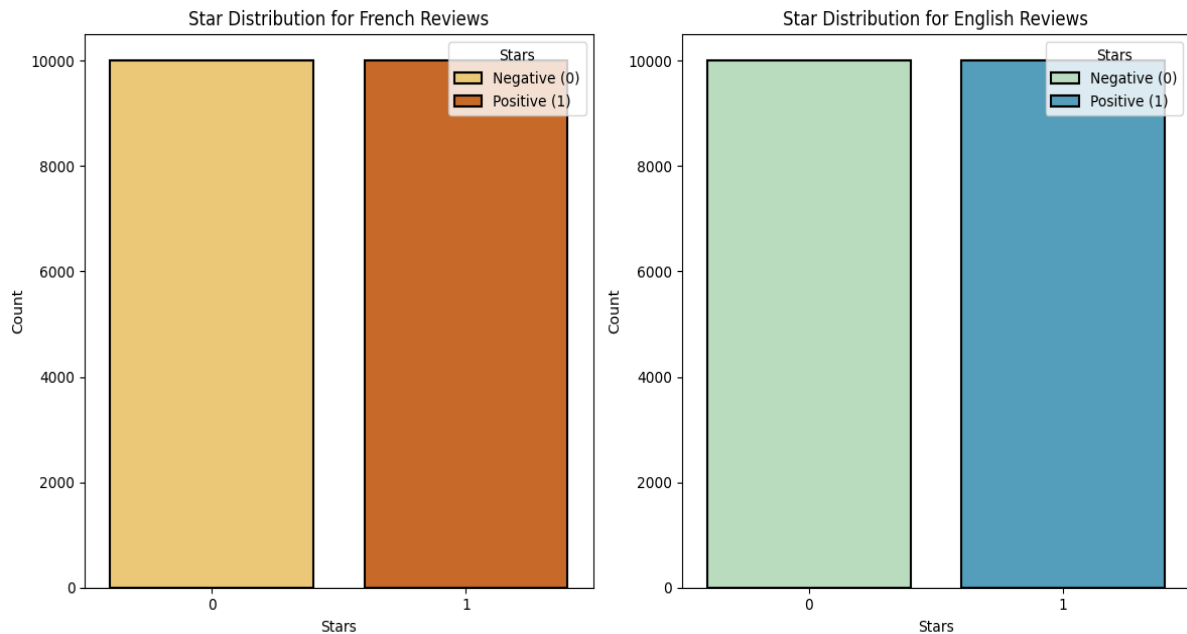
### **4.4. Génération du nuage de mots :**

En utilisant la bibliothèque Wordcloud, nous générons un nuage de mots mettant en avant les 100 termes les plus importants. Les mots les plus courants sont affichés en plus grand dans le nuage, reflétant ainsi leur fréquence d'utilisation.

Après l'application de cette bibliothèque sur les deux bases de données, nous avons obtenu les nuages suivants :



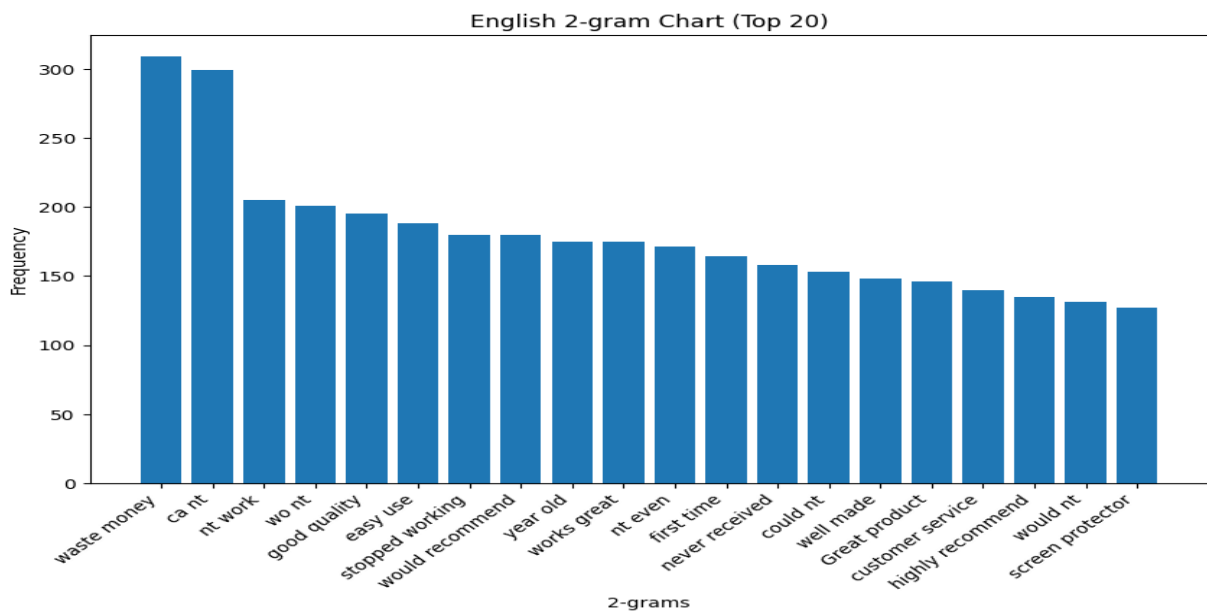
## 4.5. Diagrammes à bandes :



**Figure 14:**Diagramme de bandes de l'ensemble des données

- Ce diagramme présente la répartition des étoiles, les deux bases des données anglaise et française étant équilibrées, où l'étoile 0 représente les avis négatifs et l'étoile 1 les avis positifs.

## 4.6. Diagramme de n gramme :



**Figure 15:**Diagramme anglais de 2 gramme (top 20)

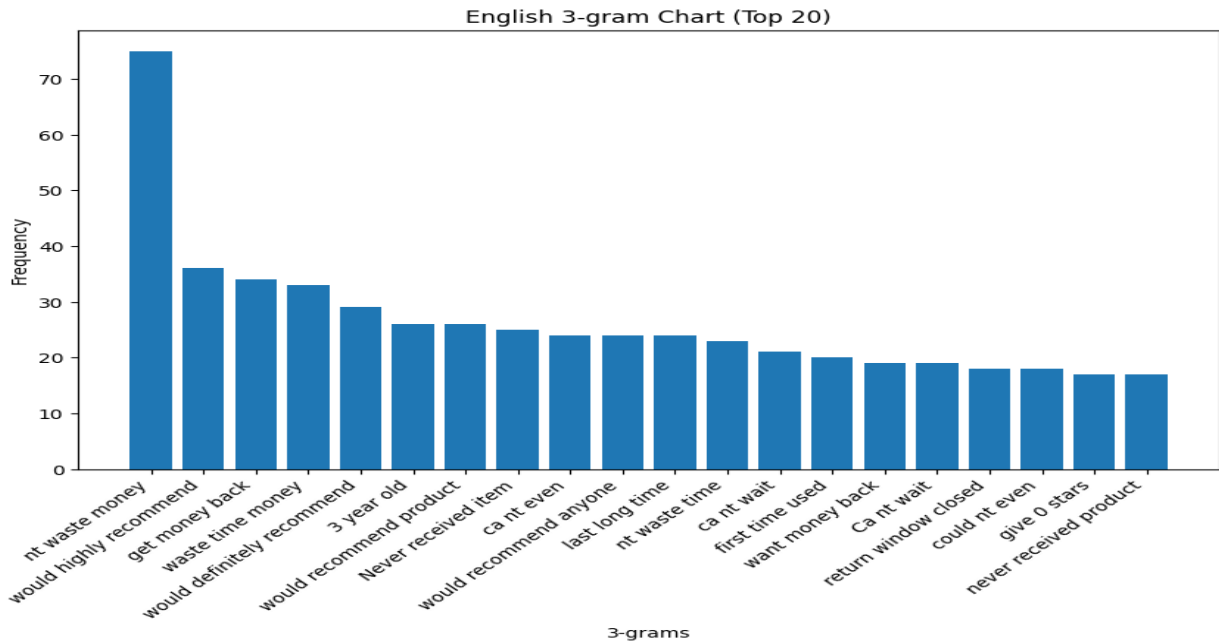


Figure 16:Diagramme anglais de 3 gramme (top 20)

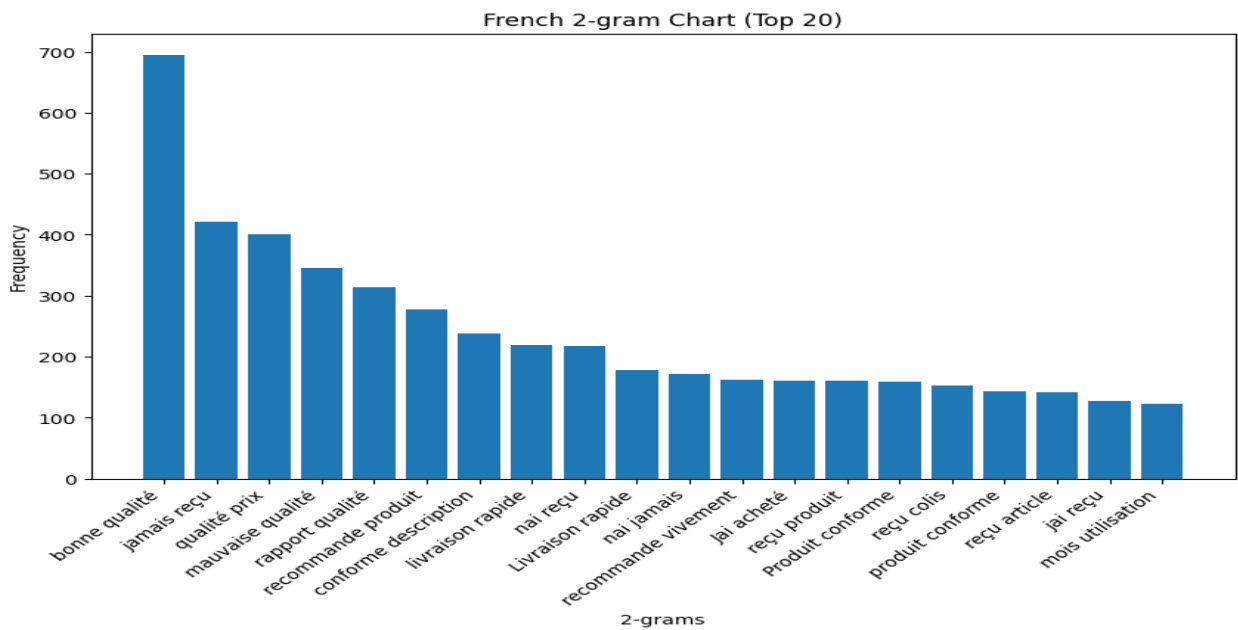
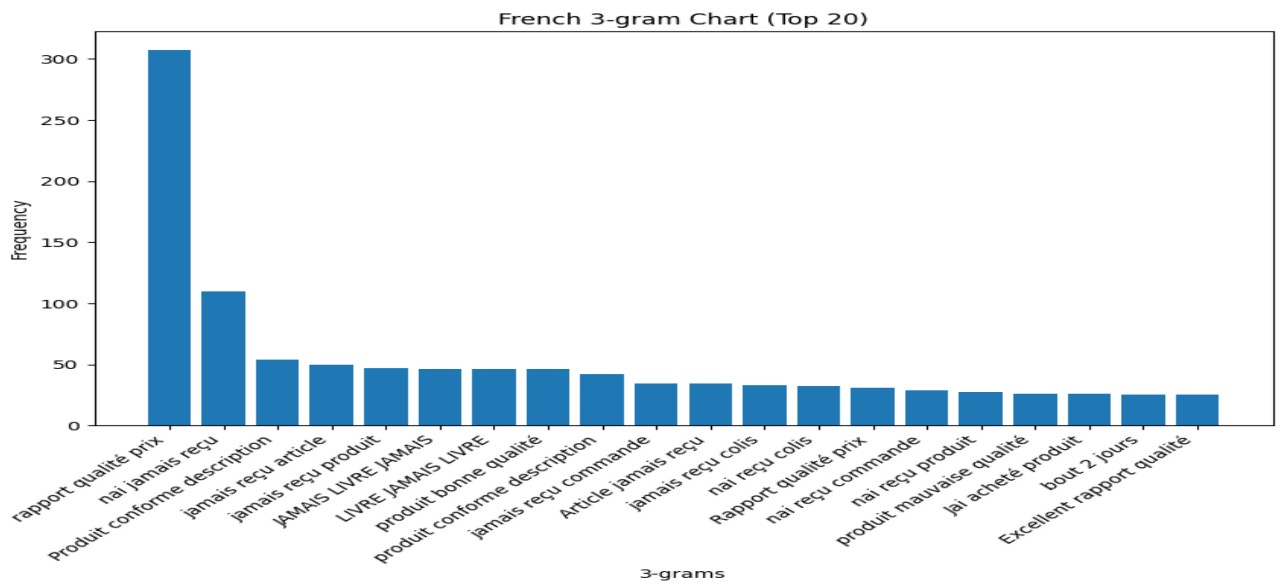


Figure 13:Diagramme français 2gramme (top 20)



**Figure 14:**Diagramme français 3 gramme (top 20)

- Les diagrammes montrent la fréquence des 20 bi- et trigrammes les plus fréquents extraits des avis clients sur un produit vendu sur Amazon. Les diagrammes sont listés dans la première colonne et leurs fréquences correspondantes dans la seconde. La fréquence de chaque diagramme est représentée par la taille de la barre correspondante dans le diagramme. Plus la barre est grande, plus le bi- et trigramme est fréquent.

#### 4.7. Application des algorithmes de classification de données :

- Dans notre application Une proportion 80% des données ont été allouées à l'ensemble d'apprentissage, tandis que le reste a été réservé pour tester et évaluer les performances des modèles de classification appliqués.



**Figure 19:**Découpage d'un ensemble de données en 80% pour l'entraînement et 20% pour le test

- Les fonctions utilisées pour préparer et nettoyer les données brutes ont été mentionnées précédemment dans le chapitre 3 section 3.2 prétraitement de données. Ces fonctions comprennent la suppression de caractères spéciaux et des Emojis, la suppression des "stopwords", la suppression des URLs, les tableaux ci-dessus nous fournissent un aperçu des résultats obtenus après certains de ces processus :

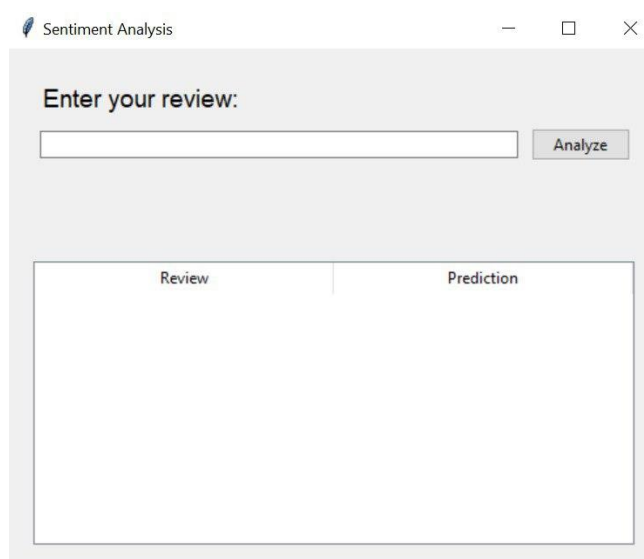


Texte brute	Texte après l'application des fonctions
"Super cadeau pour mes belles filles, elles adorent !"	Super cadeau belles filles adorent
Il aurait déjà fallu que je reçois le lustre. Il est jamais arrivé chez moi 😞	déjà fallu reçois lustre jamais arrivé
DON'T BUY IT!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! !!!It doesn't work at all. Missed the window to return it. Waste of Money. Junk	do notbuy not work Missed window return Waste Money Junk
They advertise that you can roll the hat-- -and you can!	advertise roll hat

**Tableau 5:**Exemples des commentaires après l'application des fonctions de prétraitement

#### 4.8. Génération d'une interface des résultats :

On a généré une interface qui montre une fenêtre d'analyse des sentiments. La fenêtre est vide, à l'exception des mots "Entrez votre avis", "Prédiction". Sur l'interface se trouve, un bouton "Analyser" et un tableau de résultats. Cela suggère que l'utilisateur peut saisir un avis dans la fenêtre et que l'outil analysera l'avis pour déterminer son sentiment.



**Figure 20 :** Une fenêtre d'analyse des sentiments.

## **4.9. Conclusion :**

Ce chapitre se focalise principalement sur la mise en œuvre et la réalisation de notre projet afin d'atteindre l'objectif de notre étude. Nous aborderons les outils utilisés pour le développement du projet, tels que le langage de programmation choisi, l'environnement de programmation et le matériel utilisé, Ainsi que la préparation de données pour la classification

# Chapitre 05 : Résultats et discussions

## 5.1. Introduction :

Dans ce chapitre nous présenterons et discuterons des résultats obtenus lors de nos expérimentations.

## 5.2. Les résultats obtenus :

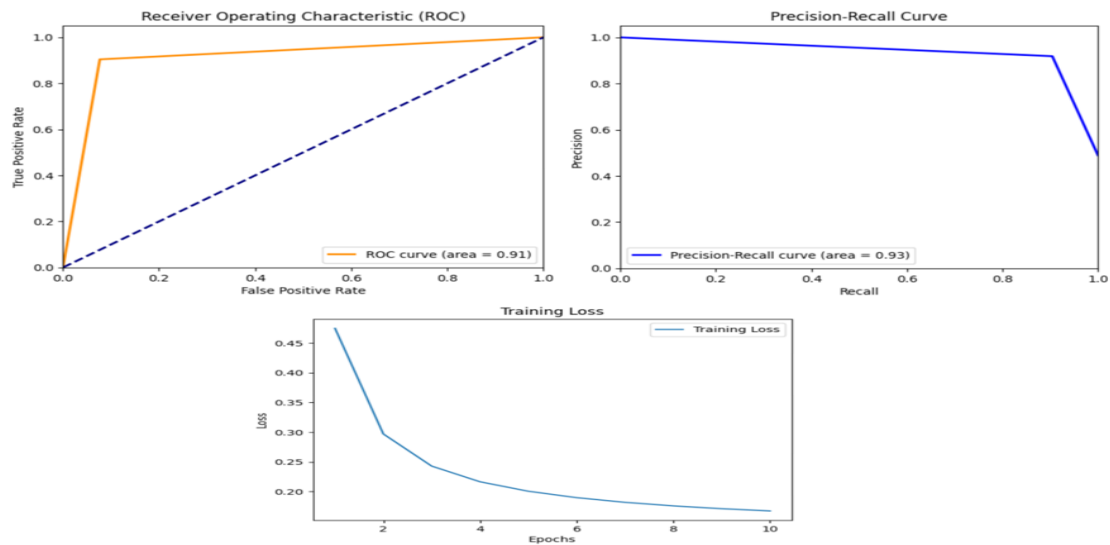
### 5.2.1. Réseau neuronal (RNA) :

Le Tableau 6 présente les performances des classificateurs de Réseau neuronal évaluées à l'aide des métriques de précision, rappel (recall), score F1 et exactitude (accuracy) avec les deux paramètres de vectorisation de texte : TF-IDF et One-hot vecteur

	Accuracy	Précision	Recall	F1_score
TF_IDF	0.91	0.91	0.91	0.91
One-hot vecteur	0.90	0.91	0.91	0.91

**Tableau 6:**Le résultat de performance de Réseau neuronal avec tf-idf et one-hot vecteur

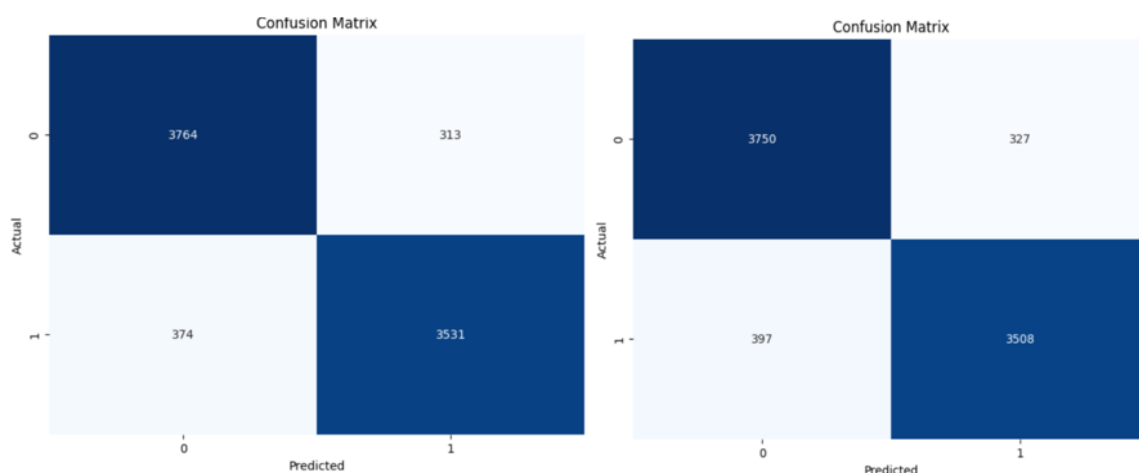
- Nous avons comparé les performances de classification en termes d'exactitude (Accuracy) ainsi que les mesures de précision, rappel (recall) et score F1, on constate que les valeurs sont presque égal avec les méthodes de vectorisation tf-idf et one-hot vecteur



**Figure 21:** Courbe ROC (gauche) et courbe PRC (droite) et courbe de Training loss de Réseau neuronal avec tf\_idf et one-hot vecteur

- La figure 21 met en lumière la performance de Réseau neuronal évaluée à l'aide des courbes **ROC**, **PRC** (courbe Précision-Rappel) et de la courbe de **Training loss**. Visuellement, la courbe **ROC** a une aire de 0,91, tandis que la courbe de **précision-rappel** a une aire de 0,93. Cela signifie que les deux modèles sont performants, mais que le modèle de précision-rappel est légèrement meilleur que le modèle ROC.

La courbe de **training loss** montre également que la perte de formation a été réduite au minimum par époque. Cela signifie qu'à mesure que le réseau neuronal apprenait davantage des données d'entraînement, la perte d'entraînement a continué à diminuer, mais à un rythme plus lent.



**Figure 22:** La matrice de confusion du Réseau neuronal avec tf\_idf(gauche) et avec one-hot vecteur (droite)

- La figure 22 illustre la matrice de confusion de la méthode de Réseau neuronal, dans cette matrice on remarque que le modèle avec tf\_idf a prédit 3764 instances True négative (TN), parmi lesquelles 313 étaient False positive (FP), et a également prédit 374 instances false négative (FN), parmi lesquelles 3531 étaient True positive (TP).

Et avec one-hot vecteur a prédit 3750 instances True négative (TN), parmi lesquelles 327 étaient False positive (FP), et a également prédit 397 instances false négative (FN), parmi lesquelles 3508 étaient True positive (TP). C'est pour cela on peut juger que cette matrice est la meilleure parmi les méthodes de classification

### 5.2.2. Les Cinq Méthodes de base de classification :

Voici une présentation des résultats obtenus dans les tableaux 7 et 8 avec les différentes méthodes de classification (TF-IDF et One-hot vecteur) en utilisant les algorithmes :

- Naïve de Bayes
- SVM
- Régression logistique
- Arbre de décision
- Forêt aléatoire

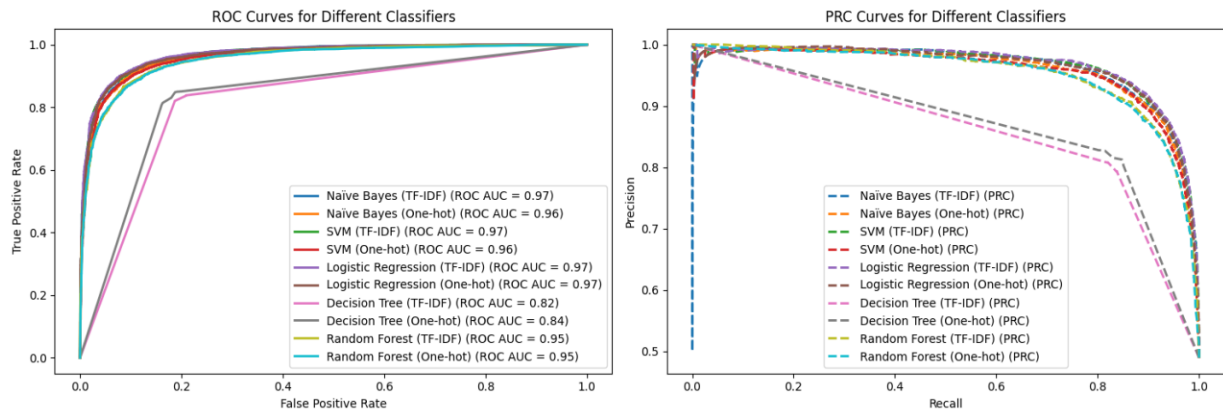
Métriques d'évaluation	Accuracy	Precision	Recall	F1-score
Naïve de Bayes	0.90	0.90	0.90	0.90
SVM	0.91	0.91	0.91	0.91
Régression logistique	0.91	0.91	0.91	0.91
Arbre de décision	0.81	0.81	0.81	0.81
Forêt Aléatoire	0.89	0.89	0.89	0.89

**Tableau 7:** Comparaison des résultats de performance des méthodes de bases avec tf-idf

Métriques d'évaluation	Accuracy	Precision	Recall	F1-score
Naïve de Bayes	0.90	0.90	0.90	0.90
SVM	0.90	0.90	0.90	0.90
Régression logistique	0.91	0.91	0.91	0.91
Arbre de décision	0.82	0.83	0.82	0.82
Forêt Aléatoire	0.89	0.89	0.89	0.89

**Tableau 8:** Comparaison des résultats de performance des méthodes de bases avec one-hot vecteur

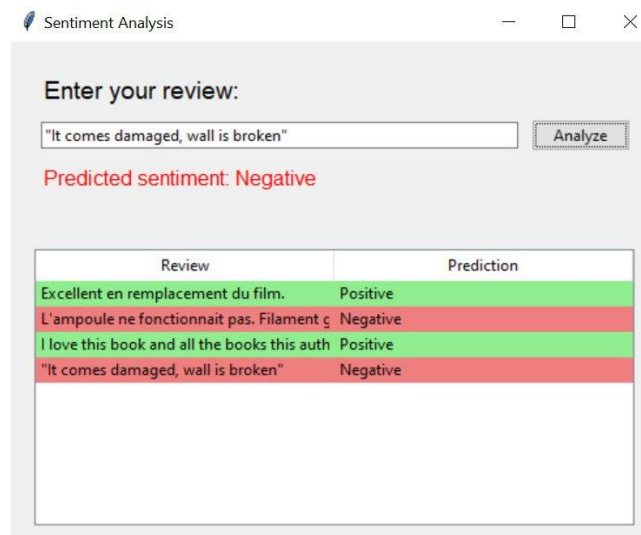
- Nous avons comparé les performances de classification en termes d'exactitude (Accuracy) ainsi que les mesures de précision, rappel (recall) et score F1, nous avons observé que les valeurs 0.90, 0.91 de tableau de la méthode tf-idf sont très proches et la valeur 0.81 est un peu petite que les autres, et dans le tableau de méthode one hot vecteur nous avons observé que la valeur plus grande est 0.91 et la valeur 0.82 la plus petite



**Figure 23:** Courbe ROC (gauche) et courbe PRC (droite) des méthodes de base avec tf\_idf et one-hot vecteur

- La Figure 23 met en lumière la performance des méthodes de base évaluée à l'aide des courbes ROC et PRC (courbe Précision-Rappel). On constate visuellement que l'évaluation de la courbe ROC révèle un AUC élevé de 0,97 avec la méthode Naïve de Bayes, SVM et Régression logistique utilisant TF-IDF, ainsi qu'un AUC de 0,97 avec la méthode Régression logistique utilisant One-hot vecteur. En revanche, on observe des valeurs plus faibles d'AUC, soit 0,82 avec TF-IDF et 0,84 avec One-hot vecteur, pour la méthode Arbre de décision.

### 5.3. Résultat d'exécution :



**Figure 24:** Résultat d'exécution des commentaires

- La figure 24 montre une capture d'écran d'une fenêtre d'analyse des sentiments. Le tableau de résultats affiche l'avis, la prédiction de sentiment et le sentiment réel. La

plupart des avis sont positifs. Cela suggère que les utilisateurs sont satisfaits des produits ou services qu'ils ont achetés. Cependant, il y a aussi avis négatifs. Ces avis devraient être examinés de plus près pour identifier les problèmes potentiels

#### **5.4. Discussion des résultats :**

En examinant les résultats des tableaux 6, 7 et 8 nous avons cherché à déterminer l'approche la plus efficace pour la classification. Les résultats obtenus en utilisant le modèle Réseau neuronal, SVM et Régression logistique avec la méthode TF-IDF ont généralement été légèrement supérieurs à ceux obtenus avec Naïve de Bayes, l'Arbre de décision et les Forêt Aléatoire. De plus, le modèle Régression logistique avec la méthode one hot vecteur a généralement surpassé les autres méthodes.

Cependant, la performance globale reste modérée, avec des scores de précision, de rappel et de F1 variant d'une classe à l'autre et d'une langue à l'autre. La comparaison révèle que les performances des différentes méthodes de classification dépendent de la langue et de la méthode utilisée.

Les résultats indiquent que le Réseau neuronal, le SVM et la Régression logistique avec TF-IDF tendent à mieux performer pour la classe Accuracy par rapport aux méthodes One-hot vecteur. Cela suggère que ces trois modèles sont plus efficaces pour capturer les caractéristiques discriminantes spécifiques à cette tâche de classification.

Dans l'ensemble, les trois tableaux montrent des performances de classification allant de modérées à bonnes, avec des taux d'exactitude compris entre 0,81 et 0,91. Ces résultats suggèrent que les méthodes et algorithmes utilisés peuvent fournir des résultats significatifs dans la tâche de classification.

#### **5.5. Conclusion :**

Dans ce chapitre, nous avons exposé tous les tests expérimentaux réalisés sur deux ensembles de données linguistique afin d'obtenir le meilleur classificateur possible pour identifier et évaluer les commentaires. Cette évaluation nous aide à prendre des décisions concernant un produit ou un service.



## **Conclusion générale :**

Dans le contexte numérique contemporain, marqué par des échanges et des interactions à l'échelle mondiale, la manipulation de données multilingues combinée à l'analyse des sentiments est devenue essentielle pour appréhender les opinions et les réactions des individus à travers le globe.

Ce travail de recherche a examiné diverses méthodes de classification telles que le Réseau neuronal , SVM ,Régression logistique , Naïve de Bayes , Arbre de décision ,et les forêts aléatoires, ainsi que deux approches de vectorisation, TF-IDF et One-Hot vecteur, pour analyser des données provenant de la plateforme Amazon. Les résultats obtenus ont révélé que malgré les différences entre ces techniques, elles ont généré des performances comparables dans la classification de ces données. Cette observation suggère que la sélection de la méthode de classification et de vectorisation peut être adaptée en fonction des besoins spécifiques du projet sans compromettre significativement les performances. De plus, cette étude souligne l'importance d'explorer et de comparer différentes approches pour obtenir des résultats robustes et fiables dans le traitement des données de grande taille et variées, telles que celles provenant d'Amazon. Ces résultats offrent des perspectives précieuses pour les praticiens et les chercheurs dans le domaine de l'apprentissage automatique et de l'analyse de données, en mettant en lumière les options disponibles pour traiter efficacement les données de commerce électronique à grande échelle.

Grâce à ces méthodes, les chercheurs et les entreprises peuvent mieux comprendre les opinions et les attitudes des utilisateurs à partir de différentes sources de données textuelles, ce qui facilite la prise de décisions éclairées et l'adaptation des stratégies en conséquence.

## **Perspectives et travaux futurs :**

En ce qui concerne les perspectives et les travaux futurs, plusieurs pistes peuvent être envisagées. Tout d'abord, une approche intéressante aurait été d'explorer l'utilisation d'un petit ensemble de données en comparaison avec le vaste ensemble de données consulté. Cette démarche aurait permis d'observer comment les méthodes étudiées interagissent avec la quantité d'informations disponibles, offrant ainsi des insights précieux sur l'efficacité des différentes approches dans des contextes de données variés.

De plus, une autre piste de recherche aurait été de reconfigurer les étoiles d'évaluation utilisées. Par exemple, il aurait été pertinent de considérer les étoiles 1 et 2 comme des évaluations négatives, tandis que les étoiles 4 et 5 auraient été considérées comme des évaluations positives, et ainsi de suite pour les autres étoiles intermédiaires. Cette réorganisation des critères d'évaluation aurait pu fournir une perspective différente sur les performances des méthodes étudiées et aider à mieux comprendre leur impact sur la qualité des résultats obtenus.

## Les références :

- [1] M.Boussaha ,M.boudiaf,« Moodle forum text mining » Mémoire de Master, Département de l'informatique, Université de Ibn khaldoun - Tiaret, 2020-2021. Consulté le: 22 janvier 2024
- [2] « namata-inbook09.pdf ». Consulté le:22 janvier 2024. [En ligne]. Disponible sur: <https://linqs.org/assets/resources/namata-inbook09.pdf>
- [3] « (PDF) A Brief Survey of Text Mining ». Consulté le:24 janvier 2024. [En ligne]. Disponible sur:  
[https://www.researchgate.net/publication/215514577\\_A\\_Brief\\_Survey\\_of\\_Text\\_Mining](https://www.researchgate.net/publication/215514577_A_Brief_Survey_of_Text_Mining)
- [4] « Information\_behavior.pdf ». Donald O .Case « information Seeking Behavior :A Review of the Research Litterrature» 2012 Consulté le: 24 janvier 2024
- [5] Charu Aggarwal, Cheng XiangZhai. « Text Classification » 2012 Consulté le:02 février 2024
- [6] pemlaW.Jordan « annotation for artificial intelligence » Consulté le:02 février 2024
- [7] S. Sarawagi, *Information extraction*. In Foundations and trends in databases, no. 1.2007,3. Hanover, MA:Now Publishers, Inc, 2007. Consulté le: 05 février 2024
- [8] « Analyse de sentiment : suivre les émotions client », Qualtrics. Consulté le: 19 février 2024. [En ligne]. Disponible sur: <https://www.qualtrics.com/fr/gestion-de-l-experience/etude-marche/analyse-sentiment/> Consulté le:15 février 2024
- [9]B. Liu . Sentiment Analysis and Opinion Mining. University of Illinois at Chicago. June 2014 Consulté le:22 février 2024
- [10]S. Kolkur, G. Dantal, et R. Mahe, « Study of Different Levels for Sentiment Analysis ». Consulté le: 22 février 2024
- [11]« Twitter Sentiment Analysis in Real-Time », MonkeyLearn Blog. Consulté le: 19 février 2024. [En ligne]. Disponible sur: <https://monkeylearn.com/blog/sentiment-analysis-of-twitter/>
- [12]V. A. Kharde et P. S. Sonawane, « Sentiment Analysis of Twitter Data: A Survey of Techniques », *IJCA*, vol. 139, n° 11, p. 5-15, avr. 2016, doi: 10.5120/ijca2016908625. Consulté le:23 février 2024
- [13]W. Medhat, A. Hassan, et H. Korashy, « Sentiment analysis algorithms and applications: A survey », *Ain Shams Engineering Journal*, vol. 5, n° 4, p. 1093-1113, déc. 2014, doi: 10.1016/j.asej.2014.04.011. Consulté le:23 février 2024
- [14]I. Guellil, F. Azouaou, H. Saâdane, et N. Semmar, « Une approche fondée sur les lexiques d'analyse de sentiments du dialecte algérien », *Revue TAL : traitement automatique des langues*, 2017, Consulté le:29 février 2024. [En ligne]. Disponible sur: <https://hal.science/hal-02012130>

[15]M. Nassima et A. Maroua, « Traiter le problème de déséquilibre de données en apprentissage automatique : le cas de la détection automatique des attitudes envers les rumeurs politiques en ligne ». Consulté le:29 fevrier 2024

[16]« document.pdf ». Consulté le: 03 mars 2024. [En ligne]. Disponible sur: <https://theses.hal.science/tel-01207633/document>

[17]« Modélisation, Classification Et Commande Par Réseaux De Neurones : Principes Fondamentaux, Méthodologie De Conception Et Illustrations Industrielles ». Consulté le: 03 mars 2024. [En Ligne]. Disponible Sur: [https://www.researchgate.net/publication/242593091\\_Modelisation\\_Classification\\_Et\\_Commande\\_Par\\_Reseaux\\_De\\_Neurones\\_Principes\\_Fondamentaux\\_Methodologie\\_De\\_Conception\\_Et\\_Illustrations\\_Industrielles](https://www.researchgate.net/publication/242593091_Modelisation_Classification_Et_Commande_Par_Reseaux_De_Neurones_Principes_Fondamentaux_Methodologie_De_Conception_Et_Illustrations_Industrielles)

[18] S. Bird, E. Klein, and E. Loper, “Natural language processing with Python: analyzing text with the natural language toolkit. ‘ O’Reilly Media, Inc.,” 2009. Consulté le:03 mars 2024

[19] « Documentation for Visual Studio Code ». Consulté le: 10 mars 2024. [En ligne]. Disponible sur: <https://code.visualstudio.com/docs>

[20] « Google Colab ». Consulté le: 15 mars 2024. [En ligne]. Disponible sur: <https://research.google.com/colaboratory/faq.html?hl=fr>

[21] « What is the Natural Language Toolkit (NLTK)? - Definition from Techopedia ». Consulté le: 20 mars 2024. [En ligne]. Disponible sur: <https://www.techopedia.com/definition/30343/natural-language-toolkit-nltk>

[22] « Pandas : la bibliothèque Python dédiée à la Data Science ». Consulté le: 22 avril 2024. [En ligne]. Disponible sur: <https://datascientest.com/pandas-python-data-science>

[23] « NumPy - ». Consulté le: 08 avril 2024. [En ligne]. Disponible sur:<https://numpy.org/>

[24] « Matplotlib — Visualization with Python ». Consulté le: 22 avril 2024. [En ligne]. Disponible sur: <https://matplotlib.org/>

[25] Zygomatic, « Générateur de nuage de mots clés gratuit en ligne et Générateur de nuage de tags. », [nuagesdemots.fr](https://www.nuagesdemots.fr/). Consulté le: 08 avril 2024. [En ligne]. Disponible sur: <https://www.nuagesdemots.fr/>

[26] « 1-Etapes du processus d’extraction de... | Download Scientific Diagram ». Consulté le: 15 avril 2024. [En ligne]. Disponible sur: [https://www.researchgate.net/figure/Etapes-du-processus-dextraction-de-connaissancesaconnaisances-connaissancesa-partir\\_fig2\\_299368487](https://www.researchgate.net/figure/Etapes-du-processus-dextraction-de-connaissancesaconnaisances-connaissancesa-partir_fig2_299368487)

[27] « Les techniques algorithmiques de l’IA | Support vector machine », Projet AJC | ACT Project. Consulté le: 20 avril 2024. [En ligne]. Disponible sur: <https://www.ajcact.org/en/2020/12/17/les-techniques-algorithmiques-de-lia-svm/>

[28] « Classification Naïve Bayésienne - Naive Bayes classifier - YouTube ». Consulté le: 20 avril 2024. [En ligne]. Disponible sur: [https://www.youtube.com/watch?app=desktop&v=\\_izUmHzI3a0](https://www.youtube.com/watch?app=desktop&v=_izUmHzI3a0)

## Annexe :

Dans cette annexe, nous fournissons des extraits de code pertinents utilisés dans le cadre de notre étude.

### Extrait 1 : Prétraitement des données

```
if language == 'french' and lang == 'french':
    with open('datasets/stopwords-fr.txt', 'r', encoding='utf-8') as
f:
    custom_stopwords = set(f.read().split())
    stop_words =
set(stopwords.words('french')).union(custom_stopwords)
    elif lang == 'english':
        stop_words = set(stopwords.words('english'))

    df[review_column] = df[review_column].progress_apply(lambda x: '
.join([token for token in x.split() if token.lower() not in stop_words]))

    preprocessed_data[lang] = df

return preprocessed_data
```

- ✓ Cette partie du code vise à supprimer les mots vides (stopwords) des données d'une colonne spécifique dans le dataframe, en fonction de la langue spécifiée. Pour le français, le code lit le fichier stopwords-fr.txt, tandis que pour l'anglais, il utilise automatiquement les mots vides

```
generate_word_cloud(english_text, 'English Word Cloud (Top 100 Words)', max_words=100)
```

- ✓ Cette fonction **generate\_word\_cloud** prend en entrée un texte en anglais (english\_text), un titre pour le nuage de mots à générer ('English Word Cloud (Top 100 Words)'), et un paramètre optionnel max\_words=100 qui spécifie le nombre maximum de mots à afficher dans le nuage de mots.

```
filtered_df = train_df[['stars', 'review_body', 'language']]

filtered_df = filtered_df.dropna()

filtered_df = filtered_df[filtered_df['language'].isin(['fr', 'en'])]

filtered_df['stars'] = filtered_df['stars'].replace({1: 0, 5: 1})

filtered_df = filtered_df[filtered_df['stars'].isin([0, 1])]
```

- ✓ Ce code prépare les données en filtrant, nettoyant et transformant les colonnes nécessaires pour une tâche de classification binaire basée sur les évaluations d'étoiles des commentaires en français et en anglais.

### Extrait 2: classification des modèles

```
classifiers = {  
    'Naïve Bayes (TF-IDF)': MultinomialNB(alpha=1.0),  
    'Naïve Bayes (One-hot)': MultinomialNB(alpha=1.0),  
    'SVM (TF-IDF)': SVC(kernel='linear', probability=True),  
    'SVM (One-hot)': SVC(kernel='linear', probability=True),  
    'Logistic Regression (TF-IDF)': LogisticRegression(max_iter=1000),  
    'Logistic Regression (One-hot)': LogisticRegression(max_iter=1000),  
    'Decision Tree (TF-IDF)': DecisionTreeClassifier(),  
    'Decision Tree (One-hot)': DecisionTreeClassifier(),  
    'Random Forest (TF-IDF)': RandomForestClassifier(),  
    'Random Forest (One-hot)': RandomForestClassifier()  
}
```

- ✓ Ce code pour les classificateurs sont des modèles d'apprentissage automatique utilisés pour la classification de données textuelles. Chaque classificateur est associé à une méthode de représentation du texte (TF-IDF et one-hot vecteur) et à un algorithme spécifique

### Extrait 3 : entraîner le modèle

```
num_epochs = 10  
train_losses = []  
for epoch in range(num_epochs):  
    model.train()  
    total_loss = 0  
    with tqdm(total=len(X_train)) as pbar:  
        for i, x_batch in enumerate(X_train):  
            optimizer.zero_grad()  
            outputs = model(x_batch.unsqueeze(0))  
            loss = criterion(outputs.squeeze(), y_train[i])  
            loss.backward()  
            optimizer.step()  
            total_loss += loss.item()  
            pbar.set_description(f'Epoch {epoch+1}/{num_epochs}, Loss:  
{total_loss/(i+1):.4f}')  
            pbar.update(1)  
    train_losses.append(total_loss / len(X_train))
```

- ✓ Ce code est une boucle d'apprentissage pour entraîner le modèle 10 fois sur l'ensemble de données

# Résumé

Ce projet de fin d'études se concentre sur la gestion efficace du volume important de commentaires et d'avis multilingues de consommateurs pour les entreprises et les porteurs de projets. En se basant sur l'analyse des sentiments et le text mining, l'étude explore différentes approches, telles que les réseaux neuronaux, les SVM, la régression logistique, le Naïve de Bayes, les arbres de décision et les forêts aléatoires, pour traiter les données en français et en anglais. Une comparaison détaillée de ces méthodes est réalisée pour déterminer la plus adaptée à l'analyse des sentiments et au text mining multilingue. De plus, deux méthodes de numérisation distinctes, tf-idf et le codage one-hot vecteur, sont expérimentées pour évaluer leur efficacité dans l'analyse des données multilingues

**Mots-clés :** Fouille de texte, analyse de sentiments, réseaux neuronaux, classification, comparaison, prédiction, données multilingue.

# Abstract

This end-of-studies project focuses on the effective management of the large volume of multilingual consumer comments and reviews for companies and project leaders. Based on sentiment analysis and text mining, the study explores different approaches, such as neural networks, SVMs, logistic regression, Bayes Naive, decision trees and random forests, to process data in French and English. A detailed comparison of these methods is made to determine the most suitable for sentiment analysis and multilingual text mining. In addition, two distinct scanning methods, tf-idf and one-hot vector coding, are being tested to assess their effectiveness in analyzing multilingual data

**Keywords:** Text mining, sentiment analysis, neural networks, classification, comparison, prediction, multilingual data

## ملخص

يركز مشروع نهاية الدراسات هذا حول الإدارة الفعالة لحجم كبير من تعليقات المستهلكين واستعراضاتهم بمختلف اللغات للشركات وقادة المشاريع. من خلال تحليل المشاعر وتعدين النصوص، تبحث الدراسة في أساليب مختلفة مثل الشبكات العصبية، و SVMs، والانحدار اللوجستي، و Bayes Naive، وأشجار القرار والغابات العشوائية لمعالجة البيانات باللغتين الفرنسية والإنجليزية. يُجرى مقارنة دقيقة بين هذه الأساليب لتحديد الأنسب لتحليل المشاعر وتعدين النصوص بلغات متعددة. بالإضافة إلى ذلك، يتم اختبار طريقتين مميزتين للمسح، وهما ترميز ناقل واحد ساخن و tf-idf ، لتقييم فعالتهما في تحليل البيانات بلغات متعددة.

**الكلمات الرئيسية:** تعدين النصوص، تحليل المشاعر، الشبكات العصبية، مقارنة، متعدد اللغات

