MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE ABBES LAGHROUR KHENCHELA
FACULTE DES SCIENCES ET DE LA TECHNOLOGIE

Département de Mathématiques et informatique

N° de série : .....

Mémoire de fin d'études

Pour l'obtention du diplôme de Master (L.M.D)

Spécialité : informatique

Option : sécurité et Technologies web

THEME

Analyse de Sentiments des malades, vers un système d'assistance médicale

Réalisés par :

Ahlam Bouterai

Randa Benamara

Encadré par :

Dr. Hichem Rahab

Année Universitaire : 2021/2022

ملخص

مع التوسع الهائل في المعلومات على الإنترنت، يعبر المستخدمون في جميع أنحاء العالم يوميًا عن آرائهم على الشبكات الاجتماعية مثل (تويتر وفايس بوك) ومنتديات المناقشة والمواقع الإلكترونية. تستثمر الشركات الطبية اليوم في تحليل هذه الآراء من أجل تحسين منتجاتها وخدماتها. تسمى عملية التعرف على آراء المستخدمين المرضى حول المنتجات أو الخدمات، سواء كانت إيجابية أو سلبية، بتحليل المشاعر. تم اقتراح العديد من الأساليب لتحليل المشاعر وتستخدم معظم هذه الأساليب تقنيات التعلم الآلي. لذلك، في هذه الدراسة، نحاول تقديم نهج لتحليل المشاعر على المواقع الطبية. يعتمد هذا الحل المقترح على التقنيات المختلفة للتعلم الآلي الخاضع للإشراف مع عدة طرق: الة الدعم الموجه، نايف بايز، غابة أشجار القرار، أشجار القرار، ك الجار الأقرب. تم استخدام أربع مدونات، الأولى تم تجميعها خلال هذا العمل، بالنسبة للثلاث مدونات الأخرى، فقد تم استخراجها من مدونة كبرى متوفرة عبر الأنترنت. تم الحصول على أفضل النتائج باستخدام المدونة المجمعة خلال هذا العمل نظرا لمراعاة التوازن بين الفئتين، الإيجابية والسلبية، خلال إنشائها.

الكلمات المفتاحية: المشاعر الطبية، تحليل المشاعر، استخلاص الآراء، التعلم الآلي، الة الدعم الموجه، نايف بايز، غابة أشجار القرار، أشجار القرار، ك الجار الأقرب.

Résumé

Avec l'expansion spectaculaire de l'information sur Internet, les utilisateurs du monde entier expriment quotidiennement leur opinion sur les réseaux sociaux tels que (Facebook et Twitter), des forums de discussions et des sites web. Aujourd'hui les entreprises médicales investissent dans l'analyse de ces opinions afin d'améliorer leurs produits et services. Le processus de reconnaissance des opinions des patients sur les produits ou services, qu'elles soient positives ou négatives, est appelé analyse des sentiments. Plusieurs approches ont été proposées pour l'analyse des sentiments et la plupart de ces approches utilisent des techniques d'apprentissage automatique. Par conséquent, dans cette étude, nous proposons une approche pour l'analyse des sentiments sur les sites web médicaux. Cette solution proposée se base sur les différentes techniques de l'apprentissage automatique supervisé avec plusieurs méthodes à savoir ; les séparateurs à vaste marge (SVM pour Support Vector Machine), les voisins les plus proches (K-NN pour K-Nearest Neighbors), les arbres de décision (DT pour Decision trees), les forêts d'arbres décisionnels (RF pour Random Forest) et le Naïve Bayes (NB pour Naive Bayes). Pour l'expérimentation nous avons utilisé un corpus créer dans le cadre de ce mémoire en plus de trois autres corpus téléchargés à partir du Web. Les résultats du notre corpus dépassent les autres trois corpus vu le respect de l'équilibrage, entre les deux classes, lors de son création.

Mots clés : Sentiments médicaux, Analyse des Sentiments, fouille d'opinion, Apprentissage automatique, SVM, K-NN, DT, RF, NB.

Abstract

With the dramatic expansion of information on the Internet, users around the world express their opinions daily on social networks such as (Facebook and Twitter), discussion forums and websites. Today, medical companies are investing in the analysis of these opinions to improve their products and services. The process of recognizing the opinions of user-patients on products or services, whether positive or negative, is called sentiment analysis. Several approaches have been proposed for sentiment analysis and most of these approaches use machine learning techniques. Therefore, in this study, we try to propose an approach for sentiment analysis on medical websites. This proposed solution is based on different techniques of supervised machine learning with several methods: Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), Decision trees (DT), Random Forest (RF) and Naive Bayes (NB). In the experimental study, we have used a collected and created a dataset in this work in addition to three other datasets collected from the Web. The best obtained results were obtained using our own dataset for its equilibrium between the positive and negative classes.

Keywords: Medical sentiments, Sentiment Analysis, opinion mining, Machine Learning, SVM, K-NN, DT, RF, NB.

Remerciements

Tout d'abord, nous remercions Dieu tout-puissant, qui nous a donné des connaissances que nous ne savions pas. Nous le louons beaucoup digne de sa grandeur, et beaucoup de bénédictions.

Nous exprimons nos sincères remerciement et notre gratitude au docteur HICHEM RAHAB, qui nous avons été honorés de proposer et d'encadrer notre mémoire. Et il ne nous a épargné aucune information ni aucun conseil.

Nous sommes également heureux d'exprimer nos remerciements et notre gratitude aux membres du jury pour avoir accepté l'évaluation de notre mémoire.

A la fin du parcours universitaires, nous exprimons nos remerciements et notre gratitude aux honorables enseignants du département de mathématiques et informatique de l'université ABBES LAGHROUR KHENCHELA.

Nous remercions également tous ceux qui nous ont aidés de près ou de loin à mener à bien ce travail.

AHLAM&RANDA

Dédicace

Je dédie ce travail à :

A mes chers parents,

A mes frères et sœurs,

A tous mes chers amis,

A toute ma famille,

A tous mes professeurs et collègues,

A tous ceux qui m'ont conseillé et encouragé,

Ahlam

Dédicace

Je dédie ce travail à :

A ma mère,

A mon père,

A mon mari,

A mon frère et mes sœurs,

A tous mes chers amis,

A toute ma famille,

A tous mes professeurs qui m'ont enseigné au primaire, au collège, au lycée et

À l'université

RANDA

Table de matière

Introduction générale ---- 16
Contexte global de mémoire ---- 16
Problématique ---- 17
Objectifs de l'étude ---- 17
Organisation du mémoire ---- 18
Chapitre 1 : fouille de sentiments médicaux ---- 21
1 Introduction ---- 21
2 Analyse des Sentiments médicaux ---- 21
2.1 Explication de l'analyse des sentiments médicaux ---- 22
2.2 L'importance et l'objectifs de l'analyse des sentiments médicaux ---- 24
2.2.1 L'analyse des sentiments médicaux aide les prestataires à améliorer la communication avec les patients ---- 24
2.2.2 L'analyse des sentiments médicaux aide à quantifier les performances des employés et des services ---- 25
2.3 Les niveaux d'analyse des sentiments ---- 25
2.3.1 Niveau du document ---- 26
2.3.2 Niveau de la phrase ---- 26
2.3.3 Niveau des aspects ---- 26
3 Types d'analyse de sentiments ---- 26
3.1 Analyse fine des sentiments (fine-grained sentiments analysis) ---- 26
3.2 Détection d'émotion (Emotion detection) ---- 27
3.3 Analyse de sentiments à base d'aspects (Aspect-Based Sentiment Analysis ABSA) ---- 27
4 Sources des sentiments médicaux ---- 27
4.1 Analyse des sentiments sur les sites web ---- 27
4.2 Analyse des sentiments sur les réseaux sociaux ---- 28
4.2.1 Le réseau social twitter ---- 29
5 Les approches de l'analyse des Sentiments ---- 30
5.1 Approche à base d'apprentissage automatique ---- 30
5.2 Approche à base de règles (Rule-based) ---- 31
5.3 Approche hybride ---- 32
5.4 Les avantages et les inconvénients d'approches ---- 32
6 Conclusion ---- 33
Chapitre 2 : l'apprentissage automatique ---- 35

1 Introduction ---- 35
2 L'intelligence artificielle ---- 35
2.1 Termes associés à l'intelligence artificielle ---- 36
2.2 L'avantage de l'apprentissage automatique ---- 38
3 Les types de l'apprentissage automatique ---- 38
3.1 Apprentissage supervisé ---- 39
3.1.1 Régression linéaire ---- 40
3.1.2 Les Séparateurs à Vaste Marge (SVM) ---- 41
3.1.3 Arbre de décision (DT) ---- 42
3.1.4 Les Forêts d'arbres décisionnels (Random Forest RF) ---- 43
3.1.5 Bayésien naïf (Naïves Bayes) ---- 44
3.1.6 K plus proches voisins (KNN) ---- 44
3.1.7 Comparaison des algorithmes d'apprentissage supervisé ---- 45
3.2 Apprentissages non-supervisé ---- 47
3.2.1 Méthode des K-moyennes (K-means) ---- 49
3.2.2 Clustering hiérarchique ---- 49
3.2.3 Algorithme Apriori ---- 50
3.3 Apprentissage semi-supervisé ---- 51
3.4 Apprentissages par renforcement ---- 52
3.5 Conclusion ---- 52
Chapitre 3 : Analyse du sentiment des revues de médicaments ---- 54
1 Introduction ---- 54
2 Présentation des outils utilisés ---- 54
2.1 Le langage Python ---- 54
2.1.1 Le NLTK ---- 56
2.1.2 Pandas ---- 56
2.1.3 Scikit-learn ---- 57
2.1.4 Streamlit ---- 57
2.1.5 Jupyter notebook ---- 57
2.1.6 Spyder ---- 57
3 L'ensemble de données ---- 58
3.1 Création de l'ensemble de données Cymbalta_drug_dataset ---- 58
3.1.1 Description de l'ensemble des données Cymbalta_drug_dataset ---- 59
3.1.2 Annotation du corpus « Cymbalta_drug_dataset » ---- 60

3.2 L'ensemble de données drugsCom ---- 60
3.2.1 Description de l'ensemble des données drugsCom ---- 61
3.2.2 Annotation du corpus « drugsCom » ---- 61
4 Architecture de notre application ---- 62
4.1 Importation de l'ensemble de données ---- 63
4.2 Nettoyage des données ---- 63
4.3 Prétraitement des données ---- 63
4.3.1 Prétraitement préliminaire ---- 64
4.3.2 Séparation de mots (Tokenization) ---- 64
4.3.3 Suppression des mots vides (stop words removal) ---- 64
4.3.4 Enracinement (Stemming) ---- 65
4.4 Extraction des caractéristiques ---- 65
4.5 La Classification ---- 66
4.6 Mesures d'évaluation ---- 67
5 Implémentation ---- 68
5.1 La division (Train/Test Split) ---- 68
5.2 L'évaluation de résultats ---- 69
5.2.1 Séparateur à Vaste Marge (SVM) ---- 69
5.2.2 Classifieur Naïve Bayes (NB) ---- 71
5.2.3 Classifieur La Forêt d'arbres décisionnels (RF) ---- 73
5.2.4 Comparaison entre les cinq classifieur ---- 80
5.3 L'application développer ---- 81
6 Conclusion ---- 82
Conclusion générale et perspectives ---- 84
Bibliographie ---- 85

Table de figures

Figure 1: L'analyse des sentiments médicaux par des émojis 22
Figure 2: Page d'accueil du site web CMS.gov 23
Figure 3: Page d'accueil du site web Drugs.com. 28
Figure 4: Exemple des Tweets **(7)** 29
Figure 5: Les approches de l'analyse des sentiments 30
Figure 6: Les étapes de l'approche à base d'apprentissage automatique. 31
Figure 7: Les différentes étapes de l'approche à base de règles. 32
Figure 8 : L'intelligence artificielle. (10) 36
Figure 9 : La relation entre l'apprentissage automatique et l'intelligence artificielle 37
Figure 10 : Les différents types d'apprentissage automatique. 39
Figure 11 : Exemples de classification et de régression (15) 40
Figure 12 : Le modèle du SVM (17) 42
Figure 13 : Le modèle du KNN (21) 45
Figure 14 : L'apprentissage non supervisé (23) 49
Figure 15 : logo de python 55
Figure 16 : Les principales étapes de l'analyse du sentiment de revues des médicamentes 63
Figure 17:exemple de séparation de mots d'un commentaire 64
Figure 18: Liste de mots vides de l'Anglais de NLTK 65
Figure 19: Fenêtre principale de l'application 81

Table de tableaux

Tableau 1: Comparaison entre l'd'approche à base d'apprentissage automatique et l'approche lexique. (1) 32
Tableau 2: Les avantages et les inconvénients des algorithmes d'apprentissage supervisé. (17) (19) (20) (21) 46
Tableau 3 : Description des variables de Cymbalta_drug_dataset 59
Tableau 4 : Exemples de données de Cymbalta_drug_dataset 59
Tableau 5 : Statistiques de Cymbalta_drug_dataset 60
Tableau 6 : Statistiques de l'ensemble de données drugsCom 61
Tableau 7 : Matrice de confusion 67
Tableau 8 : Matrice de confusion de SVM sur Cymbalta_drug_dataset 69
Tableau 9 : Performances pour SVM sur Cymbalta_drug_dataset 69
Tableau 10 : Matrice de confusion de SVM sur drugsCom_Reduced1 70
Tableau 11 : Performances pour SVM sur drugsCom_Reduced1 70
Tableau 12 : Matrice de confusion de SVM sur drugsCom_Reduced2 70
Tableau 13 : Performances pour SVM sur drugsCom_Reduced2 70
Tableau 14 : Matrice de confusion de SVM sur drugsCom_Reduced3 71
Tableau 15 : Performances pour SVM sur drugsCom_Reduced3 71
Tableau 16 : Matrice de confusion de NB sur Cymbalta_drug_dataset 71
Tableau 17 : Performances pour NB sur Cymbalta_drug_dataset 71
Tableau 18 : Matrice de confusion de NB sur drugsCom_Reduced1 72
Tableau 19 : Performances pour NB sur drugsCom_Reduced1 72
Tableau 20 : Matrice de confusion de NB sur drugsCom_Reduced2 72
Tableau 21 : Performances pour NB sur drugsCom_Reduced2 73
Tableau 22 : Matrice de confusion de NB sur drugsCom_Reduced3 73
Tableau 23 : Performances pour NB sur drugsCom_Reduced3 73
Tableau 24 : Matrice de confusion de RF sur Cymbalta_drug_dataset 74
Tableau 25 : Performances pour RF sur Cymbalta_drug_dataset 74
Tableau 26 : Matrice de confusion de RF sur drugsCom_Reduced1 74
Tableau 27 : Performances pour RF sur drugsCom_Reduced1 74
Tableau 28 : Matrice de confusion de RF sur drugsCom_Reduced2 75
Tableau 29 : Performances pour RF sur drugsCom_Reduced2 75
Tableau 30 : Matrice de confusion de RF sur drugsCom_Reduced3 75
Tableau 31 : Performances pour RF sur drugsCom_Reduced3 75
Tableau 32 : Matrice de confusion de DT sur Cymbalta_drug_dataset 76
Tableau 33 : Performances pour DT sur Cymbalta_drug_dataset 76
Tableau 34 : Matrice de confusion de DT sur drugsCom_Reduced1 76
Tableau 35 : Performances pour DT sur drugsCom_Reduced1 77
Tableau 36 : Matrice de confusion de DT sur drugsCom_Reduced2 77
Tableau 37 : Performances pour DT sur drugsCom_Reduced2 77
Tableau 38 : Matrice de confusion de DT sur drugsCom_Reduced3 77
Tableau 39 : Performances pour DT sur drugsCom_Reduced3 78
Tableau 40 : Matrice de confusion de KNN sur Cymbalta_drug_dataset 78
Tableau 41 : Performances pour KNN sur Cymbalta_drug_dataset 78
Tableau 42 : Matrice de confusion de KNN sur drugsCom_Reduced1 79
Tableau 43 : Performances pour KNN sur drugsCom_Reduced1 79

Table de tableaux

Tableau 44 : Matrice de confusion de KNN sur drugsCom_Reduced2 79
Tableau 45 : Performances pour KNN sur drugsCom_Reduced2..... 79
Tableau 46 : Matrice de confusion de KNN sur drugsCom_Reduced3 80
Tableau 47 : Performances pour KNN sur drugsCom_Reduced3..... 80
Tableau 48: Comparaison des différents ensembles de données en termes de F_measure 81

Introduction générale

Introduction générale

Contexte global de mémoire

Avec l'expansion spectaculaire du World Wide Web et la réponse rapide aux l'information sur Internet, de plus en plus d'utilisateurs s'impliquent dans ce domaine. Internet facilite Internet facilite l'échange de connaissances et d'informations entre les clients et l'entreprises. C'est d'autant plus vrai pour l'utilisation des sites Internet, des forums de discussions et des différents réseaux sociaux.

Les utilisateurs du monde expriment quotidiennement leurs opinions sur des forums de discussions (médicaux, photographie, …), des sites web de services et de produits (commerciaux, médicaux, …), et des réseaux sociaux (Facebook, Twitter, etc.). De nos jours, la plupart des entreprises médicales ont un essentiel besoin de vérification de leurs services et produits. Ces opérations dépendent du point de vue des consommateurs sur ces services ou produits. Par conséquent, l'obtention de l'information utile à partir de ces opinions devienne très difficile en termes de temps et d'effort qu'elle requiert.

Les sondages des utilisateurs sont désormais plus accessibles et plus faciles à utiliser. Ces informations sont généralement en mode texte, de sorte que les nouvelles technologies telles que le Web mining et le web sémantique facilitent l'analyse du texte et conduisant à l'extraction de connaissances. Tout cela est considéré comme un processus appelé analyse des sentiments.

L'analyse des sentiments, également connu sous le nom d'opinion mining, est une étude informatique des opinions, des sentiments, et des attitudes à propos de sujets textuels, d'entités, de personnes et d'événements, exprimé à travers du texte. Ceci est destiné à affecter des sentiment prédéfinies (très négatifs, négatifs, neutres, positifs, très positifs, etc.) aux données textuelles. Ce processus vise à mieux comprendre l'opinion publique sur diverses sujettes. De nombreuses études ont montré que l'analyse des sentiments est d'un grand intérêt pour les personnes qui se concentrent sur l'opinion publique, pour de nombreuses raisons personnelles, commerciales, médicales, etc.

Actuellement, les sites web de services et de produits médicaux sont considérés parmi les sites web les plus populaires et les plus utilisés. Ils permettent aux patients de partager

leurs commentaires et d'exprimer leurs opinions sur les services médicaux tels que ; les rendez-vous, l'imagerie médicale, les consultations des médecins et les produits médicales tels que ; tension mètre, médicaments, etc.

L'analyse des sentiments est nécessaire pour les entreprises pharmaceutiques et de services médicaux afin d'améliorer leurs services et produits, augmentant ainsi leurs bénéfices. Cependant, les systèmes d'apprentissage automatique peuvent être difficiles à mettre en œuvre en raison de la complexité de l'interprétation du langage humain. Le traitement des sentiments est si complexe qu'il faut utiliser d'autres domaines comme le Traitement Automatique du Langage Naturel (TALN) et l'apprentissage automatique avec ces différents classificateurs.

Problématique

Aujourd'hui, avec le développement énorme d'internet, les patients sont de plus en plus engagés dans les communautés de santé tel que les forums de discussion médicaux. Cela pour recueillir les informations de santé, partager de l'expérience sur les médicaments, les traitements, le diagnostic ou pour l'interaction avec autrui avec des conditions de santé similaires. Le sujet du présent mémoire s'intéresse à l'analyse de sentiments liés à l'état de santé des patients dans les réseaux sociaux et /ou les forums de discussion spécialisés. Le but de ce travail sera alors d'aider à l'amélioration des conditions des patients ainsi les produits médicaux offerts par les entreprises spécialisées.

Objectifs de l'étude

Les sites d'évaluation en ligne et les forums d'opinion contiennent une mine d'informations sur les préférences et les expériences des utilisateurs dans plusieurs domaines (commercial, politique, médicale, social, économique, etc.). Ces informations peuvent être évaluées l'aide d'approches d'exploration de données telles que l'analyse des sentiments pour obtenir des informations précieuses.

Ce travail passe en revue les opinions d'utilisateurs en ligne dans le domaine médicale. Les opinions des clients ou des patients en ligne dans ce domaine contiennent des informations
Sur plusieurs aspects, tels que les commentaires sur l'efficacité des médicaments et les effets secondaires, ce qui rend l'analyse automatique très intéressante mais également difficile. Cependant, l'analyse des sentiments de divers aspects indésirables des médicaments peut

améliorer la surveillance de la santé publique en fournissant des informations précieuses, en aidant à la prise de décision et en révélant des expériences collectives.

Dans ce travail, nous utilisons les données obtenues en explorant un sites web d'évaluation de médicaments en ligne, pour effectuer plusieurs tâches d'évaluation de médicaments. Tout d'abord, effectuez une analyse des sentiments pour prédire les sentiments liés à la satisfaction globale, aux effets secondaires et à l'efficacité à partir des opinions des utilisateurs sur des médicaments particuliers. Pour remédier à la pénurie de données annotées, nous avons choisi de collecter et d'annoter nos propres données. Dans le cadre de ce mémoire nous avons sollicités le site web www.askapatient.com pour collecter l'ensemble de données, Cymbalta_drug_dataset des commentaires des patients envers le médicament Cymbalta utilisé dans le traitement de la dépression. Dans notre corpus nous étions intéressés par l'équilibrage entre les classes positive et négative. En plus, nous avons collecté trois autres ensembles de données ; drugsCom_Reduced1, drugsCom_Reduced2, et drugsCom_Reduced3.

Pour l'annotation de l'orientation sentimentales des différents ensembles de données nous nous basons sur les notes d'évaluation des auteurs de commentaires envers leurs expériences avec les médicaments. Dans notre corpus, Cymbalta_drug_dataset, le système d'évaluation du site web www.askapatient.com est sur une échelle de 5 étoiles. Nous avons considéré un commentaire ayant une note de plus de trois étoiles comme positif. Un commentaire de moins de trois étoiles est considéré comme négatif. Les commentaires avec 3 étoiles sont considérés neutre et éliminés de notre corpus.

Pour les autres corpus, drugsCom_Reduced1, drugsCom_Reduced2, et drugsCom_Reduced3, le système d'évaluation du site est sur une échelle de 10 étoiles. Nous avons annoté un commentaire ayant une note de plus de cinq étoiles comme positif. Un commentaire de moins de cinq étoiles est considéré comme négatif. Les commentaires avec cinq étoiles sont annotés neutre et éliminés des ensembles de données.

Une interface graphique est développée à la fin de ce travail pour faciliter l'exploitation des modèles obtenus par les utilisateurs finaux.

Organisation du mémoire

Après cette introduction générale, le reste de notre travail est structuré comme suit :

- Le premier chapitre : dans le premier chapitre nous présentons un aperçu de l'analyse des sentiments médicaux et de l'importance de l'analyser, puis expliquons les niveaux et les types d'analyse des sentiments. Ensuite, nous expliquons les sources de ces sentiments et donnons des exemples de sources dans lesquelles les sentiments sont analysés. Enfin, nous introduisons des méthodes d'analyse des sentiments. Enfin nous introduisons des approches de l'analyse des sentiments.
- Par la suite, dans le deuxième chapitre, nous présenterons l'apprentissage automatique et sa relation avec le domaine de l'intelligence artificielle. Nous présenterons ainsi les différents types de l'apprentissage automatique avec des algorithmes de chaque type. Dans ce travail, nous nous intéresserons à l'apprentissage supervisée, et donc nous présenterons ses algorithmes avec leurs avantages et leurs inconvénients.
- Ensuite, le troisième chapitre définira les outils de programmation et l'implémentation de notre travaille. Aussi les différents ensembles de données, et les résultats obtenus.
- Finalement, nous clôturons ce mémoire par une conclusion générale comportant des perspectives pour des prochaines travaux.

CHAPITRE 1 :
Fouille de sentiments médicaux

Chapitre 1 : fouille de sentiments médicaux

1 Introduction

Avec l'avènement des technologies Web 2.0 et un nombre croissant de sites Web et réseaux sociaux et de forums Web, où les internautes actuels (par exemple, les patients) ont la possibilité d'ajouter leurs commentaires, évaluation ou avis sur les réseaux sociaux. De plus, les patients consultent souvent les forums de discussion en demandant l'avis de leurs amis et en consultant des opinions positives, négatives et neutres sur un sujet particulier avant de prendre la décision d'utiliser des produits ou services médicaux. Cela donne une opinion sur un sujet particulier, l'étude de ces opinions est appelée analyse des sentiments. Dans ce chapitre nous allons expliquer et exposer le domaine de fouille de sentiments médicaux.

2 Analyse des Sentiments médicaux

C'est un sous-domaine du NLP (Natural Languge Processing) ou TALN (Traitement Automatique des Langages Naturels). Il s'intéresse à l'interprétation et à la classification des sentiments (positives, négatives, neutres, etc.) dans les données textuelles à l'aide de techniques d'analyse de texte. L'Analyse des Sentiments (Sentiment Analysis) a permet de déterminer le sentiment derrière une suite de mots. Par exemple les entreprises médicales peuvent identifier les sentiments des clients ou des patients envers les produits tels que les médicaments (ZYRTEC, ABILFY, BACLOFEN, …), les marques des produits ou les services à travers des conversations et des commentaires. (1)

L'analyse de sentiments, aussi connu sous le nom fouille d'opinions (opinion mining), révèle des opinions, des attitudes et des émotions basées sur ce que les clients ou les patients disent d'une chose particulière. Les modèles d'analyse des sentiments se concentrent non seulement sur la polarité (neutre, négatif, positif), mais aussi sur les émotions et les sentiments (triste, heureux, en colère, etc.), l'intention (qu'elle soit intéressée) et l'urgence (qu'elle soit urgente). (2) La Figure 1 illustre l'analyse des sentiments médicaux par des émojis.

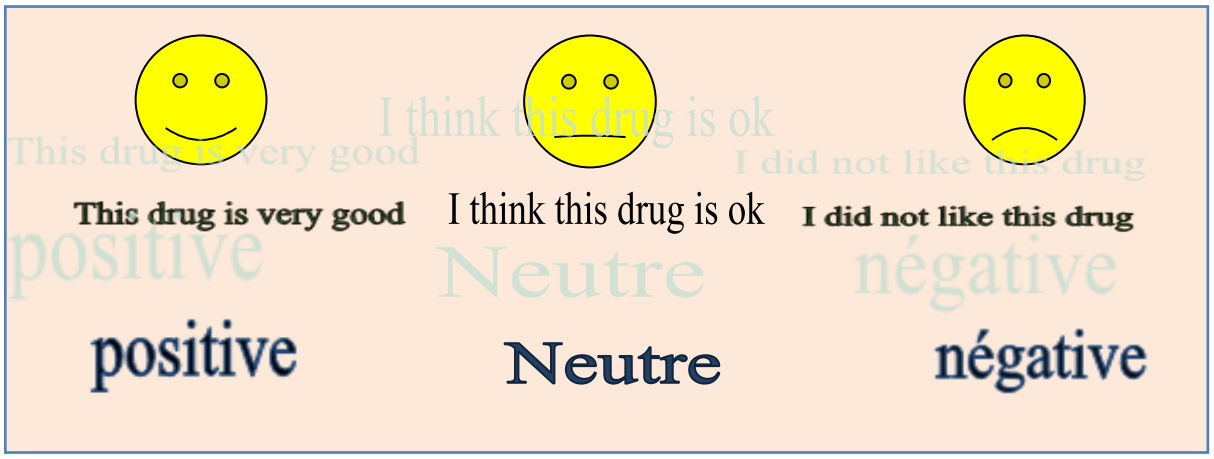

Figure 1: L'analyse des sentiments médicaux par des émojis

2.1 Explication de l'analyse des sentiments médicaux

Dans le domaine médical, l'analyse des sentiments des patients aide les prestataires à acquérir un avantage concurrentiel par rapport à leurs concurrents en les aidant à améliorer leurs services en fonction de leurs opinions et de leurs commentaires. Les résultats de ces analyses donnent aux prestataires des informations plus détaillées sur le traitement dont ils ont besoin, le type et la qualité de service que les clients et les patients souhaitent et le type d'hôpital pour lequel ils recherchent un traitement. (2)

L'analyse des sentiments des patients est souvent effectuée dans le cadre d'un sondage d'évaluation des consommateurs hospitaliers (HCAHPS, Hospital Consumer Assessment of Healthcare Providers and Systems) auprès des prestataires et des systèmes de soins de santé. Ce sondage est un « sondages publique nationale normalisée sur les opinions des patients sur les soins hospitaliers ». (2)

HCAHPS est un outil de sondages et une méthode de collecte de donnée utilisés pour mesurer la perception d'un patient de l'expérience hospitalière, voir Figure 2. (2) De nombreux hôpitaux ont recueilli des informations sur la satisfaction des patients pour leur usage interne, mais en ce que concerne les expériences de soins des patients qui permettent des comparaisons valides entre les hôpitaux aux niveau local, il n'y avait pas de norme nationale pour la collecte et divulgation des informations. La suivante représente une site web qui permet aux patients de partager leurs expériences hospitalières. (3)

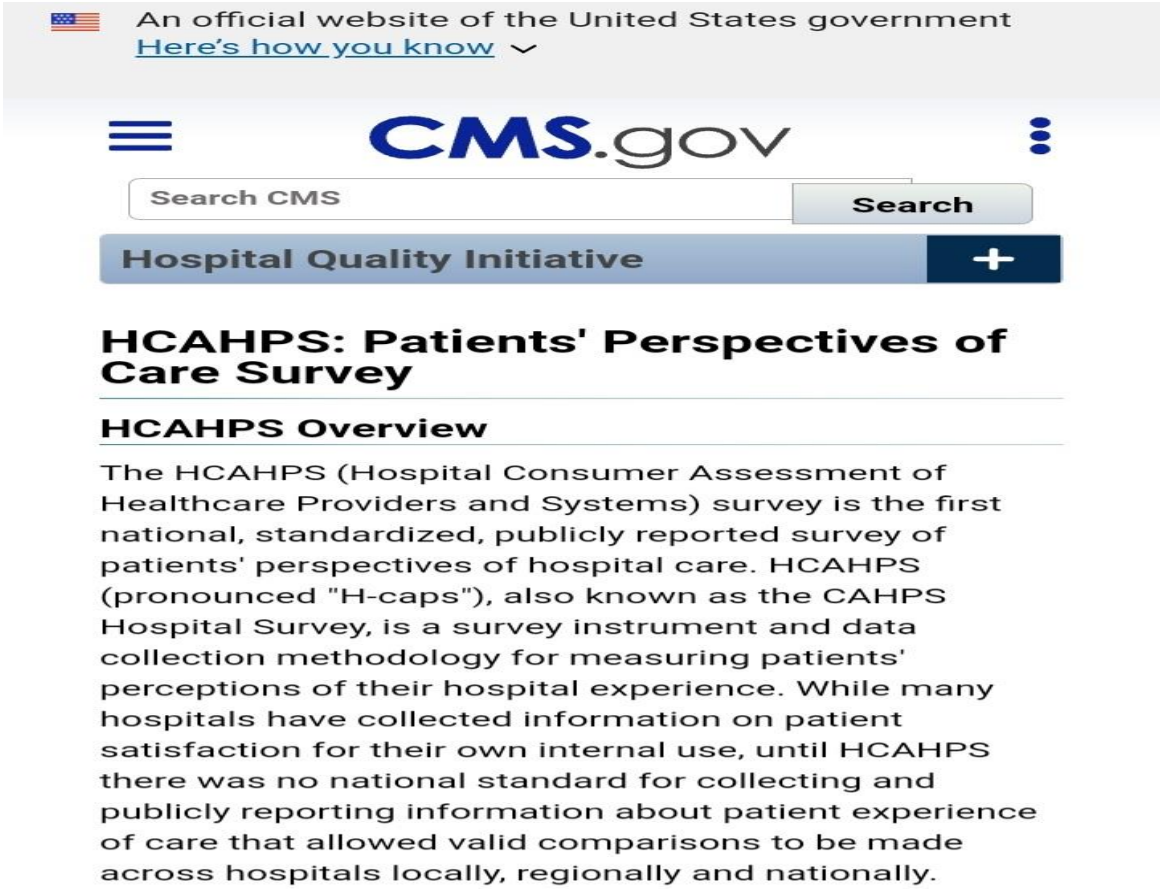

Figure 2: Page d'accueil du site web CMS.gov

Trois objectifs clés ont façonné HCAHPS :

- Premièrement, l'étude vise à fournir des données sur les perspectives des soins aux patients qui permettent des comparaisons objectives et significatives entre les hôpitaux sur des questions importantes pour les consommateurs. (3)
- Deuxièmement, la publication des résultats de l'enquête incite les hôpitaux à améliorer la qualité des soins. (3)
- Troisièmement, les rapports publics contribuent à accroître la responsabilisation en matière de soins de santé en rendant la qualité des soins hospitaliers plus transparente en échange d'un investissement public. Avec ces objectifs à l'esprit. (3)

2.2 L'importance et l'objectifs de l'analyse des sentiments médicaux

Les patients ont des sentiments très forts à propos des soins médicaux qu'ils reçoivent. Presque toutes les interactions avec un prestataire ou un hôpital provoquent une réaction positive ou négative. Par conséquent, l'analyse des sentiments dans le domaine de la santé est d'une grande valeur. (2)

Les connaissances acquises grâce à l'analyse de sentiment des patients permettent aux prestataires de soins de santé de combler les lacunes de communication entre les établissements et les patients. Cela peut optimiser l'expérience du patient et améliorer les résultats commerciaux à plus grande échelle. (2) L'importance de l'analyse des sentiments médicaux est donc expliquer dans ce qui suit.

2.2.1 L'analyse des sentiments médicaux aide les prestataires à améliorer la communication avec les patients

La communication est importante dans sa mission de fournir des soins de santé de qualité. Les prestataires de santé améliorent l'expérience du patient et la littératie en santé lors de la mise en œuvre de stratégies de communication sur la santé. (2)

Les patients veulent des rendez-vous et des changements de rendez-vous faciles, un accès à l'information, des réponses aux questions, et plus encore. Ils détestent tous passer des appels ou être pris dans un arbre téléphonique. Ils veulent tous communiquer rapidement et facilement avec leurs fournisseurs de soins de santé et leurs médecins. (2)

L'analyse des sentiments des soins de santé peut indiquer au prestataire si ces causes ou d'autres irritent le patient. Cela donne aux prestataires un aperçu du point de vue du patient et leur permet de voir ce qui est important pour eux et dans quelle mesure ils communiquent avec eux. (2)

L'un des avantages de l'analyse des sentiments est la possibilité de classer les commentaires des patients dans différentes. Cette segmentation permet une analyse très détaillée. Les prestataires peuvent examiner la communication avec les médecins, la communication avec les infirmières, la réponse du personnel hospitalier, etc. Cela donne aux prestataires une

meilleure compréhension des facteurs qui influencent la façon dont il communique avec le patient. Les méthodes de base de l'analyse des sentiments sont :

- ✓ Analyser les données recueillies pour trouver les plaintes et les compliments les plus courants.
- ✓ Segmenter les informations en fonction des médecins, des services et d'autres indicateurs pour identifier les opportunités d'amélioration des soins aux patients. (2)

2.2.2 L'analyse des sentiments médicaux aide à quantifier les performances des employés et des services

L'analyse des sentiments dans l'industrie médicale permet aux prestataires d'identifier les domaines de supériorité (et d'échec potentiel) en termes de service aux patients. Cela comprend des catégories importantes de stratégies de communication avec les patients. (2)

L'analyse des sentiments dans l'industrie médicale permet aux prestataires de catégoriser les commentaires des patients en fonction de la personne (infirmière ou médecin), de l'emplacement et du processus. En évaluant les commentaires comme positifs ou négatifs, les hôpitaux peuvent quantifier l'expérience du patient et identifier les domaines à améliorer. (2)
Voici quelques exemples de ce que patients peuvent dire de leur expérience.

- ➢ "Une infirmière gentille et attentionnée. (Positif)
- ➢ " Bon service client, mais les réservations doivent être automatisées. "(Neutre)
- ➢ " J'ai attendu pendant des heures pour parler à la réceptionniste. (Négatif)

En analysant les sentiments du patient, le prestataire peut recueillir ces commentaires en fonction de leur fréquence et de leur intensité. Cela permet d'identifier les domaines dans lesquels l'entreprise fonctionne de bonne manière, a besoin d'amélioration et manque de compétences, etc. (2)

2.3 Les niveaux d'analyse des sentiments

La recherche sur l'analyse des sentiments est menée à trois principaux niveau d'analyse.

- ✓ Niveau du document (Document Level en Anglais).
- ✓ Niveau de la phrase (Sentence Levels en Anglais).
- ✓ Niveau des aspects (Entity and aspect level en Anglais).

2.3.1 Niveau du document

Détermine la polarité de l'ensemble du texte. Le texte est censé ne représenter qu'une opinion sur une seule entité (comme un seul produit ou service). (4)

2.3.2 Niveau de la phrase

Détermine la polarité de chaque phrase contenue dans le texte. Chaque phrase du texte est censée représenter une opinion sur une entité unique. (4)

2.3.3 Niveau des aspects

Effectue une analyse plus détaillée que les autres niveaux. Les opinions sont basées sur l'idée qu'elles sont composées d'émotions et d'objectifs (opinions). Par exemple, la phrase « le médicament est très bon mais doit encore tenir compte des effets secondaires et du prix » évalue trois aspects : le médicament (positif), des effets secondaires (négatif), et le prix (négatif) (4).

3 Types d'analyse de sentiments

Il existe de nombreux types d'analyses de sentiments allant des systèmes qui se concentrent sur la classification de la polarité (positif, négatif, neutre) aux systèmes qui détectent des émotions (en colère, heureux, triste, etc.) ou identifient des intentions (par exemple, intéressé, pas intéressé). Dans la section suivante, nous aborderons les types les plus importants. (4)

3.1 Analyse fine des sentiments (fine-grained sentiments analysis)

Au lieu de parler de phrases négatives, positives ou neutres, nous considérons les catégories suivantes ; Positive, Très positive, Très négative, Négative et Neutre.

Certains systèmes offrent également différentes classifications de polarité en identifiant si le sentiment positif ou négatif est associé à un sentiment particulier, tel que la colère, la tristesse ou des inquiétudes (sentiments négatifs) ou du bonheur, de l'amour ou de l'enthousiasme (sentiments positifs). (4)

3.2 Détection d'émotion (Emotion detection)

La détection des émotions vise à détecter des émotions telles que le bonheur, la frustration, la colère, la tristesse, etc. De nombreux systèmes de détection d'émotions sont basés sur l'utilisation de lexiques de sentiments (c'est-à-dire des listes des émotions) ou sur des algorithmes d'apprentissage automatique complexes. (4)

3.3 Analyse de sentiments à base d'aspects (Aspect-Based Sentiment Analysis ABSA)

Au lieu de classer le sentiment général d'un texte en positif ou en négatif, l'analyse de sentiments à base d'aspects permet d'analyser le texte afin d'identifier différents aspects et de déterminer le sentiment correspondant pour chacun. Les résultats sont plus détaillés, intéressants et précis car l'analyse à base d'aspects examine de manière précise les informations contenues dans un texte. (4)

4 Sources des sentiments médicaux

Les données médicales (sentiments) à analyser sont collectées à partir de diverses sources telles que les blogs, les forums de discussions, les articles médicaux, les sites web, et les réseaux sociaux (Twitter, Facebook, LinkedIn, etc.)

4.1 Analyse des sentiments sur les sites web

Les sites web médicaux permettent à leurs utilisateurs de poster leur commentaires et opinions sur les produits ou services médicaux. Ces sentiments sont extraits de site web, et analysés afin d'aide d'autres patients à prendre leurs propres décisions, concernant ces produits ou services. La Figure 3 représenter un site web, qui contient des sentiments des patients sur les médicaments.

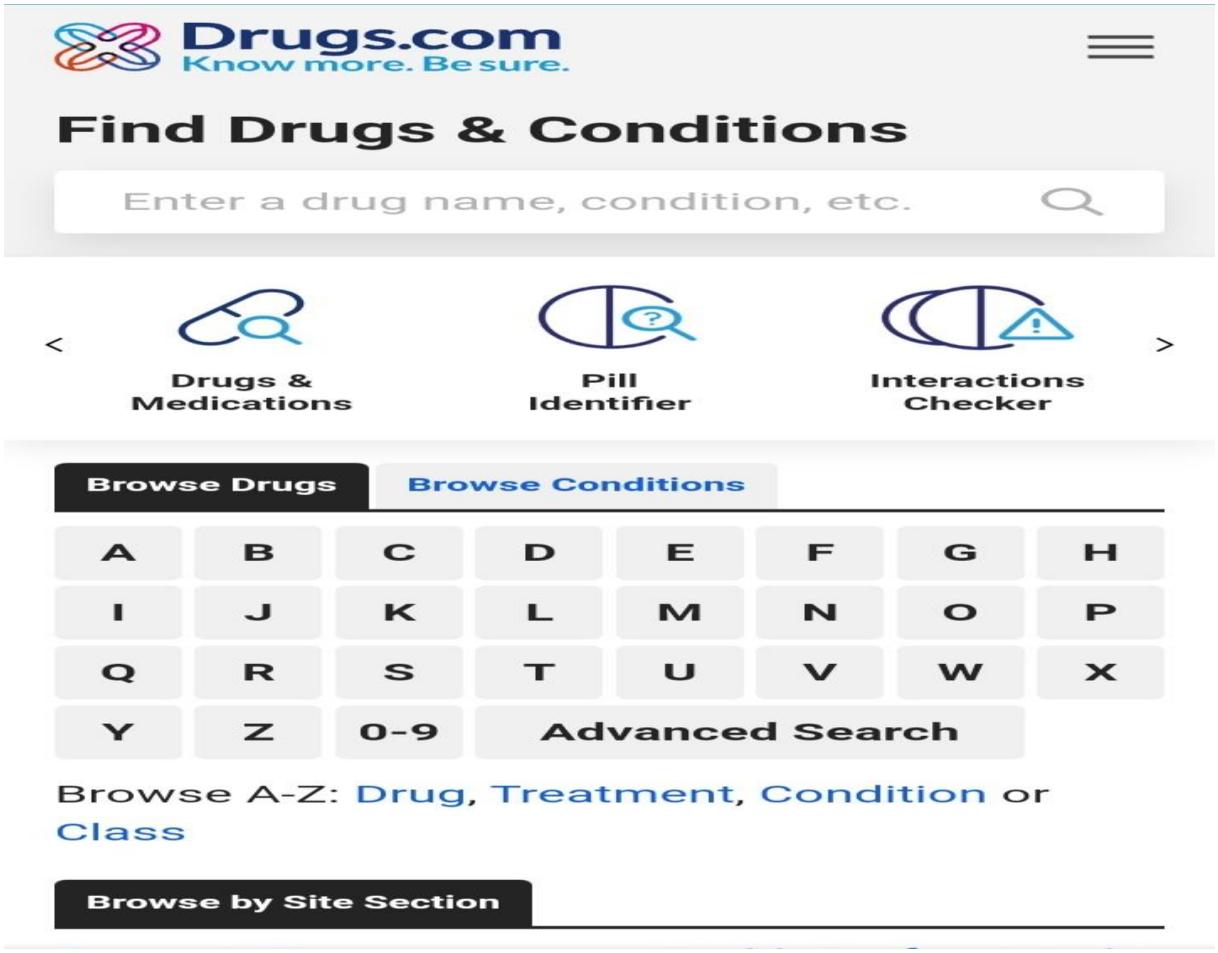

Figure 3: Page d'accueil du site web Drugs.com.

4.2 Analyse des sentiments sur les réseaux sociaux

L'utilisation des médias sociaux par les utilisateurs-patients est fondamentale, car elle leur permet de partager des opinions et rechercher des informations sur divers sujets. Sur les différentes plateformes de médias sociaux, les utilisateurs-patients peuvent émettre des sentiments et participer à des discussions. Certains utilisateurs partagent des informations sur des sujets relatifs à la santé, comme leurs expériences avec des centres de santé publiques et des établissements de soins médicaux. Ils peuvent ainsi apprécier un service, recommander un médecin ou des cliniques. Ils peuvent aussi se plaindre de certains aspects particuliers dudit établissement, tels que la qualité du service, le fonctionnement dans le département des urgences (heure d'attente, compétence et attitude des soignants). Ces données peuvent être d'une grande valeur si elles sont extraites, traitées et analysées. Cela a intéressés plusieurs concepteurs de logiciels. Il existe plusieurs applications

informatiques qui permettent de faire l'analyse des données issues des médias sociaux et la détection de sentiments exprimés dans les conversations. (5)

4.2.1 Le réseau social twitter

Twitter est un service de microblogue où les utilisateurs peuvent envoyer et lire de courts messages de 140 caractères appelés "tweets". Plusieurs tweets non structurés, en texte libre, relatifs aux soins de santé sont partagés sur Twitter, qui devient un domaine populaire pour la recherche sur les soins de santé. Le sentiment est une métrique couramment utilisée pour étudier les opinions positives ou négatives dans ces messages. L'exploration des méthodes utilisées pour l'analyse des sentiments dans la recherche sur les soins de santé sur Twitter peut nous permettre de mieux comprendre les options disponibles pour les recherches futures dans ce domaine. (6)

Les sentiments médicaux sur Twitter ont trop de valeur, pour cela les entreprises médicaux cherchent à analyser ces sentiments et à extraire les véritables intentions derrière ces sentiments. Ces sentiments sont exprimés dans des conversations et des tweets. L'objectif de mener une analyse de ces sentiments sur Twitter est de connaître les opinions des patients sur les produits et services et de classer en opinions positives, négatives et neutres. La technique d'analyse des sentiments peut être appliquée à toutes les langues du monde. La Figure 4 montrer un tweet écrit en arabe concernant le vaccin utilisé pour Covid-19.

Figure 4: Exemple des Tweets (7)

5 Les approches de l'analyse des Sentiments

Les approches de classification de sentiments peuvent être catégorisées comme suit (voir Figure 5) :

- Approche à base d'apprentissage automatique (Machine Learning-based approach) : systèmes qui s'appuient sur des techniques d'apprentissage automatique à partir de données. (4)
- Approche à base de règles (Rule-based approach) : systèmes qui effectuent une analyse des sentiments basée sur un ensemble de règles. (4)
- Approche hybride : systèmes combinant à la fois des approches automatiques. (4)

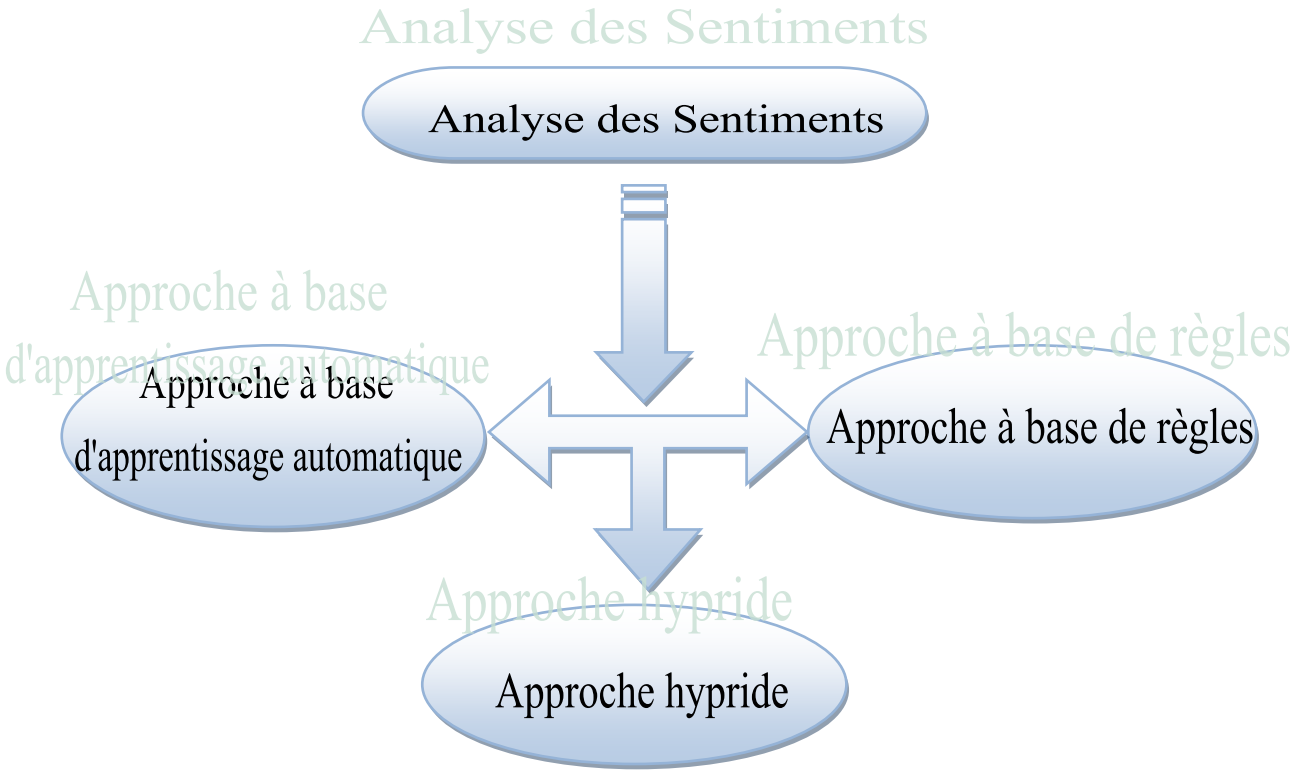

Figure 5: Les approches de l'analyse des sentiments

5.1 Approche à base d'apprentissage automatique

Les approches à base d'apprentissage automatique reposent sur des techniques d'apprentissage automatique (Machine Learning), c.f, dans Figure 6. La tâche d'analyse des sentiments est généralement modélisée comme un problème de classification dans lequel un classificateur est alimenté avec un texte et renvoie la catégorie correspondante, par exemple, positif, négatif et neutre (en cas d'analyse de polarité). (4)

Figure 6: Les étapes de l'approche à base d'apprentissage automatique.

Dans ces approches, la machine est entraînée à détecter des modèles dans un corpus en la faisant apprendre sur un premier corpus test. Ce type d'apprentissage est similaire à l'apprentissage humain des expériences passées pour acquérir de nouvelles connaissances afin d'améliorer sa capacité d'effectuer des tâches dans le monde réel. Dans l'apprentissage automatique, la machine apprend à partir de données collectées dans le passé, qui représentent des expériences passées dans certaines applications du monde réel.

5.2 Approche à base de règles (Rule-based)

L'approche à base de règles (ou l'approche lexique) définit un ensemble de règles dans un type de langage de programmation (script) qui identifie la subjectivité, la polarité ou le sujet d'une opinion voir Figure 7. Cette approche peut utiliser diverses entrées, telles que : (4)

- Techniques classiques de NLP, telles que la racinisation, tokenisation, POS-tagging et Chunking. (4)
- Autres opérations basées sur le lexique, ils utilisent le dictionnaire des sentiments avec des mots d'opinion et les faire correspondre avec les données pour déterminer la polarité. (4)

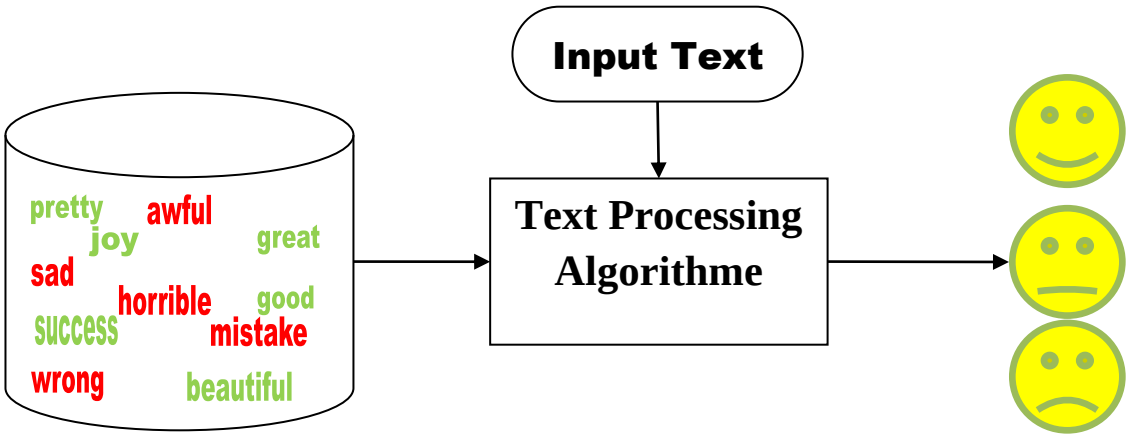

Figure 7: Les différentes étapes de l'approche à base de règles.

5.3 Approche hybride

Le concept de méthode hybrides est très intuitif ; combine simplement le meilleur des deux approches, celui basé sur des règles et celui basé sur l'apprentissage automatique. Généralement, en combinant les deux approches, les méthodes peuvent améliorer la précision. (4)

5.4 Les avantages et les inconvénients d'approches

Ce Tableau 1 suivent montre les avantages et les inconvénients de l'approche automatique et l'approche à base de règles que nous avons vues précédemment.

Tableau 1: Comparaison entre l'd'approche à base d'apprentissage automatique et l'approche lexique. (1)

Approches	**Avantages**	**Inconvénients**
Approche à base de règles	-Ne demande aucune donnée d'entrainement ou des données étiquetées et ceci permet d'introduire moins d'opération de calcul.	-Moins de capacités de classification en fonction ou du domaine. -Exige l'existence de ressources linguistique puissantes qui ne sont pas toujours disponibles

Approche à base d'apprentissage automatique	-Il peut être transformé en ce que le domaine demande pour mieux travailler. -Un dictionnaire n'est pas nécessaire. -Donne de meilleurs résultats en termes de haute précision de classification.	-peut être affecté par les variations de classes et aussi par l'effet des changements linguistiques. -Les modèles qui sont entrainés sur un domaine spécifique, dans la plupart des cas ne fonctionnent pas avec un autre.

6 Conclusion

Nous avons étudié dans ce chapitre, l'analyse des sentiments dans le domaine médical est un aspect très important, car elle aide les entreprises à réaliser à quel point les clients sont satisfaits du produit, ce que leur permet d'apporter des modifications et d'améliorer ce produit, et les aide également à développer.

Dans le prochain chapitre, nous présenterons l'apprentissage automatique, ses types et certaines de ses algorithmes, ainsi que leurs avantages et leurs inconvénients.

CHAPITRE 2 :

L'apprentissage automatique

Chapitre 2 : l’apprentissage automatique

1 Introduction

L'apprentissage automatique (ou machine Learning) est la discipline qui consiste à appliquer des algorithmes à des ensembles de données pour extraire des modèles. Celles-ci peuvent à leur tour être appliquées à des données similaires à des fins prédictives, il est possible de formuler une approximation de la relation entre tous les champs d'entrée et les valeurs individuelles. Cette formule peut ensuite être appliquée à de nouvelles entrées pour prédire la valeur associée. Cette approche diffère lorsque l'application est développée sur la base de règles prédéfinies. Alors que les concepts de base de l'apprentissage automatique existent depuis un certain temps, le domaine a récemment pris de l'ampleur, en partie à cause de l'amélioration des performances des processeurs, en particulier des graphiques, et en partie à cause de la disponibilité d'une grande quantité d'informations. Ces deux éléments sont essentiels pour des prévisions précises. Étant donné qu’il existe déjà suffisamment de littérature sur l’histoire de l’apprentissage automatique.

Dans un autre part, l’apprentissage automatique est une branche très essentielle de l’intelligence artificielle performante dédié à la résolution de problèmes divers, qui peuvent aller du filtrage d’une collection de photos aux défis mondiaux les plus urgents (en termes de santé ; environnement, par exemple). Dans ce chapitre on va définir d’abord l’intelligence artificielle puis nous donnons une définition de l’apprentissage automatique avec une mention de ses types et les algorithmes spécifiques qu’il utilise.

2 L’intelligence artificielle

En termes simples, l’intelligence artificielle (IA) fait référence à des systèmes ou machines qui imitent l’intelligence humaine pour effectuer des taches et qui peuvent s’améliorer en fonction des informations collectées gras à l’itération, voir Figure 8. L’intelligence artificielle se manifeste sous plusieurs formes. Voici quelques exemples : (8)

- Les chatbots utilisent l’IA pour comprendre les problèmes des clients plus rapidement et répondre plus efficacement. (8)

- Les assistants intelligents utilisent l'IA pour analyser les informations critiques à partir de grands ensembles de données en texte libre afin d'améliorer la planification. (8)
- Les moteurs de recommandation peuvent suggérer automatiquement des émissions télévisées en fonction des habitudes des téléspectateurs. (8)

L'intelligence artificielle est davantage liée au processus et à la capacité de réflexion et d'analyse de données approfondies au maximum qu'à un format ou des fonctions particuliers. Bien que l'IA évoque des images de robots ultra-performants ressemblant à des humains et envahissant le monde, l'IA n'est pas destinée à nous remplacer. Elle vise à améliorer de manière significative les capacités et les contributions humaines. Cela est, en fait, un atout commercial très précieux. (8)

L'intelligence artificielle (IA) est un processus d'imitation de l'intelligence humaine qui repose sur la création et l'application d'algorithmes exécutés dans un environnement informatique dynamique. Son but **est** de permettre à des ordinateurs de penser et d'agir comme des êtres humains ; L'apprentissage automatique est une discipline de l'intelligence artificielle qui efforce de trouver un moyen de créer des programmes informatiques qui s'améliorent automatiquement avec l'expérience. (9)

Figure 8 : L'intelligence artificielle. (10)

2.1 Termes associés à l'intelligence artificielle

L'IA est devenue un terme fourre-tout pour les applications qui effectuent des tâches complexes nécessitant auparavant une intervention humaine, comme communiquer avec les

clients en ligne ou jouer aux échecs. Le terme est souvent utilisé de manière interchangeable avec les domaines qui composent l'IA tels que l'apprentissage automatique (Machine Learning en Anglais) et l'apprentissage profond (Deep Learning en Anglais). Il y a cependant des différences. Par exemple, l'apprentissage automatique (voir la Figure 9) est axé sur la création de systèmes qui apprennent ou améliorent leurs performances en fonction des données qu'ils traitent. Il est important de noter que, même si l'intégralité de l'apprentissage automatique repose sur l'IA, cette dernière ne se limite pas à l'apprentissage automatique. (8)

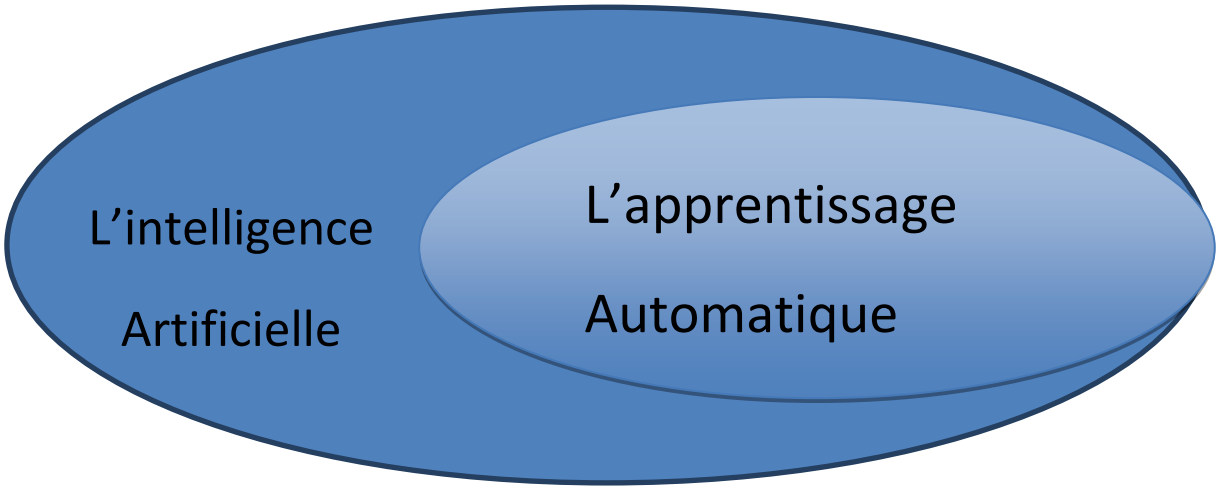

Figure 9 : La relation entre l'apprentissage automatique et l'intelligence artificielle

La définition de l'apprentissage automatique selon Wikipédia[1] est : L'apprentissage automatique (en anglais : machine Learning, litt.« apprentissage machine »), apprentissage artificiel ou apprentissage statistique est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'« apprendre » à partir de données, c'est à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune comme illustré sur Figure 9. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes. (11)

L'apprentissage automatique, également appelé apprentissage machine ou apprentissage artificiel et en anglais machine Learning, est une forme d'intelligence artificielle (IA) qui permet à un système d'apprendre à partir des données et non à l'aide d'une programmation explicite. Cependant, l'apprentissage automatique n'est pas un processus simple. Au fur et à mesure que les algorithmes ingèrent les données de formation, il devient possible de créer des

[1] fr.wikipedia.org

modèles plus précis basés sur ces données. Un modèle d'apprentissage automatique est le résultat généré lorsque vous entraînez votre algorithme d'apprentissage automatique avec des données. Après la formation, lorsque vous fournissez des données en entrée à un modèle, vous recevez un résultat en sortie. Par exemple, un algorithme prédictif crée un modèle prédictif. Ensuite, lorsque vous fournissez des données au modèle prédictif, vous recevez une prévision qui est déterminée par les données qui ont servi à former le modèle (12).

2.2 L'avantage de l'apprentissage automatique

L'avantage de l'apprentissage automatique est qu'il permet d'utiliser des algorithmes et des modèles pour prédire les résultats. L'astuce consiste à s'assurer que les spécialistes des données utilisent les bons algorithmes, les données les plus appropriées (précises et propres) et les meilleurs modèles d'exécution. Si tous ces éléments se coordonnent harmonieusement, il devient alors possible de former en continu le modèle et d'exploiter les résultats en apprenant à partir des données. L'automatisation de ce processus de modélisation, de formation du modèle et de test débouche sur des prédictions précises qui accompagnent utilement le changement métier (12).

3 Les types de l'apprentissage automatique

L'apprentissage automatique (Machine Learning en anglais) est utilisé en intelligence artificielle et en science et analyse des données (Analytics and Data Science). Il existe différents types d'apprentissage automatique (voir Figure 10) ; Le supervisé, le non-supervisé, semi-supervisé et celui par renforcement (13).

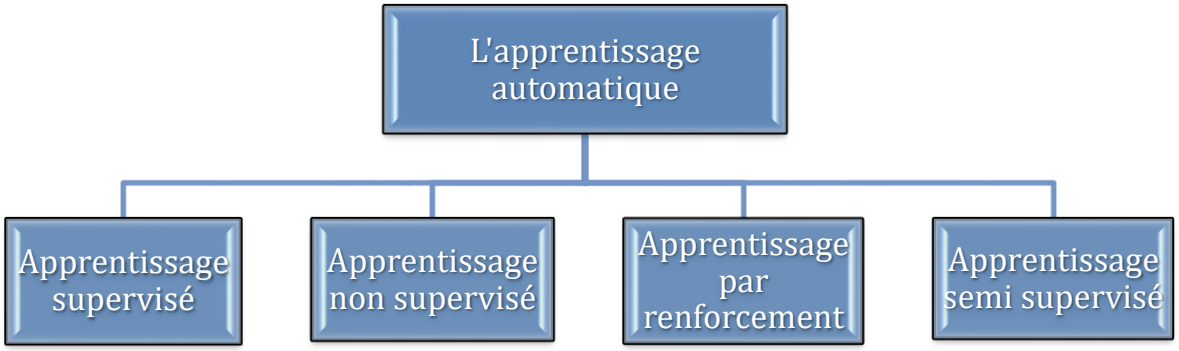

Figure 10 : Les différents types d'apprentissage automatique.

3.1 Apprentissage supervisé

Apprentissage supervisé (Supervised Learning en Anglais) est une tâche d'apprentissage automatique consistant à apprendre une fonction de prédiction à partir d'exemples annotés, au contraire de l'apprentissage non supervisé. On distingue les problèmes de régression des problèmes de classification. Ainsi, on considère que les problèmes de prédiction d'une variable quantitative sont des problèmes de régression tandis que les problèmes de prédiction d'une variable qualitative sont des problèmes de classification (14).

L'apprentissage supervisé commence généralement par un ensemble de données bien défini et une certaine compréhension de la façon dont ces données sont classifiées. L'apprentissage supervisé a pour but de déceler des modèles au sein des données et de les appliquer à un processus analytique. Ces données comportent des caractéristiques associées à des libellés qui définissent leur signification. Vous pouvez, par exemple, créer une application d'apprentissage automatique capable de faire la distinction entre plusieurs millions d'animaux, en se basant sur des images et des descriptions écrites (12).

Les exemples annotés constituent une base d'apprentissage, et la fonction de prédiction apprise peut aussi être appelée « hypothèse » ou « modèle ». On suppose cette base d'apprentissage représentative d'une population d'échantillons plus large et le but des méthodes d'apprentissage supervisé est de bien généraliser, c'est-à-dire d'apprendre une fonction qui fasse des prédictions correctes sur des données non présentes dans l'ensemble d'apprentissage (14).

En apprentissage supervisé, on distingue entre deux types de tâches (voir Figure 11) :

- La classification ; des problèmes de classification surviennent lorsque les variables de sortie appartiennent aux catégories telle que les catégories suivantes ; "Blanc ","Noir " ou "maladie " et "pas malade ". (15)Exemples ;
  - Opération financière et bancaires pour détecter la fraude à la carte crédit (fraude, not fraude)
  - Détection d'E-mail non sollicité (spam, pas spam).
  - Utilisé dans le domaine de marketing pour analyser les sentiments textuelles (satisfait, pas satisfait)
  - Dans le domaine médical, il utilisé pour prédire si un client est malade ou pas malade.
- La Régression ; si les variables des sorties sont des valeurs réelles, vous aurez un problème de régression telle que "Dollar" ou "Poids" (15). Exemples ; Prédire les prix des maisons, Prédire les cours des bourses

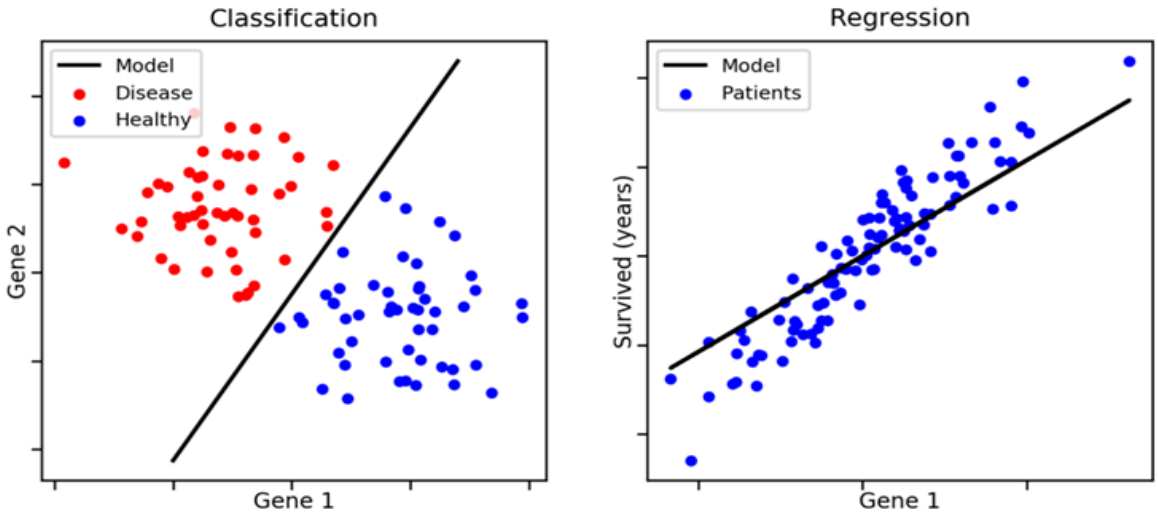

Figure 11 : Exemples de classification et de régression (15)

Les modèles d'apprentissage supervisé demandent beaucoup de travail préparatoire aux data scientistes. Les ensembles de données en entrée doivent être étiquetés, tandis qu'il faut indiquer les paramètres de sortie, les résultats attendus. Il faut également ajuster la précision pendant le processus d'apprentissage. (16)

3.1.1 Régression linéaire

Les algorithmes de régression linéaire sont les plus utilisés par les équipes de data science. Il s'agit d'effectuer des corrélations simples entre deux variables dans un jeu de données. Un

ensemble d'entrées et les sorties correspondantes sont examinés et quantifiés pour montrer une relation, par exemple comment le changement d'une variable affecte une autre. Les régressions linéaires sont représentées sous forme de lignes sur un graphique (16) .

La popularité de la régression linéaire s'explique par sa simplicité. L'algorithme est facilement explicable, relativement transparent et il y a peu de paramètres à configurer. Bien connu dans la pratique des statistiques, ce type d'algorithmes est souvent utilisé pour prévoir des ventes ou des risques (16).

3.1.2 Les Séparateurs à Vaste Marge (SVM)

Les Séparateurs à vastes marges SVM (connus aussi sous machines à vecteurs de support) sont des algorithmes qui séparent les données en classes. Pendant l'entraînement, un SVM trouve une ligne qui sépare les données d'un jeu en classes spécifiques et maximise les marges (les distances entre les frontières de séparation et les échantillons les plus proches) de chaque classe. Après avoir appris les lignes de classification, le modèle peut ensuite les appliquer aux nouvelles données (16).

Les spécialistes placent les SVM dans la catégorie des « classificateurs linéaires » : l'algorithme est idéal pour identifier des classes simples qu'il sépare par des vecteurs, nommés hyperplans, tel qu'il est présenté en Figure 12. Il est également possible de programmer l'algorithme pour les données non linéaires, que l'on ne peut pas séparer clairement par des vecteurs. Mais, avec des données d'entraînement hypercomplexes ; visages, traits de personnalité, génomes et matériel génétique – les systèmes de classes deviennent plus petits et plus difficiles à identifier et nécessitent un peu plus d'assistance humaine (16).

Les Séparateurs à vastes marges sont très utilisés dans la finance. Elles offrent une grande précision sur les données actuelles et futures. Les modèles associés peuvent servir à comparer virtuellement les performances financières relatives, la valeur et les retours sur investissement. Les SVM dits non linéaires sont souvent mis à contribution pour classifier des images (vision par ordinateur) ou des mots, des phrases et des entités (NLP) (16) . Les SVM peuvent être de deux types :

SVM linéaire : Les SVM linéaires sont utilisés pour les données linéairement séparables, ce qui signifie que si un ensemble de données peut être classé en deux classes en utilisant une

seule ligne droite, alors ces données sont appelées données linéairement séparables, et le classifieur utilisé est appelé classifieur SVM linéaire (17).

SVM non-linéaires : Les SVM non linéaires sont utilisés pour les données non linéairement séparées, ce qui signifie que si un ensemble de données ne peut pas être classé en utilisant une ligne droite, alors ces données sont qualifiées de données non linéaires et le classificateur utilisé est appelé classificateur SVM non linéaire (17).

Les SVM sont utilisés pour les problèmes de classification de texte telles que l'attribution de catégorie, la détection du spam ou encore l'analyse des sentiments. Ils sont également couramment utilisés pour les problèmes de reconnaissance d'image, particulièrement en reconnaissance de forme et en classification de couleur. Les SVM jouent également un rôle essentiel dans de nombreux domaines de la reconnaissance manuscrite des symboles, tels que les services d'automatisation postale (18).

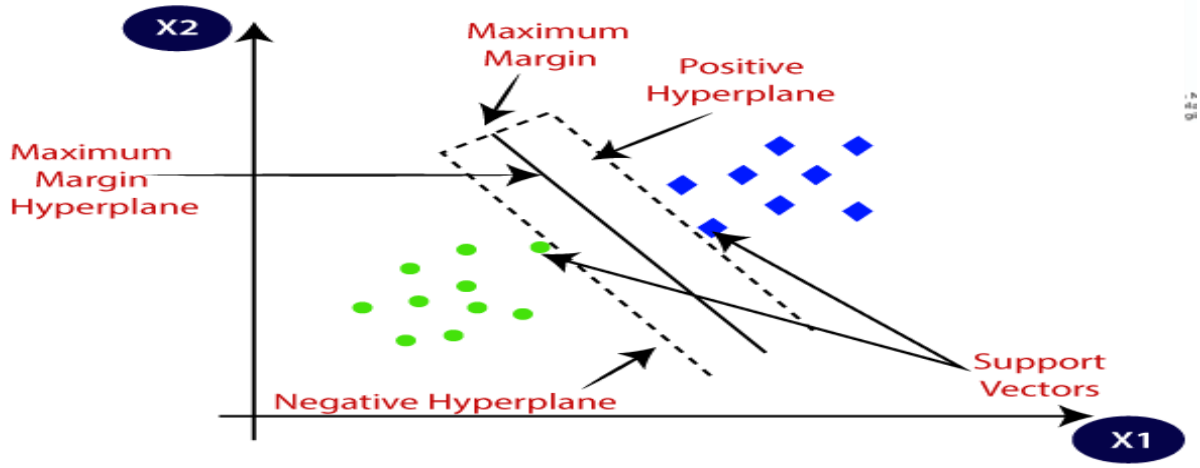

Figure 12 : Le modèle du SVM (17)

3.1.3 Arbre de décision (DT)

Un algorithme d'arbre de décision représente graphiquement les données en branches pour montrer les résultats possibles de diverses actions. Il classifie et prédit les variables de réponse en fonction des décisions passées. Cette méthode visuelle a fait ses preuves. Les résultats des arbres de décision sont faciles à expliquer. Les décisions et leurs impacts probables sur un résultat final sont aisément visibles, même lorsque les ensembles de données en entrée s'avèrent incomplets (16).

Les arbres de décisions deviennent difficiles à lire quand ils sont associés à de gros volumes de données et à des variables complexes. C'est pourquoi ils sont utilisés pour les décisions à faibles enjeux, comme l'anticipation des variations de taux d'emprunt ou les réactions du marché si une entreprise modifie un élément important d'un de ses produits (16).

3.1.4 Les Forêts d'arbres décisionnels (Random Forest RF)

Une Forêt d'arbres décisionnels est une technique d'apprentissage automatique utilisée pour résoudre les problèmes de régression et de classification. Elle utilise l'apprentissage d'ensemble, qui est une technique combinant plusieurs classificateurs pour fournir des solutions à des problèmes complexes. Cet algorithme compose de plusieurs arbres de décision. La "forêt" générée par l'algorithme de la Forêt d'arbres décisionnels est formée par agrégation en sac ou par agrégation bootstrap. Le bagging est un méta-algorithme d'ensemble qui améliore la précision des algorithmes d'apprentissage automatique (19).

L'algorithme établit le résultat sur la base des prédictions des arbres de décision. Il prédit en prenant la moyenne ou la moyenne des résultats de plusieurs arbres. L'augmentation du nombre d'arbres accroît la précision du résultat (19).

Une Forêt d'arbres décisionnels élimine les limites d'un algorithme d'arbre de décision. Elle réduit le surajustement des ensembles de données et augmente la précision. Elle génère des prédictions sans nécessiter de nombreuses configurations dans les paquets (comme scikit-learn) (19).

Voici quelques-unes des applications de la Forêt d'arbres décisionnels :

Banque **;** La Forêt d'arbres décisionnels est utilisée dans le secteur bancaire pour prédire la solvabilité d'un demandeur de prêt. Cela aide, dans l'établissement de crédit, à prendre une bonne décision quant à l'octroi ou non du prêt au client. Les banques utilisent également l'algorithme de Forêt d'arbres décisionnels pour détecter les fraudeurs. (19)

Soins de santé **;** Les professionnels de la santé utilisent des systèmes de Forêt d'arbres décisionnels pour diagnostiquer les patients. Les patients sont diagnostiqués en évaluant leurs antécédents médicaux. Les dossiers médicaux antérieurs sont examinés afin d'établir le bon dosage pour les patients (19).

3.1.5 Bayésien naïf (Naïves Bayes)

L'algorithme de bayésien naïf (Naïve Bayes) est un algorithme d'apprentissage supervisé, basé sur le théorème de Bayes et utilisé pour résoudre les problèmes de classification. Il est principalement utilisé dans la classification de textes qui comprend un ensemble de données d'entraînement à haute dimension (20).

Le classificateur bayésien naïf est l'un des algorithmes de classification les plus simples et les plus efficaces, qui permet de construire des modèles d'apprentissage automatique rapides, capables de faire des prédictions rapides. Il s'agit d'un classificateur probabiliste, ce qui signifie qu'il prédit sur la base de la probabilité d'un objet (20).

Certains exemples populaires de l'algorithme de bayésien naïf sont ; Evaluation du crédit bancaire, la classification des données médicales, il peut être utilisé pour les prédictions en temps réel car le classificateur bayésien naïf est un apprenant avide, il est utilisé dans la classification de textes tels que le filtrage du spam et l'analyse des sentiments et la classification d'articles, etc. (20).

3.1.6 K plus proches voisins (KNN)

K plus proches voisins est l'un des algorithmes d'apprentissage automatique les plus simples, basé sur la technique d'apprentissage supervisé. Cet algorithme suppose la similarité entre le nouveau cas/données et les cas disponibles et place le nouveau cas dans la catégorie qui est la plus similaire aux catégories disponibles (21).

L'algorithme KNN stocke toutes les données disponibles et classifie un nouveau point de données sur la base de la similarité. Cela signifie que lorsqu'une nouvelle donnée apparaît, elle peut être facilement classée dans une catégorie bien adaptée en utilisant l'algorithme KNN. Cet algorithme peut être utilisé aussi bien pour la régression que pour la classification, mais il est surtout utilisé pour les problèmes de classification. (21)

L'algorithme KNN est un algorithme non-paramétrique, ce qui signifie qu'il ne fait aucune hypothèse sur les données sous-jacentes. Il est également appelé un algorithme d'apprentissage paresseux parce qu'il n'apprend pas immédiatement à partir de l'ensemble d'apprentissage, mais il stocke l'ensemble de données et au moment de la classification, il effectue une action sur l'ensemble de données. Cet algorithme KNN, lors de la phase

d'apprentissage, se contente de stocker l'ensemble de données et, lorsqu'il reçoit de nouvelles données, il les classe dans une catégorie qui est très similaire aux nouvelles données (21).

Supposons qu'il existe deux catégories, voir Figure 13, à savoir la catégorie A et la catégorie B, et que nous avons un nouveau point de données x1, ce point de données se situera au milieu de ces catégories. Pour résoudre ce type de problème, nous avons besoin d'un algorithme KNN. Avec l'aide de KNN, nous pouvons facilement identifier la catégorie ou la classe d'un ensemble de données particulier (21).

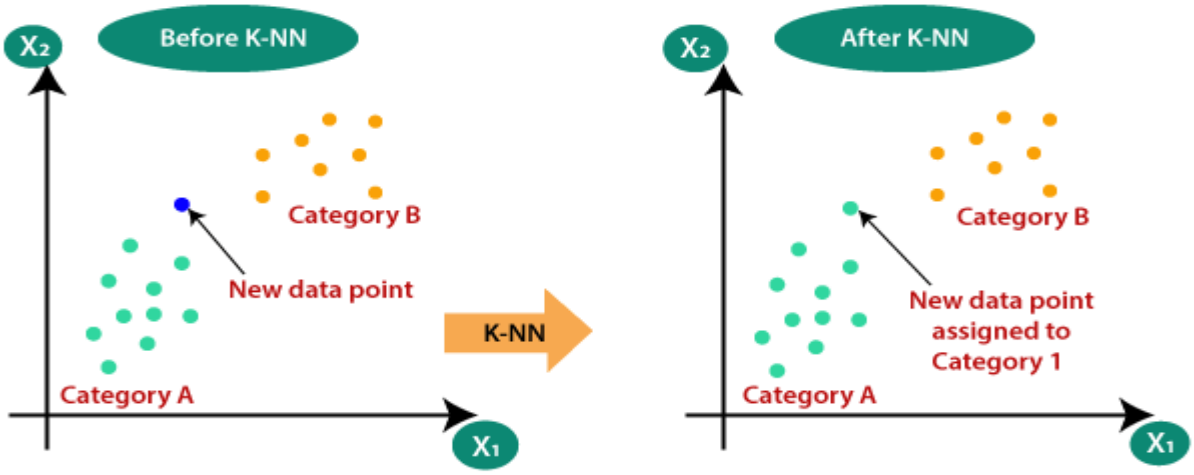

Figure 13 : Le modèle du KNN (21)

Le fonctionnement du KNN peut être expliqué sur la base de l'algorithme ci-dessous : (21)

- Étape 1 : Sélectionner le nombre K de voisins.
- Étape 2 : Calculer la distance euclidienne du nombre K de voisins.
- Étape 3 : Prenez les K voisins les plus proches selon la distance euclidienne calculée.
- Étape 4 : Parmi ces K voisins, compter le nombre de points de données dans chaque catégorie.
- Étape 5 : Attribuez les nouveaux points de données à la catégorie pour laquelle le nombre de voisins est maximal.
- Étape 6 : Notre modèle est prêt.

3.1.7 Comparaison des algorithmes d'apprentissage supervisé

Le Tableau 2 montre les avantages et les inconvénients des algorithmes d'apprentissage supervisé que nous avons vus précédemment.

Tableau 2: Les avantages et les inconvénients des algorithmes d'apprentissage supervisé. (17) (19) (20) (21)

Algorithmes	Avantages	Inconvénients
SVM	-Sa grande précision de prédiction -il fonctionne bien sur de petits ensemble de données (dataset) -Ils peuvent être plus efficace car ils utilisent un sous-ensemble de points d'entraînements.	-Ne convient pas à des jeux de données plus volumineux, car le temps d'entraînement avec les SVM peut être long. -Moins efficace sur les jeux de données contenant des valeurs aberrantes et du bruit.
KNN	-Il est simple à mettre en œuvre. -Il est robuste aux données de formation bruyantes. -Il peut être plus efficace si les données d'entraînement sont importantes.	-Il faut toujours déterminer la valeur de K, ce qui peut être complexe. -Le coût de calcul est élevé en raison du calcul de la distance entre les points de données pour tous les échantillons d'apprentissage.
NB	-Naïve Bayes est l'un des algorithmes de ML les plus rapides et les plus simples pour prédire une classe d'ensembles de données. -Il peut être utilisé pour les classifications binaires et multi-classes. -Il donne de bons résultats dans les prédictions multi-classes par rapport aux autres algorithmes. -C'est le choix le plus populaire pour les problèmes de classification de texte.	-Naïve Bayes suppose que toutes les caractéristiques sont indépendantes ou non liées, il ne peut donc pas apprendre la relation entre les caractéristiques.

RF	-Elle peut effectuer des tâches de régression et de classification. - Elle produit de bonnes prédictions qui peuvent être comprises facilement. -Elle peut traiter efficacement de grands ensembles de données. -Elle fournit un niveau de précision plus élevé dans la prédiction des résultats que l'algorithme d'arbre de décision.	L'utilisation d'une Random Forest nécessite davantage de ressources pour le calcul. -Il consomme plus de temps par rapport à un algorithme d'arbre de décision.
DT	-Il est simple à comprendre car il suit le même processus que celui que suit l'homme lorsqu'il prend une décision dans la vie réelle. -Il peut être très utile pour résoudre les problèmes liés à la prise de décision. -Il aide à réfléchir à toutes les issues possibles d'un problème. -Il est moins nécessaire de nettoyer les données que d'autres algorithmes.	-L'arbre de décision contient beaucoup de couches, ce qui le rend complexe. -Il peut présenter un problème de surajustement, qui peut être résolu à l'aide de l'algorithme forêt d'arbres décisionnels. -Pour plus d'étiquettes de classe, la complexité de calcul de l'arbre de décision peut augmenter.

3.2 Apprentissages non-supervisé

Quand le système ou l'opérateur ne dispose que d'exemples, mais non d'étiquette, et que le nombre de classes et leur nature n'ont pas été prédéterminés, on parle d'apprentissage non supervisé (Clustering en Anglais). Aucun expert n'est requis. L'algorithme doit découvrir par lui-même la structure plus ou moins cachée des données. Le partitionnement de données, data clustering en anglais, est un algorithme d'apprentissage non supervisé. Le système doit ici dans l'espace de description (l'ensemble des données) cibler les données selon leurs attributs disponibles, pour les classer en groupes homogènes d'exemples. (22)

L’apprentissage non supervisé est un sous-domaine de l’apprentissage automatique qui identifie des clusters ou des groupes basés sur des données non étiquetées avec peu d’intervention humaine, voir Figure 14.

L'apprentissage non supervisé est utilisé lorsque le problème nécessite une quantité massive de données non étiquetées. Par exemple, les applications de réseaux sociaux, telles que Twitter, Instagram et Snapchat, exploitent toutes de très grandes quantités de données non étiquetées. Pour comprendre le sens de ces données, il est nécessaire d'utiliser des algorithmes qui classifient les données en fonction des tendances ou des clusters qu'ils décèlent. L'apprentissage non supervisé mène un processus itératif, analysant les données sans intervention humaine. (12)

La similarité est généralement calculée selon une fonction de distance entre paires d'exemples. C'est ensuite à l'opérateur d'associer ou déduire du sens pour chaque groupe et pour les motifs (*patterns* en Anglais) d'apparition de groupes, ou de groupes de groupes, dans leur « espace ». Divers outils mathématiques et logiciels peuvent l'aider. On parle aussi d'analyse des données en régression (ajustement d'un modèle par une procédure de type moindres carrés ou autre optimisation d'une fonction de coût). Si l'approche est probabiliste (c'est-à-dire que chaque exemple, au lieu d'être classé dans une seule classe, est caractérisé par un jeu de probabilités d'appartenance à chacune des classes), on parle alors de « *soft clustering* » (par opposition au « *hard clustering* »). (22)

Cette méthode est souvent source de sérendipité ex : Pour un épidémiologiste qui voudrait dans un ensemble assez large de victimes de cancer du foie tenter de faire émerger des hypothèses explicatives, l'ordinateur pourrait différencier différents groupes, que l'épidémiologiste chercherait ensuite à associer à divers facteurs explicatifs, origines géographique, génétique, habitudes ou pratiques de consommation, expositions à divers agents potentiellement ou effectivement toxiques (métaux lourds, toxines telle que l'aflatoxine, etc.). (22)

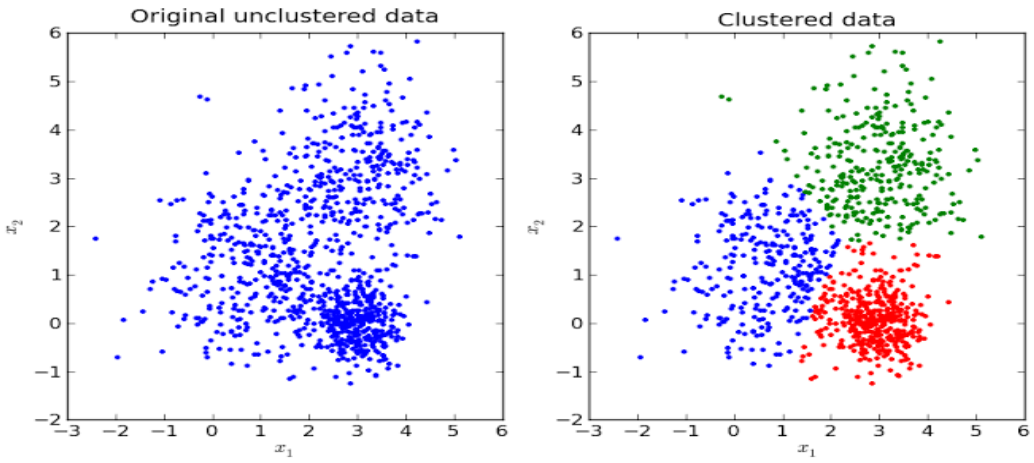

Figure 14 : L'apprentissage non supervisé (23)

3.2.1 Méthode des K-moyennes (K-means)

La méthode des K-moyennes, ou K-means clustering, est l'un des algorithmes de clustering les plus utilisés. Dans cette méthode, les points de données sont assignés à K groupes. Dans ce cas, K représente le nombre de groupes créés en fonction de la distance entre le noyau de chaque groupe. Ce noyau est aussi appelé le centroïde et est soit choisi au hasard, soit spécifié par le data scientist, en fonction des données. Une fois que le nombre de groupes (K) et les centroïdes ont été identifiés, le modèle assigne chaque nouveau point au noyau le plus proche et le groupe dans le cluster correspondant. La méthode la plus répandue pour calculer la distance entre un point et un noyau est le carré de la distance Euclidienne. (24)

L'un des principaux défis avec la méthode des K-moyennes est le fait que le nombre de clusters doit être spécifié avant le début du modèle. Ceci peut être compliqué par moment, surtout lorsque la quantité de données est très importante. Si une grande valeur pour K est choisie, le modèle rendra de plus petits groupes. A l'inverse, si K prend une petite valeur, alors le modèle rendra des grands groupes. (24)

Cette méthode est la plus souvent utilisée pour la classification de documents, la segmentation d'images et la segmentation marketing. (24)

3.2.2 Clustering hiérarchique

Le clustering hiérarchique (hierarchical clustering en Anglais) est un autre algorithme de clustering, qui créé une structure s'apparentant à un arbre. On appelle cette structure un dendrogramme. Ce type de clustering peut être divisé en deux catégories : les classifications

descendantes hiérarchiques et les classifications ascendantes hiérarchiques. Dans le cas d'une classification descendante hiérarchique, tous les points commencent par être assignés à un même groupe puis, lorsque le modèle est affiné, les points sont séparés en clusters jusqu'à ce qu'il y est un cluster pour chaque point. A l'inverse, dans le cas d'une classification ascendante hiérarchique, chaque point commence par être considéré comme son propre groupe puis, lorsque le modèle est affiné, des paires de clusters sont combinés, en fonction de leurs similarités, en un grand groupe contenant toutes les observations.

Comme pour la méthode des K-moyennes, les mesures de distances sont utilisées pour évaluer la similarité entre les points. Il existe quatre principales méthodes pour mesurer la similarité : (24)

Single linkage ; Dans cette méthode, la distance entre deux clusters correspond à la distance minimale entre deux points de chaque cluster : (24)

$$D(c_1, c_2) = \min D(x_1, x_2)$$

Complete linkage ; Dans cette deuxième méthode, la distance entre deux clusters correspond à la distance maximale entre deux points de chaque cluster : (24)

$$D(c_1, c_2) = \max D(x_1, x_2)$$

Average linkage ; Dans cette troisième méthode, la distance entre deux clusters correspond à la moyenne des distances entre toutes les paires de points dans chaque groupe : (24)

$$D(c_1, c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\sum\sum D(x_1, x_2)$$

Ward's linkage ; Dans cette quatrième méthode, la distance entre deux clusters correspond à l'augmentation de la somme des carrés, après que chaque cluster a été combiné. Le but est de minimiser la variance totale entre clusters. (24)

Dans ces quatre méthodes, la distance Euclidienne est la mesure d'évaluation la plus utilisée pour calculer les distances entre points. (24)

3.2.3 Algorithme Apriori

L'algorithme Apriori est l'un des algorithmes d'apprentissage non supervisé, qui peut être classifié comme une règle associative. En effet, cette technique utilise une approche ascendante, dans laquelle les points ou les collections de points les plus fréquents sont

identifiés et utilisés pour établir des règles d'association. Cet algorithme est basé sur l'idée qu'un sous-groupe d'un groupe fréquent est également un groupe fréquent. (24)

Les algorithmes Apriori ont gagné en popularité lorsqu'ils ont commencé à être utilisés pour l'analyse du panier de consommation ou pour les recommandations musicales dans les applications les plus répandues. En effet, cet algorithme permet d'établir la probabilité qu'un individu achète ou écoute un élément X sachant que il/elle a acheté ou écouté l'élément Y. Ainsi, ce modèle nécessite les comportements passés d'un individu afin de faire des prédictions sur les comportements futurs. (24)

3.3 Apprentissage semi-supervisé

L'apprentissage semi-supervisé est une classe de techniques d'apprentissage automatique qui utilise un ensemble de données étiquetées et non étiquetées. Il se situe ainsi entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non-supervisé qui n'utilise que des données non étiquetées. Il a été démontré que l'utilisation de données non étiquetées, en combinaison avec des données étiquetées, permet d'améliorer significativement la qualité de l'apprentissage. (25)

Un autre intérêt provient du fait que l'étiquetage de données nécessite souvent l'intervention d'un utilisateur humain. Lorsque les ensembles de données deviennent très grands, cette opération peut s'avérer fastidieuse. Dans ce cas, l'apprentissage semi-supervisé, qui ne nécessite que quelques étiquettes, revêt un intérêt pratique évident. (25)

Un exemple d'apprentissage semi-supervisé est le Co apprentissage, dans lequel deux classifier apprennent un ensemble de données, mais en utilisant chacun un ensemble de caractéristiques différentes, idéalement indépendantes. Si les données sont des individus à classer en hommes et femmes, l'un pourra utiliser la taille et l'autre la pilosité par exemple (25).

Les méthodes d'apprentissage semi-supervisé combinent données étiquetées et non étiquetées. Les algorithmes de ce type se nourrissent de certaines informations grâce à des catégories labélisées, des suggestions et des exemples. Ensuite, ils créent leurs propres labels en explorant les données par eux-mêmes, en suivant un schéma rudimentaire ou les indications de data scientists (16).

3.4 Apprentissages par renforcement

En intelligence artificielle, plus précisément en apprentissage automatique, l'apprentissage par renforcement consiste, pour un agent autonome (robot, etc.), à apprendre les actions à prendre, à partir d'expériences, de façon à optimiser une récompense quantitative au cours du temps. L'agent est plongé au sein d'un environnement, et prend ses décisions en fonction de son état courant. En retour, l'environnement procure à l'agent une récompense, qui peut être positive ou négative. L'agent cherche, au travers d'expériences itérées, un comportement décisionnel (appelé stratégie ou politique, et qui est une fonction associant à l'état courant l'action à exécuter) optimal, en ce sens qu'il maximise la somme des récompenses au cours du temps (26).

L'apprentissage par renforcement est un modèle d'apprentissage comportemental. L'algorithme reçoit un feedback de l'analyse des données et guide l'utilisateur vers le meilleur résultat. L'apprentissage par renforcement diffère des autres types d'apprentissage supervisé, car le système n'est pas formé avec un ensemble de données exemple. Au lieu de cela, le système apprend plutôt par le biais d'une méthode d'essais et d'erreurs. Par conséquent, une séquence de décisions fructueuses aboutit au renforcement du processus, car c'est lui qui résout le plus efficacement le problème posé (12).

4 Conclusion

Dans ce chapitre, nous avons expliqué l'apprentissage automatique en tant que l'un des outils les plus importants de l'intelligence artificiel et sa relation avec ce domaine. Le rôle principal des algorithmes d'apprentissage automatique est de prédire correctement la sortie. On peut dire qu'il apprend et se développe à partir de l'ensemble de données, jusqu'à ce qu'il forme un modèle précis capable de prédire la sortie avec le taux d'erreur le plus faible possible. Les algorithmes d'apprentissage automatique sont principalement classés par méthode d'apprentissage automatique en 4 méthodologies : les algorithmes d'apprentissage supervisé, les algorithmes d'apprentissage non supervisé, les algorithmes d'apprentissage semi supervisé et les algorithmes d'apprentissage par renforcement. Les techniques d'apprentissage automatique sont plus simples et efficaces pour le domaine d'analyse des sentiments médicaux.

CHAPITRE 3 :

Analyse du sentiment des revues de médicaments

Chapitre 3 : Analyse du sentiment des revues de médicaments

1 Introduction

Aujourd'hui, l'analyse des sentiments est d'une grande importance dans de nombreux domaines comme le domaine médicale. Actuellement, les réseaux sociaux et les sites web regorgent d'avis et de commentaires d'utilisateurs sur les produits pharmaceutiques, et les services médicaux. L'analyse de ces sentiments et opinions est très importante pour satisfaire les besoins des consommateurs. Afin de catégoriser les commentaires et les besoins des consommateurs, nous utilisons des algorithmes d'apprentissage automatique.

Dans ce dernier chapitre de notre mémoire nous présentons notre contribution dans la prédiction médicale dans le cas de l'analyse des sentiments des patients sur les revues des médicaments. Nous fournirons également tous les outils et packages utilisés dans ce travail. Nous présenterons les ensembles de données que nous avons utilisé et expliquerons toutes les étapes nécessaires que nous avons suivies pour nettoyer, traiter et préparer les données avant d'appliquer les algorithmes d'apprentissage automatique supervisé. Nous présenterons et comparerons également les résultats obtenus. A la fin nous présenterons une application web que nous avons créés afin de tester des modèles sur des données concrètes.

2 Présentation des outils utilisés

2.1 Le langage Python

Python est un langage de programmation de haut niveau, interactif, interprété et orienté objet. Il utilise également moins de formule ou de syntaxe (voir Figure 15) . (27) Python a de nombreuses bibliothèques dans divers domaine (tels que l'intelligence artificiel, l'analyse des données, la cyber sécurité, etc.).

- Interactif : cela signifie que vous pouvez directement interagir et interpréter les instructions à l'aide de l'invite de commande Python lors de l'écriture de votre programme.

- Interprété : Python est traité au moment de l'exécution par un interpréteur, ce qui signifie que vous n'avez pas besoin de compiler votre programme avant de l'exécuter. Identique aux langages de programmation Matlab, PHP et PERL.
- Orienté objet : Python prend également en charge l'approche orientés objet ou les techniques de programmation qui encapsulent le code dans les objets. (27)

Python possède plusieurs fonctionnalités uniques parmi les langages de programmation, notamment les suivant :

- Le langage de programmation python possède une grammaire et des scripts très faciles à apprendre
- Le langage de programmation python dispose d'un système de gestion automatique des données et de la mémoire
- Le langage de programmation Python dispose de nombreuses fonctionnalités de prise en charge et facilite la tâche de ses utilisateurs (27).

Figure 15 : logo de python

Les avantages du langage de programmation Python sont les suivants :

- ✓ Facile à apprendre : Facile à apprendre est déjà attaché comme l'un avantages du langage de programmation Python parmi d'autres langages de programmation. Ce langage de programmation Python a une syntaxe assez simple et facile à comprendre.
- ✓ Facile à utiliser : Un autre avantage du langage de programmation Python, ce langage de programmation est un langage facile à utiliser pour développer un produit, que ce soit un Web, des logiciels, des applications Web, des jeux vidéo, etc. En plus d'avoir une lisibilité élevée du code, donc le code est facile à comprendre, ce langage de programmation a une bibliothèque très grande et étendue. Les différents types de bibliothèques contiennent beaucoup d'équipements et de fonctionnalités qui sont très

extraordinaires, de sorte que la facilité de création de programmes est celle offerte par le langage de programmation.

- ✓ Soutenez bien l'Internet des objets (Internet of Things IoT) : L'une des forces du langage de programmation Python est qu'il supporte très bien l'écosystème de l'Internet des objets. L'Internet des objets est une technologie qui relie les objets autour de nous ou de notre environnement en un réseau qui se connecte les uns aux autres. Une technologie qui transporte tout ce qui est connecté dans un réseau Internet est inséparable de la nécessité de langages de programmation dans le développement du système. Et le langage de programmation Python offre un très bon support pour cette technologie. (27)

2.1.1 Le NLTK

La bibliothèque NLTK (Natural Language Toolkit) est une bibliothèque open source. Cette bibliothèque est une suite qui contient des bibliothèques et des programmes pour le traitement statistique du langage. Il s'agit de l'une des bibliothèques NLP les plus puissantes, qui contient des packages permettant le traitement automatique du langage naturel (28). Dans ce travail nous avons utilisé le « SnowballStemmer » et la liste des mots vides de la bibliothèque NLTK.

2.1.2 Pandas

La bibliothèque Pandas est une bibliothèque de manipulation et l'analyse des données écrite pour le langage de programmation Python. En particulier, elle fournit des structures de données et des opérations pour manipuler des séries temporelles et des tableaux numériques. (29)

- La bibliothèque fournit ce que l'on appelle Data Frame pour l'importation et la manipulation des données. (30)
- La bibliothèque fournit les capacités nécessaires pour importer des données à partir de fichiers dans des différents formats (telle que format csv, format Excel, etc.). (30)
- La bibliothèque facilite les opérations de prétraitement des données telles que le nettoyage des données, le traitement des valeurs vides qu'elles contiennent et l'exécution d'opérations exploratoires sur les données. (30)

2.1.3 Scikit-learn

La bibliothèque Scikit-learn est une bibliothèque python dédiée à l'analyse de données. Il s'agit d'une bibliothèque facile à utiliser et offre des fonctionnalités puissantes. Scikit-learn s'intègre naturellement dans l'ensemble des outils d'analyse de données.

Scikit-learn, également connu sous le nom de Sklearn, est la bibliothèque d'apprentissage automatique la plus robuste de python. Elle offre un choix d'outils efficace d'apprentissage automatique que nous avons utilisé dans ce mémoire à savoir : les classifieurs SVM, NB, DT, RF, et KNN.

Cette bibliothèque est écrite principalement en Python est basée sur NumPy, SciPy et Matplotlib (31).

2.1.4 Streamlit

L'environnement (En anglais, Framework) Streamlit est une environnement open source en python. Streamlit va nous aider à créer des applications web d'apprentissage automatique. Elle est compatible avec les principales bibliothèques python que nous avons utilisé dans le cas de ce travail telles que ; Scikit-learn, Pandas, NumPy, etc.

2.1.5 Jupyter notebook

Le Jupyter Notebook est une application Web open source qui vous permet de créer et de partager des documents contenant du code en direct, des équations, et visualisations. Elle utilise notamment le nettoyage et la transformation des données, la visualisation des données, l'apprentissage automatique et bien plus encore (32).

Le plus souvent, Jupyter Notebook est utilisé dans un environnement Python. Ils ont des sorties très interactives et peuvent être facilement partageables (32).

2.1.6 Spyder

Spyder (En Anglais, Scientific python developpement environment) est un environnement scientifique puissant et open source écrit en Python. Il convient à l'apprentissage automatique

ainsi qu'à l'analyse de données. Il est compatible avec les bibliothèques python telles que ; NLTK, NumPy, Pandas, Sklearn, Pickle, Streamlit, etc.

3 L'ensemble de données

La quantité et la qualité des données ont un impact significatif sur la précision et l'efficacité du modèle. Plus leur nombre et leur fiabilité sont élevés, plus le résultat obtenu est précis. Dans le cadre de ce mémoire nous avons utilisé deux ensembles de données. Le premier ensemble de données est collecté et annoté par dans ce travail de master. Afin de comparer les résultats de performance de notre corpus avec d'autres corpus de référence, nous avons utilisé l'ensemble de données drugsCom télécharger à partir de site web Kaggle[2].

3.1 Création de l'ensemble de données Cymbalta_drug_dataset

Le site web que nous avons exploré s'appelle askapatient.com, ce qui permet aux patients de partager leurs expériences médicamenteuses et comparaison avec les expériences d'autre patients. Ce site a reçu le Web by Award 2012 du meilleur site web dans la catégorie pharmaceutique. Le askapatient.com[3] a été créé pour trois raisons principales, qui sont les suivantes ; Satisfaction des patients, analyse des sentiments et pharmacovigilance.

Les données utilisées pour la création de notre corpus « Cymbalta_drug_dataset » contiennent des commentaires écrits par des patients en fonction de leur expérience sur l'utilisation du médicament « Cymbalta ». Le médicament Cymbalta est un antidépresseur inhibiteur sélectif de la recapture de la sérotonine et de la norépinéphrine (ISRSN). La duloxétine agit sur les substances chimiques du cerveau qui peuvent être déséquilibrées chez les personnes souffrant de dépression. Ce médicament est utilisé pour traiter le trouble dépressif majeur chez les adultes. Il est également utilisé pour traiter le trouble anxieux général chez les adultes et les enfants âgés d'au moins 7 ans. (33)

Nous avons collecté les données manuellement car il est interdit d'accéder au site à l'aide de robots d'exploration ou de bots automatisés.

Les principales étapes de création de Cymbalta_drug_dataset sont :

[2] www.kaggle.com
[3] www.askapatient.com

- D'abord, nous explorons le site web dont nous avons parlé plus tôt, puis choisissons une catégorie de traitement parmi les dix meilleurs catégories (Top 10 Catégories) des traitements. Cette catégorie de traitement que nous avons choisie s'appelle la dépression.
- Par la suite, nous avons choisi un médicament appelé Cymbalta qui est prescrit pour la dépression.
- Après, nous recueillons des ensembles des commentaires (COMMENTS) et des évaluations (RATING), Chaque commentaire est accompagné d'une note 1 à 5.
- Puis, nous créons l'ensemble de données, Cymbalta_drug_dataset, dans MS Excel où on a stocké les données que nous avons collectées précédemment.
- Enfin, nous convertissons l'ensemble de données Excel en format CSV à savoir Cymbalta_drug_dataset.csv.

3.1.1 Description de l'ensemble des données Cymbalta_drug_dataset

Le Tableau 3 décrit les variables contenues dans l'ensemble de données utilisé dans notre travail. Dans le Tableau 4 : Exemples de données de Cymbalta_drug_dataset des exemples de commentaires du corpus Cymbalta_drug_dataset sont présentés.

Tableau 3 : Description des variables de Cymbalta_drug_dataset

Variables	**Description**	**Types**
COMMENTS	Ce sont les sentiments et les opinions des patients sur le médicament après avoir utilisé le médicament	Object
RATING	Evaluation du médicament par les patients (1 à 5).	Int 64

Tableau 4 : Exemples de données de Cymbalta_drug_dataset

RATING	**COMMENTS**
1	This drug causes dependence. Almost impossible to get off of NEVER start taking it
2	Don't takeit.
3	Helps pain immensely, but I'm zonked out most of the day and when I'm

	up I'm eating like a hog.
4	I feel much more like myself. Although I am not motivated yet the horrible darkness has gone and I have a lot of good days
5	Works extremely well with aripiprazole 5mg. The constant negative and worried thinking has gone. I feel peaceful and even joyfu

3.1.2 Annotation du corpus « Cymbalta_drug_dataset »

Ces commentaires sont accompagnés des notes d'évaluation de 1 jusqu'à 5. Ces notes sont postées par les visiteurs du site web avant d'écrire leurs commentaires. Nous avons utilisé ces notes d'évaluation pour l'annotation de l'orientation sentimentale des commentaires. L'annotation faite dans ce travail est alors automatique. Dans l'annotation nous avons considéré un commentaire accompagné d'une note inférieur à trois (rating 3) comme étant négatif. Pour un commentaire qui possède une note supérieure à trois, il est considéré comme positif. Les commentaires ayant une note de trois sont considéré neutres et sont éliminés de notre ensemble de données. Le Tableau 5 présente les statistiques du Cymbalta_drug_dataset. On peut remarquer que le nombre de commentaires positifs est légèrement inférieur au nombre de commentaires négatifs. Donc, on peut considérer que notre corpus est équilibré.

Tableau 5 : Statistiques de Cymbalta_drug_dataset

Sentiment	**Nombre de commentaires**	
Négatif	1005	54.00%
Positif	856	46.00%
Total	1861	100%

3.2 L'ensemble de données drugsCom

Le corpus drugsCom est utilisé dans le Hackathon de l'hiver 2018 de Kaggle Université Club. Le corpus est actuellement disponible publiquement (34).

3.2.1 Description de l'ensemble des données drugsCom

Le corpus drugsCom offre des commentaires des patients sur des médicaments spécifiques accompagnés des conditions liées à savoir : l'identificateur du commentaire (uniqueID), le nom du médicament (drugName), la condition d'utilisation (condition), le commentaire (review), l'évaluation (rating), la date du commentaire, et le nombre d'utilisateurs qui trouvent le commentaire utile (usefulCount).

Le corpus drugsCom comporte plus de 70000 commentaires. Pour les limites de performances de notre PC nous nous limitons à trois sous-ensembles de 10000 commentaires chacun nommés ; drugsCom_reduced1, drugsCom_reduced2, et drugsCom_reduced3. Le Tableau 6 présente les statistiques des sous-ensembles de données générés à partir de corpus drugsCom.

Tableau 6 : Statistiques de l'ensemble de données drugsCom

	drugsCom_reduced1		drugsCom_reduced2		drugsCom_reduced3	
Positif	4890	73.32%	4865	72.70%	5001	73.66%
Négatif	1779	26.68%	1827	27.30%	1788	26.34%
Total	6669	100%	6692	100%	6789	100%

Depuis ce tableau on peut voir que les trois sous-ensembles de données ne nous sont pas équilibrés. A la différence de l'ensemble de données Cymbalta_drug_dataset, on remarque dans ces trois ensembles de données que le nombres de commentaires positifs est dominat.

3.2.2 Annotation du corpus « drugsCom »

L'évaluation (rating) des commentaires de drugsCom sont étalé sur une échelle de 10 étoiles reflétant la satisfaction du patient envers le médicament. Comme nous sommes intéressé par une classification binaire, i.e. Positif vs. Négatif, une annotation binaire est faite. Dans le premier temps, les commentaires ayant une note d'évaluation de 5 sont considérer neutres et éliminer de notre corpus. Dans l'annotation nous avons considéré les commentaires avec plus de 5 points comme positifs, et ceux avec moins de 5 points comme négatifs.

4 Architecture de notre application

Dans le cadre d'analyse des sentiments, nous choisissons l'analyse des sentiments dans les avis des patients sur les médicaments. Elle consiste à classer ces opinions et les sentiments des patients à propos de leur expérience médicamenteuse en positifs et négatifs. Ce processus aide les patients à prendre une décision concernant l'achat et l'utilisation de médicaments et de services médicaux. L'analyse de sentiments aide également à déterminer les avantages et les inconvénients de ces médicamentes, permettant aux entreprises médicales d'améliorer ses produits et services pharmaceutiques.

Les principales étapes de l'analyse des sentiments dans notre application sont ; l'importation de l'ensemble de données, le nettoyage de données, le prétraitement des documents textuels, l'extraction des caractéristiques, et la classification de données (voir Figure 16).

Nous avons d'abord commencé par la création l'ensemble de données (dataset), Cette étape vise à collecter les données nécessaires pour la création de l'ensembles de données à savoir les commentaires sur les médicaments.

Après, nous avons procédé par un prétraitement des commentaires de l'ensembles de données. Le prétraitement comporte le nettoyage des données et la suppression des données inutiles. Ensuite, nous avons effectué une Tokenisation, suppression des mots vides et un stemming. Tout cela pour réduire l'espace de représentation des commentaires et améliorer la qualité des modèles de classification générés.

Puis, nous avons effectué une extraction des caractéristiques. Cette étape est importante car le type d'entités extraites et la manière dont elles sont construites influence sur la performance des méthodes d'apprentissage automatique. L'extraction des caractéristiques est faite dans le cadre de ce mémoire via la méthode TF-IDF.

Dans l'étape suivante, nous avons passé à l'analyse de sentiments à travers une classification à base d'apprentissage automatique. Les méthodes d'apprentissage automatique supervisés utilisés dans ce travail sont ; SVM, NB, RF, DT et K-NN.

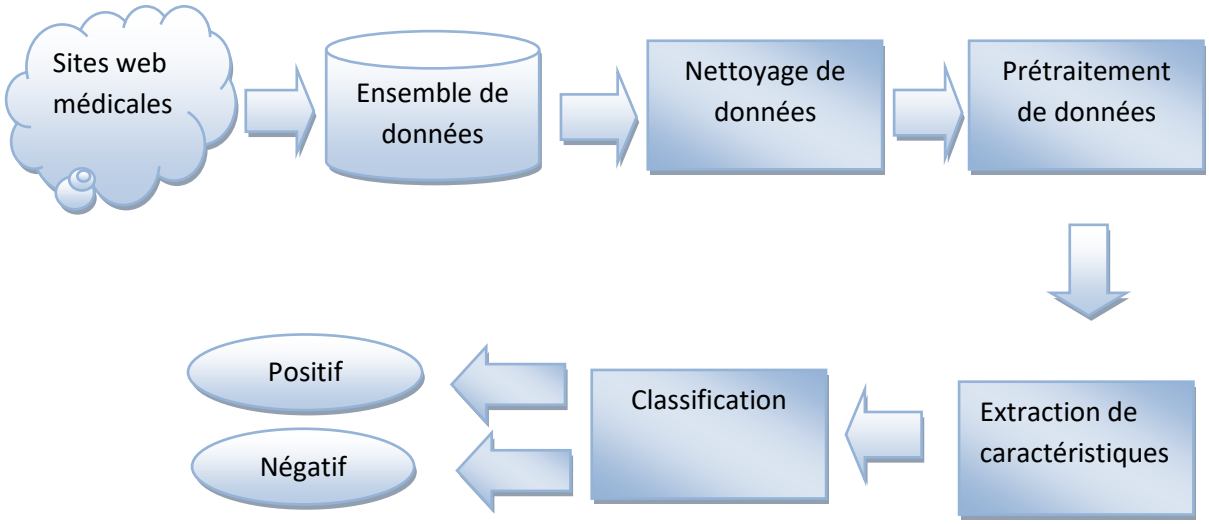

Figure 16 : Les principales étapes de l'analyse du sentiment de revues des médicamentes

4.1 Importation de l'ensemble de données

Dans cette étape, l'ensemble de données, i.e. le dataset à utiliser, va être importé dans notre système pour subir un ensemble de traitement. Dans cette étape, nous avons importé notre ensemble de données créé dans le cadre de ce travail de mémoire.

4.2 Nettoyage des données

Le nettoyage des données (data cleaning) est une étape très importante pour améliorer la qualité des données et gérer les données brutes. Dans cette étape, nous avons filtré toutes les lignes vides. Nous avons également supprimé toutes les lignes avec des valeurs manquantes, et les ligne en double.

4.3 Prétraitement des données

C'est est une étape importante dans l'analyse des sentiments, pour obtenir les meilleures performances des outils de traitement automatique de la langue. La qualité des données doit être vérifiée avant d'appliquer les algorithmes d'apprentissage automatique. Dans la première étape, nous avons effectué un prétraitement préliminaire pour garder seulement les caractères de l'alphabet. Ensuite, nous avons effectué le processus de division du texte en unités appelée tokens utilisant les espaces les singes de ponctuation. Puis nous avons supprimer les mots

vides (comme a, the, etc.) utilisant la bibliothèque NLTK. Enfin nous avons effectué la technique d'enracinement (stemming).

Le prétraitement de données comprend les étapes suivantes ; le prétraitement préliminaire, la séparation des mots, la suppression des mots vides, et l'enracinement.

4.3.1 Prétraitement préliminaire

Dans cette étape, nous avons supprimé les caractères spéciaux (tel que les symboles, les chiffres, les signes de ponctuation, etc.) utilisant la bibliothèque des expressions régulières en python (bibliothèque re). Nous avons préservé uniquement les caractères appartenant à l'alphabet. Nous avons également converti les caractères du texte en lettres minuscules.

4.3.2 Séparation de mots (Tokenization)

La Séparation de mots (tokenization) est le processus qui consiste à décomposer un texte donné en unités appelées tokens correspondants aux mots individuels. Au cours du processus de tokenisation, certains caractères comme les signes de ponctuation peuvent être éliminés. Les tokens deviennent alors l'entrée pour les étapes suivantes tels que la suppression de mots vides et l'enracinement (35). La Figure 17 présente un exemple de séparation de mots d'un commentaire de notre corpus, i.e., Cymbalta_drug_dataset.

```
['this', 'drug', 'should', 'not', 'be', 'on', 'the', 'market', 'it',
'is', 'pure', 'hell', 'and', 'even', 'worse', 'when', 'you', 'try',
'to', 'get', 'off', 'of', 'it', 'it', 'did', 'absolutely', 'nothing',
'for', 'my', 'depression', 'i', 'am', 'praying', 'that', 'i', 'will',
'be', 'able', 'to', 'wean', 'off', 'of', 'it', 'as', 'soon', 'as',
'possible', 'in', 'the', 'meantime', 'it', 'is', 'pure', 'hell',
'please', 'do', 'not', 'even', 'consider', 'it', 'the', 'drug',
'maker', 'should', 'be', 'sued']
```

Figure 17:exemple de séparation de mots d'un commentaire

4.3.3 Suppression des mots vides (stop words removal)

Les mots vides (aussi appelé les mots d'arrêt) sont des mots qui sont répétés fréquemment et n'ont aucun effet sur les sentiments dans le texte. Essentiellement les mots vide sont un ensemble de mots couramment utilisés dans une langue. La suppression des mots vides permet d'économiser de l'espace dans le vecteur de mots, ce qui permet d'obtenir de bons résultats et d'optimiser le temps d'exécution.

Dans cette étape, nous supprimons les mots vides de l'Anglais qui ne sont pas importants dans le processus de l'analyse des sentiments. Dans ce mémoire nous avons utilisé la bibliothèque NLTK qui offre la liste des mots vides présenté dans la Figure 18.

```
{"you'll", 'at', 'the', 'these', 'during', "that'll", 'against', 'so',
'should', 're', 'his', 'your', 'there', 'theirs', 'not', 'herself',
'an', "isn't", 'myself', 'very', 'ain', 'more', 'hers', 'it', 'her',
'do', 'about', "hasn't", "shan't", 'had', 'too', 'have', 'after',
"you've", "she's", "should've", 'own', 'mustn', 'for', 'few', 'a',
'who', 'down', 'me', 'or', 'both', 'don', 'ma', 'weren', 'where',
'couldn', 'having', "hadn't", 'now', 'what', 'i', 'this', 'than', 'd',
'its', "won't", 'being', 'shan', 'we', 'once', 've', 'will', 'they',
'has', 'when', "mightn't", 'themselves', 'needn', 'am', 'but', 'whom',
'doesn', 'yourself', 'was', 'ourselves', 'be', 'hadn', 'y', 'over',
'ours', 'm', 'shouldn', 'which', 'from', 'been', 'he', 'with', 'no',
'o', 'll', 'off', 'isn', 'to', "it's", 'on', 'won', "you'd", 'here',
'can', "didn't", 'further', "aren't", 'she', 'did', "couldn't",
'above', 'their', 'between', 'before', 'all', 'is', 'each', 'them',
'and', 'you', "doesn't", 'if', 'same', "haven't", 'any', 'while',
"weren't", 'didn', 'haven', 'through', 'under', 'some', 's', 'nor',
"wouldn't", 'hasn', 'again', "you're", 't', 'aren', 'wouldn', 'only',
'that', 'were', 'himself', "don't", 'out', 'in', 'my', 'by', 'mightn',
'into', 'itself', 'doing', 'then', 'wasn', 'why', 'such', 'how',
'other', "wasn't", "shouldn't", 'those', 'as', 'are', 'yours',
'because', 'our', 'him', 'until', 'most', 'yourselves', 'up',
"needn't", "mustn't", 'just', 'below', 'of', 'does'}
```

Figure 18: Liste de mots vides de l'Anglais de NLTK

4.3.4 Enracinement (Stemming)

Enracinement est une étape très importante dans le domaine de l'analyse des sentiments. Dans cette opération les mots similaires seront remplacés par leur racine (stem), cela pour réduire le nombre de vocabulaire utilisé. Dans ce travail, nous avons utilisé le processus d'enracinement afin d'éliminer la redondance dans le vocabulaire, ce qui aide à obtenir des résultats efficaces utilisant les algorithmes d'apprentissage automatique.

Dans ce travail nous avons utilisé l'algorithme Snowball d'enracinement. Snowball est un algorithme d'enracinement très utile offert par la bibliothèque NLTK. Il supporte l'Anglais en plus de 15 autres langues (36).

4.4 Extraction des caractéristiques

Les techniques d'extraction de caractéristiques permettant de convertir un ensemble de données textuelles en vecteurs de caractéristiques. Ces vecteurs vont être exploité par les algorithmes d'apprentissage automatique.

Il existe dans la littérature de nombreuses techniques pour convertir des données textuelles en vecteurs.

Dans cette étape nous avons utilisé une technique très populaire pour transformer notre ensemble de données en vecteurs de caractéristiques numériques. Cette représentation est nécessaire pour les algorithmes de l'apprentissage automatique supervisés, où chaque document est représenté sous la forme d'un vecteur de caractéristiques pondérées. La technique d'extraction de caractéristiques que nous avons utilisée est TF-IDF qui signifie « Fréquence du terme - Fréquence de document inverse » (1).

TF-IDF (Fréquence du terme - Fréquence de document inverse) Il s'agit d'une technique de pondération couramment utilisée pour la recherche et l'exploration de l'information. TF-IDF c'est une méthode statistique, Utilisé pour évaluer l'importance d'un mot dans un documents dans un ensemble de documents. L'importance du mot augmente proportionnellement au nombre de fois qu'il apparaît dans le fichier, Mais en même temps, il diminue inversement avec la fréquence à laquelle il apparaît dans le corpus. (37) TF-IDF est calculé avec la formule suivants :

$$TF - IDF = TF \times IDF$$

TF Fréquence du terme (Term Frequency en anglais) explique et montre l'importance du terme dans le document. Si la fréquence du terme est plus élevée dans le document, l'importance est plus grande. TF est calculé avec la formule suivante :

$$TF(k) = \frac{le\ nombre\ de\ fois\ que\ le\ terme\ k\ apparaît\ dans\ le\ document}{la\ somme\ des\ termes\ dans\ le\ document}$$

IDF Fréquence de document inverse (Inverse Document Frequency en anglais) explique et montre l'importance du terme dans ensembles des documents. IDF se concentre sur l'analyse de l'apparence du terme dans tous les documents :

$$IDF(k) = \log \frac{le\ nombre\ total\ de\ documents}{le\ nombre\ de\ documents\ contenant\ le\ terme\ k}$$

4.5 La Classification

Pour la classification des sentiments des malades nous avons utilisé cinq classifieurs qui sont ; les Séparateurs à Vaste Marge (SVM), le Naïve Bayes (NB), Randon Forest (RF), les arbres de décisions (DT), le K-voisins les plus proches KNN.

4.6 Mesures d'évaluation

La performance des algorithmes d'apprentissage automatiques est directement liée à leur capacité à prédire un résultat. Ainsi lorsque l'on cherche à comparer les résultats d'algorithmes avec la réalité, on utilise la matrice de confusion.

La matrice de confusion est un outil dans le domaine de l'apprentissage automatique, utilisée pour tester les performances des algorithmes de classification. Cette matrice est un tableau à 4 valeurs représentant les différentes informations sur les valeurs attendues par le classifieur et les valeurs réelles (résultant de l'annotation de l'ensemble de données).

Chaque ligne dans le Tableau 7 représente la catégorie réelle et chaque colonne représente la catégorie attendue.

Tableau 7 : Matrice de confusion

	Négative : 0	Positive: 1
Négative : 0	True Négative: TN	False Négative : FN
Positive: 1	False Positive : FP	True Positive: TP

La signification de TP, TN, FP et FN est comme suit :

- **TN** : True Négative
- **TP**: True Positive.
- **FP** : False Positive
- **FN** : False Négative.

TN : la prédiction est négative et c'est la réalité (l'ensemble de données)

- Signifie qu'un commentaire est réellement négative et elle a été prédite qu'elle est négative.

TP : la prédiction est positive et c'est la réalité.

- Signifie qu'un commentaire est réellement positive et elle a été prédite qu'elle est positive.

FP : la prédiction est positive mais ce n'est pas la réalité.

- Signifie qu'un commentaire est réellement négative et elle a été prédite qu'elle est positive.

FN : la prédiction est négative mais ce n'est pas la réalité

- Signifie qu'un commentaire est réellement positive et elle a été prédite qu'elle est négative.

Donc, les métriques nécessaires pour analyser ce tableau (matrice de confusion sont ; précision, rappel, score F1.

- Précision : capacité du modèle de classification à ne renvoyer que des cas Lié, autrement dit, ici, le nombre de vrais positifs divisé par la somme de nombre des vrais positifs (TP) et le nombre de faux positifs (FP).

$$précision = \frac{TP}{TP+FP}$$

- Rappel : (recall en anglais) est la capacité du modèle de classification identifier tous les cas pertinents. Il se calcule comme le nombre de vrais positifs Divisé par le nombre de vrais positifs plus le nombre de faux négatifs.

$$\text{rappel} = \frac{\text{TP}}{\text{TP+FN}}$$

- **Le score F1 :** le score F1(F_measure en anglais) est une métrique qui combine rappel et précision avec La moyenne harmonique, en tenant compte des deux métriques dans l'équation suivante.

$$F1 = 2 * \frac{Précision * rappel}{Précision + rappel}$$

5 Implémentation

Nous présenterons dans cette section l'implémentation de notre mémoire par les algorithmes d'apprentissage supervisé utilisés dans l'analyse du sentiment des revues de médicaments. Nous adoptons le F_measure comme un facteur de comparaison entre les classifieurs utilisés.

Dans ce travail nous avons utilisé les méthodes python confusion_matrix et classification_report pour accéder à l'ensemble des métriques de performance, et tout cela grâce à la bibliothèque Sklearn

5.1 La division (Train/Test Split)

Avant d'utiliser l'algorithmes d'apprentissage automatique supervisé, nous avons divisé les données de dataset en deux parties ; la première partie d'entraînement pour entraîner le modèle et la deuxième partie pour tester le modèle et évaluer ses performances. Dans la

première étape, nous avons divisé les données comme suit : Utilisez 90% des données pour apprentissage et 10% pour les tests.

Dans cette étape nous utilisons un état aléatoire (en anglais, random-state) pour nous assurer que nous obtenons la même répartition à chaque fois que nous exécutons notre script. Si nous ne définissons pas l'état aléatoire, il sera trié en fonction du temp, ce qui conduit à la des résultats différents dans la plupart de nos exécutions.

5.2 L'évaluation de résultats

5.2.1 Séparateur à Vaste Marge (SVM)

Nous avons utilisé la classifieur Séparateur à Vaste Marge (SVM) pour analyser les sentiments des patients à propos des revues des médicaments et les classer comme positif ou négative.

Les résultats de SVM sur les différents corpus utilisés sont donnés comme suit :

Le Tableau 8 présente la matrice de confusion du modèle généré par le classifier SVM sur l'ensemble de données Cymbalta_drug_dataset.

Tableau 8 : Matrice de confusion de SVM sur Cymbalta_drug_dataset

	Négative	Positive
Négative	83	17
Positive	16	71

Le Tableau 9 présente les performances du modèle généré par le classifier SVM sur l'ensemble de données Cymbalta_drug_dataset.

Tableau 9 : Performances pour SVM sur Cymbalta_drug_dataset

	Précision	Rappel	F_measure
Négative : 0	0.84	0.83	0.83
Positive : 1	0.81	0.82	0.81
Moyenne	0.825	0.825	0.82

Le Tableau 10 présente les résultats de matrice de confusion du modèle généré par le classifier SVM sur l'ensemble de données drugsCom_Reduced1.

Tableau 10 : Matrice de confusion de SVM sur drugsCom_Reduced1

	Négative	Positive
Négative	80	9
Positive	109	469

Les résultats de performances du modèle généré par le classifier SVM sur l'ensemble de données drugsCom_Reduced1, sont présentés dans le Tableau 11 .

Tableau 11 : Performances pour SVM sur drugsCom_Reduced1

	Précision	Rappel	F_measure
Négative : 0	0.42	0.90	0.58
Positive : 1	0.98	0.81	0.89
Moyenne	0.7	0.86	0.74

Dans le Tableau 12 nous avons présenté les résultats de matrice de confusion du modèle généré par le classifier SVM sur l'ensemble de données drugsCom_Reduced2.

Tableau 12 : Matrice de confusion de SVM sur drugsCom_Reduced2

	Négative	Positive
Négative	80	12
Positive	113	465

Dans le Tableau 13 nous avons présenté les résultats de matrice de confusion du modèle généré par le classifier SVM sur l'ensemble de données drugsCom_Reduced3

Tableau 13 : Performances pour SVM sur drugsCom_Reduced2

	Précision	Rappel	F_measure
Négative : 0	0.41	0.87	0.56
Positive : 1	0.97	0.80	0.88
Moyenne	0.69	0.84	0.72

Dans le Tableau 14 nous avons présenté les résultats de matrice de confusion du modèle généré par le classifier SVM sur l'ensemble de données drugsCom_Reduced3.

Tableau 14 : Matrice de confusion de SVM sur drugsCom_Reduced3

	Négative	Positive
Négative	75	17
Positive	110	477

Dans le Tableau 15 nous avons présenté les résultats de matrice de confusion du modèle généré par le classifier SVM sur l'ensemble de données drugsCom_Reduced3.

Tableau 15 : Performances pour SVM sur drugsCom_Reduced3

	Précision	Rappel	F_measure
Négative : 0	0.41	0.82	0.54
Positive : 1	0 .97	0.81	0.88
Moyenne	0.69	0.82	0.71

5.2.2 Classifieur Naïve Bayes (NB)

Nous avons utilisé la classifieur naïve bayes (NB) pour analyser les sentiments des patients à propos des revues des médicaments et les classer comme positif ou négative.

Les résultats de NB sur les différents corpus utilisés sont donnés comme suit :

Le Tableau 16 présente la matrice de confusion du modèle généré par le classifier NB sur l'ensemble de données Cymbalta_drug_dataset

Tableau 16 : Matrice de confusion de NB sur Cymbalta_drug_dataset

	Positive	Négative
Positive	84	20
Négative	15	68

Le Tableau 17 présente la performance du modèle généré par le classifier NB sur l'ensemble de données Cymbalta_drug_dataset

Tableau 17 : Performances pour NB sur Cymbalta_drug_dataset

	Précision	Rappel	F_measure
Négative : 0	0.85	0.81	0.83
Positive : 1	0.77	0 .82	0.80
Moyenne	0.81	0.82	0.82

Dans le Tableau 18 nous avons présenté les résultats de matrice de confusion du modèle généré par le classifier NB sur l'ensemble de données Cymbalta_drug_dataset.

Tableau 18 : Matrice de confusion de NB sur drugsCom_Reduced1

	Positive	Négative
Positive	1	1
Négative	188	477

Dans le Tableau 19 nous avons présenté les résultats de performance du modèle généré par le classifier NB sur l'ensemble de données drugsCom_Reduced1

Tableau 19 : Performances pour NB sur drugsCom_Reduced1

	Précision	Rappel	F_measure
Négative : 0	0.01	0.50	0.01
Positive : 1	1.00	0.72	0 .83
Moyenne	0.51	0.61	0.42

Dans le Tableau 20 nous avons présenté les résultats de matrice de confusion du modèle généré par le classifier NB sur l'ensemble de données drugsCom_Reduced2.

Tableau 20 : Matrice de confusion de NB sur drugsCom_Reduced2

	Positive	Négative
Positive	3	1
Négative	190	476

Dans le Tableau 21 nous avons présenté les résultats de performance du modèle généré par le classifier NB sur l'ensemble de données drugsCom_Reduced2.

Tableau 21 : Performances pour NB sur drugsCom_Reduced2

	Précision	Rappel	F_measure
Négative : 0	0.02	0.75	0.03
Positive : 1	1.00	0.71	0.83
Moyenne	0.51	0.73	0.43

Dans le Tableau 22 nous avons présenté les résultats de matrice de confusion du modèle généré par le classifier NB sur l'ensemble de données drugsCom_Reduced3

Tableau 22 : Matrice de confusion de NB sur drugsCom_Reduced3

	Positive	Négative
Positive	0	0
Négative	185	494

Dans le Tableau 23 nous avons présenté les résultats de performance du modèle généré par le classifier NB sur l'ensemble de données drugsCom_Reduced3.

Tableau 23 : Performances pour NB sur drugsCom_Reduced3

	Précision	Rappel	F_measure
Négative : 0	0.00	0.00	0.00
Positive : 1	1.00	0.73	0.84
Moyenne	0.50	0.37	0.42

5.2.3 Classifieur La Forêt d'arbres décisionnels (RF)

Nous avons utilisé la classifieur Forêt d'arbres décisionnels (RF) pour analyser les sentiments des patients à propos des revues des médicaments et les classer comme positif ou négative.

Les résultats de RF sur les différents corpus utilisés sont donnés comme suit :

Dans le Tableau 24 nous avons présenté les résultats de matrice de confusion du modèle généré par le classifier RF sur l'ensemble de données Cymbalta_drug_dataset.

Tableau 24 : Matrice de confusion de RF sur Cymbalta_drug_dataset

	Négative	Positive
Négative	83	22
Positive	16	66

Dans le Tableau 25 nous avons présenté les résultats de performance du modèle généré par le classifier RF sur l'ensemble de données Cymbalta_drug_dataset.

Tableau 25 : Performances pour RF sur Cymbalta_drug_dataset

	Précision	Rappel	F_measure
Négative : 0	0.84	0.79	0.81
Positive : 1	0.75	0.80	0.78
Moyenne	0.80	0.80	0.80

Dans le Tableau 26 nous avons présenté les résultats de matrice de confusion du modèle généré par le classifier RF sur l'ensemble de données drugsCom_Reduced1

Tableau 26 : Matrice de confusion de RF sur drugsCom_Reduced1

	Négative	Positive
Négative	35	7
Positive	154	471

Dans le Tableau 27 nous avons présenté les résultats de performance du modèle généré par le classifier RF sur l'ensemble de données drugsCom_Reduced1

Tableau 27 : Performances pour RF sur drugsCom_Reduced1

	Précision	Rappel	F_measure
Négative : 0	0.19	0.83	0.30
Positive : 1	0.99	0.75	0.85
Moyenne	0.59	0.79	0.58

Dans le Tableau 28 nous avons présenté les résultats de matrice de confusion du modèle généré par le classifier RF sur l'ensemble de données drugsCom_Reduced2.

Tableau 28 : Matrice de confusion de RF sur drugsCom_Reduced2

	Négative	Positive
Négative	58	1
Positive	135	476

Dans le Tableau 29 nous avons présenté les résultats de performances du modèle généré par le classifier RF sur l'ensemble de données drugsCom_Reduced2.

Tableau 29 : Performances pour RF sur drugsCom_Reduced2

	Précision	Rappel	F_measure
Négative : 0	0.47	0.50	0.48
Positive : 1	0.81	0.79	0.80
Moyenne	0.64	0.65	0.64

Dans le Tableau 30 nous avons présenté les résultats de matrice de confusion du modèle généré par le classifier RF sur l'ensemble de données drugsCom_Reduced3

Tableau 30 : Matrice de confusion de RF sur drugsCom_Reduced3

	Négative	Positive
Négative	34	3
Positive	151	491

Dans le Tableau 31 nous avons présenté les résultats de performance du modèle généré par le classifier RF sur l'ensemble de données drugsCom_Reduced3.

Tableau 31 : Performances pour RF sur drugsCom_Reduced3

	Précision	Rappel	F_measure
Négative : 0	0.18	0.92	0.31
Positive : 1	0.99	0.76	0.86

Moyenne	0.60	0.84	0.59

5.2.3.1 Classifieur Arbre de Décision DT

Nous avons utilisé la classifieur arbre de décision DT pour analyser les sentiments des patients à propos des revues des médicaments et les classer comme positif ou négative.

Les résultats de DT sur les différents corpus utilisés sont donnés comme suit :

Le Tableau 32 montre les résultats de la matrice de confusion du modèle généré par le classifieur DT sur l'ensemble de données Cymbalta_drug_dataset.

Tableau 32 : Matrice de confusion de DT sur Cymbalta_drug_dataset

	Positive	Négative
Négative	64	36
Positive	35	52

Le Tableau 33 montre les résultats de la matrice de confusion du modèle généré par le classifieur DT sur l'ensemble de données Cymbalta_drug_dataset

Tableau 33 : Performances pour DT sur Cymbalta_drug_dataset

	Précision	Rappel	F_measure
Négative : 0	0.65	0.64	0.81
Positive : 1	0.59	0.60	0.78
Moyenne	0.62	0.62	0.80

Le Tableau 34 montre les résultats de performance du modèle généré par le classifieur DT sur l'ensemble de données Cymbalta_drug_dataset

Tableau 34 : Matrice de confusion de DT sur drugsCom_Reduced1

	Positive	Négative
Négative	84	86
Positive	105	392

Tableau 35 : Performances pour DT sur drugsCom_Reduced1

	Précision	Rappel	F_measure
Négative : 0	0.44	0.49	0.47
Positive : 1	0.82	0.79	0.80
Moyenne	0.63	0.64	0.64

Les résultats de la matrice de confusion, du modèle généré par le classifier DT sur l'ensemble de données drugsCom_Reduced2 sont présenté dans le Tableau 36.

Tableau 36 : Matrice de confusion de DT sur drugsCom_Reduced2

	Positive	Négative
Négative	90	91
Positive	103	386

Dans le Tableau 37, nous avons montré les résultats de performance du modèle généré par le classifier DT sur l'ensemble de données drugsCom_Reduced2.

Tableau 37 : Performances pour DT sur drugsCom_Reduced2

	Précision	Rappel	F_measure
Négative : 0	0.47	0.50	0.48
Positive : 1	0.81	0.79	0.80
Moyenne	0.64	0.65	0.64

Nous avons présenté les résultats de matrice de confusion, pour le modèle que généré par le classifier DT sur l'ensembles de données drugsCom_Reduced3 dans le Tableau 38.

Tableau 38 : Matrice de confusion de DT sur drugsCom_Reduced3

	Positive	Négative
Négative	57	107

Positive	128	387

Dans le Tableau 39, nous avons présenté les résultats de la performance du modèle généré par le classifier DT sur l'ensemble de données drugsCom_Reduced3.

Tableau 39 : Performances pour DT sur drugsCom_Reduced3

	Précision	Rappel	F_measure
Négative : 0	0.31	0.35	0.33
Positive : 1	0.78	0.75	0.77
Moyenne	0.55	0.55	0.55

5.2.3.2 Classifieur K Plus Proche Voisins (KNN)

Nous avons utilisé la classifieur K Pus Proche Voisins (KNN) pour analyser les sentiments des patients à propos des revues des médicaments et les classer comme positif ou négative.
Les résultats de KNN sur les différents corpus utilisés sont donnés comme suit :

Les résultats de la matrice de confusion pour le modèle généré par le classifier KNN sur l'ensemble de données sont présentés dans le

Tableau 40.

Tableau 40 : Matrice de confusion de KNN sur Cymbalta_drug_dataset

	Négative	Positive
Négative	91	28
Positive	8	60

Les résultats de performance obtenus pour le modèle généré par le classifier KNN sur l'ensemble de données Cymbalta_drug_dataset sont présentés dans le Tableau 41.

Tableau 41 : Performances pour KNN sur Cymbalta_drug_dataset

	Précision	Rappel	F_measure
Négative : 0	0.92	0.76	0.83

Positive : 1	0.68	0.88	0.77
Moyenne	0.8	0.82	0.8

Le Tableau 42 présente la performance pour modèle généré par le classifier KNN sur l'ensemble de données drugsCom_Reduced1

Tableau 42 : Matrice de confusion de KNN sur drugsCom_Reduced1

	Négative	Positive
Négative	30	23
Positive	159	455

Le Tableau 43 présente la performance pour modèle généré par le classifier KNN sur l'ensemble de données drugsCom_Reduced1

Tableau 43 : Performances pour KNN sur drugsCom_Reduced1

	Précision	Rappel	F_measure
Négative : 0	0.16	0.57	0.25
Positive : 1	0.95	0.74	0.83
Moyenne	0.56	0.66	0.54

Les résultats de la matrice de confusion pour le modèle généré par le classifier KNN sur l'ensemble de données drugsCom_Reduced2 sont présentés dans le Tableau 44

Tableau 44 : Matrice de confusion de KNN sur drugsCom_Reduced2

	Négative	Positive
Négative	65	46
Positive	128	431

Les résultats de performance obtenus pour le modèle généré par le classifier KNN sur l'ensemble de données drugsCom_Reduced2 sont présentés dans le Tableau 45

Tableau 45 : Performances pour KNN sur drugsCom_Reduced2

	Précision	Rappel	F_measure
Négative : 0	0.34	0.59	0.43
Positive : 1	0.90	0.77	0.83
Moyenne	0.62	0.68	0.63

La matrice de confusion du modèle généré par le classifier KNN sur l'ensemble de données drugsCom_Reduced3 est présentée dans le Tableau 46

Tableau 46 : Matrice de confusion de KNN sur drugsCom_Reduced3

	Négative	Positive
Négative	56	37
Positive	129	457

Le Tableau 47 présente la performance pour modèle généré par le classifier KNN sur l'ensemble de données drugsCom_Reduced3.

Tableau 47 : Performances pour KNN sur drugsCom_Reduced3

	Précision	Rappel	F_measure
Négative : 0	0.30	0.60	0.40
Positive : 1	0.93	0.78	0.85
Moyenne	0.62	0.69	0.63

A partir des résultats des cinq classifieurs sur l'ensemble de données Cymbalta_drug_dataset nous avons remarqué que les performances des deux classes sont proches avec une petite amélioration dans la classe négative. Par contre dans les corpus, drugsCom_reduced1, drugsCom_reduced2, et drugsCom_reduced3, la classe positive donne de très performants résultats par rapport à la classe négative.

5.2.4 Comparaison entre les cinq classifieur

Dans cette étude, nous avons utilisé cinq algorithmes SVM, NB, RF, DT, KNN pour analyser les sentiments des patients à propos des revues des médicaments et les classer comme positif

ou négatif. Les performances de tous les classifieurs utilisés ont été comparées selon l'échelle de F_mesure dans le Tableau 48.

Tableau 48: Comparaison des différents ensembles de données en termes de F_measure

	SVM	NB	RF	DT	KNN
Cymbalta_drug_dataset	**0.82**	**0.82**	**0.80**	**0.80**	**0.80**
drugsCom_Reduced1	**0.74**	0.42	0.58	**0.64**	0.54
drugsCom_Reduced2	**0.72**	0.43	**0.64**	**0.64**	0.63
drugsCom_Reduced3	**0.71**	0.42	0.59	0.55	**0.63**

Après avoir obtenu les résultats qui présentés dans le Tableau 48. Nous remarquons que le modèle SVM donne les meilleures performances sur les différents ensembles de données. Pour les autres classifieurs, les résultats diffèrent d'un corpus à un autre.

5.3 L'application développer

Pour faciliter aux utilisateurs finaux d'utiliser de nos modèles générer sur cas réels nous développer une interface graphique utilisant l'environnement Streamlit. La Figure 19 représente l'interface graphique.

Figure 19: Fenêtre principale de l'application

6 Conclusion

Dans ce dernier chapitre, nous présentons le langage de programmation Python ainsi que les différentes bibliothèques utilisées. Nous avons expliqué les détails des ensembles de données des revues de médicaments utilisées dans ce travail de Master. L'annotation de l'orientation sentimentales des quatre corpus est faite, dans ce mémoire, sur la base des évaluations des auteurs de ces commentaires. Ensuite, nous avons appliqué des algorithmes d'apprentissage automatique supervisé ; Les K Plus Proches Voisins (KNN), les Arbres de Décision (DT), les Séparateurs à Vaste Marge (SVM), Les Forêts d'Arbres Décisionnels (RF) et Naïve Bayes (NB) pour analyser des sentiments des patients. Nous avons calculé le F_mesure de classification pour chacun des algorithmes utilisés, afin de choisir le meilleur algorithme de classification. Après la fin de notre étude, nous avons obtenu que le meilleur algorithme de classification est le SVM pour tous les différents corpus.

Aussi, l'ensemble de données Cymbalta_drug_dataset donne les meilleurs résultats par rapport aux autres ensembles de données. Les performances des classes positive et négative sont proches dans Cymbalta_drug_dataset, qui est équilibrée en termes de nombres de documents positifs et négatifs. Dans l'autre partie, dans les trois corpus ; drugsCom_reduced1, drugsCom_reduced2, et drugsCom_reduced3, la classe positive est dominante en nombre de commentaires. Ce déséquilibre influence sur les résultats de classification, où la classe positive largement dépasse la classe négative en performance.

Conclusion générale et perspectives

Conclusion générale et perspectives

L'analyse des sentiments est une technique du Traitement Automatique de Langage Naturelle (TALN). Cette technique classe les sentiments textuels en positifs ou négatifs, et cela se fait à travers des modèles construits spécifiquement à cet effet. L'analyse de ces sentiments dans le domaine médical est devenue un sujet très important, car elle aide les patients, à prendre une décision concernant l'achat d'un produit ou la demande d'un service auprès d'une entreprise particulière. Elle aide également les entreprises à améliorer la qualité de leurs services et produits en définissant les avantages et les inconvénients de ces produits ou services en fonction des avis des malades sur des réseaux sociaux, des forums de discussion, des sites web, etc. Cela aide également les entreprises, à savoir ce que les clients ont besoin de produits ou de services qui n'étaient pas fournis auparavant, ce qui entraîne un profit économique pour ces entreprises.

L'apprentissage automatique est l'un des branches de l'intelligence artificielle qui vise à apprendre aux ordinateurs à penser d'une manière similaire à le pensée humaine. Il s'agit également de lui faire effectuer les tâches requises par lui-même sans savoir à le programmer littéralement. Il existe différents types d'apprentissage automatique ; apprentissage supervisé, apprentissage non supervisé, apprentissage semi supervisé et apprentissage par renforcement. Chaque type a ses propres algorithmes.

Dans l'étude expérimentale nous avons utilisé quatre ensembles de données qui concernent les avis des patients sur les médicaments. Le premier ensemble de données, Cymbalta_drug_dataset est créé dans le cadre de ce travail de mémoire. Pour la création de Cymbalta_drug_dataset nous avons interrogé le site web askapatient.com pour collecter les commentaires sur le médicament Cymbalta prescrit pour traiter la dépression. Nous avons annoté ce corpus sur la base des notes d'évaluation des visiteurs de site de leurs expériences avec le médicament Cymbalta. Nous avons pris soin, dans la collecte des commentaires de notre corpus, d'avoir un maximum d'équilibre entre les classes positive et négative. Les trois autres ensembles de données, drugsCom_reduced1, drugsCom_reduced2, et drugsCom_reduced3, constituent des sous-ensembles d'un grand corpus, à savoir drugsCom. Ces trois corpus sont déséquilibrés, où le nombres de commentaires positifs largement dépasse celui des commentaires négatifs.

Dans l'expérimentation, le classifieur SVM donne les meilleurs résultats dans les quatre corpus. Pour notre corpus, Cymbalta_drug_dataset, le classifieur NB donne une performance similaire au SVM de 0.82. Concernant le corpus drusCom_Reduced1, le classifieur DT vient après SVM avec un résultat de classification de 0.64. Pour le corpus drugsCom_Reduced2, les classifieurs RF et DT vient après SVM avec un résultat similaire de classification de 0.64. Concernant le dernier corpus, drugsCom_reduced3, le classifieur KNN performe le mieux après SVM avec un F_measure 0.63.

La conclusion la plus importante de ce mémoire de Master est l'importance de l'équilibrage des ensembles de données en termes des documents dans les déférentes classes. Cet équilibrage influence directement les résultats de classification. Le corpus créé dans ce travail à savoir Cymbalta_drug_dataset donne les meilleurs résultats, et ces résultats sont équilibrés entre les deux classe, positive et négative. Pour les autres trois ensembles de données les résultats de la classe positive, qui est la plus dominante dans ces corpus, dépassent largement ceux la classe négative.

Perspectives

Pour les perspectives de ce travail, nous proposons pour l'amélioration futur de cette étude les points suivants :

- Augmentation de la taille de l'ensemble de données des sentiments des patients sur les médicaments.
- L'apprentissage avec d'autre classifieurs est spécialement des classifieur d'apprentissage profond (deep learning).
- L'amélioration de l'interface graphique pour être facilement utilisable par les utilisateurs finaux.

Bibliographie

Bibliographie

1. Habes, Yasmine. *Application des méthodes d'Apprentissage Automatique dans l'Analyse des Sentiments des Tweets Arabes.* Département d'informatique. Blida : Université Saad Dahlab blida 1, 21 janvier 2021. Mémoire de Master.

2. WELL, STAFF. The Importance of Sentiment Analysis In Healthcare. *WELL.* [En ligne] 23 mars 2021. https://wellapp.com/blog/sentiment-analysis-in_healthcare/.

3. HCAHPS: Patients' Perspectives of Care Survey. *CMS.gov.* [Online] U.S. Centers for Medicare & Medicaid Services. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instrumentes/HospitalQualityInits/HospitalHCHPS.

4. Hadji, Mehdi. Analyse de sentiments : Générali. *Medium.* [En ligne] 22 8 2019. https//link.medium.com/159W75N5jqb.

5. Mba Engonga, Ulrick Serge. *Approches d'analyse des données issues des médias sociaux appliquées au domaine de la santé.* canada : Mémoire de master à l'UNIVERSITÉ DE SHERBROOKE, 10 2015.

6. Gohil, Sunir, Vuik, Sabine and Darzi, Ara . Sentiment Analysis of Health Care Tweets: Review of the Methods Used. *JMIR Public Health Surveill.* 4 23, 2018.

7. مي جوة. *في الفن.* [متصل] 26 12, 2022. [تاريخ الاقتباس: 27 5, 2022.]
.https://www.google.com/amp/s/www.filfan.com/news/detailsamp/127659

8. Qu'est-ce que l'intelligence artificielle—IA ? *ORACLE.* [En ligne] https://www.oracle.com/dz/artificial-intelligence/what-is-ai/.

9. Qu'est-ce que l'intelligence artificielle? *NetApp.* [En ligne] 2022. http://www.netapp.com/fr/artificial-intelligence/what_is-artificial-intelligence/.

10. *HI4TECK.* [En ligne] 25 1 2020. https://www.hi4teck.com/2022/01/artificial-intelligence.html?m=1.

11. Apprentissage automatique . *WIKIPEDIA.* [En ligne] 1 5 2022. [Citation : 2022 3 28.] https://fr.m.wikipedia.org/wiki/Apprentissage_automatique#.

12. Le machine learning et la science des données. *IBM.* [En ligne] https://www.ibm.com/fr-fr/analytics/machine-learning#.

13. Les 3 étapes essentielles de l'apprentissage automatique (Machine Learning). *SPIRIA.* [En ligne] 2022. https://www.spiria.com/fr/blogue/intelligence-artificielle/3_etapes-essentielles-apprentissage-automatique-machine-learning/.

14. Apprentissage supervisé. *WIKIPEDIA.* [En ligne] 17 5 2022. [Citation : 28 5 2022.] https://fr.wikipedia.org/wiki/Apprentissage_supervis%C3%A9.

15. ISMAILI, Zakariyaa. *ANALYTICS & INSIGHTS.* [En ligne] 28 janvier 2022. [Citation : 27 5 2022.] https://analyticsinsights.io/apprentissage-supervise-vs-non-supervise/.

16. Gaétan, Raoul. Machine Learning : les 9 types d'algorithme les plus prtinents en enterprise. *LEMAGIT.* [En ligne] 08 juin 2020. [Citation : 27 5 2022.] https://www.lemagit.fr/conseil/Machine-Learning-les-9-typs-dalgorithmes-les-plus-pertinents-en-enterprise.

17. Support Vector Machine Algorithm. *javatpoint.* [En ligne] 2021-2022. [Citation : 4 27 2022.] https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

18. Hausmane , Issarane. *ANALYTICS & INSIGHTS.* [En ligne] 2022. [Citation : 28 5 2022.] https://www.google.com/amp/s/analyticsinsights.io/les-svm-support-vector-machine/amp/.

19. Mbaabu, Onesmus. Introduction to Random Forest in Machine Learning. *Section.* [En ligne] 11 December 2020. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/.

20. Naïve Bayes Classifier Algorithm. *javatpoint.* [En ligne] 2011-2021. https://www.javatpoint.com/machine-learning-naive-bayes-classifier.

21. K-Nearest Neighbor(KNN) Algorithm for Machine Learning. *javatpoint.* [En ligne] [Citation : 29 5 2022.] https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.

22. Apprentissage automatique. *le parisien sensagent.* [En ligne] [Citation : 28 5 2022.] https://dictionnaire.sensagent.leparisien.fr/Apprentissage%20automatique/fr-fr/.

23. avogadro. Apprentissage supervisé et non supervisé. *itrain.fr.* [En ligne] 9 5 2020. [Citation : 9 5 2022.] https://www.google.com/amp/s/www.itrain.fr/amp/apprentissage-supervis%25C3%25A9-et-non-supervis%25C3%25A9.

24. Allmang, Amandine. Principaux algorithmes d'apprentissage non supervisé. *Linedata.* [En ligne] 2022. https://fr.linedata.com/principaux-algorithmes-dapprentissage-non-supervise.

25. Apprentissage semi-supervisé. *WIKIPEDIA.* [En ligne] 20 décembre 2021. https://fr.wikipedia.org/wiki/Apprentissage_semi-supervis%C3%A9.

26. colin. Apprentissage par renforcement - Reinforcement Learning. *Colin Bouvry-[DokuWik].* [En ligne] 17 2 2019. [Citation : 28 5 2022.] https://colinbouvry.com/dokuwiki/doku.php?id=reinforcement_learning.

27. Comrendre Python avec les forces et les faiblesses de Python, le save-vous? *ALTITUDETVM.* [En ligne] [Citation : 29 5 2022.] https://altitudetvm.com/fr/komputer/1232_pengertian-python-beserta-kelebihan-dan-kekurangan-python-sudah_tahu.html.

28. Daniel , Johanson. NLTK Tutorial: What is NLTK Library in Python. *Guru99.* [En ligne] 14 5 2022. https://www.guru99.com/nltk-tutorial.html.

29. Hyméros. *WIKIPEDIA.* [En ligne] 1 12 2021. https://fr.m.wikipedia.org/wiki/Pandas.

30. ابراهيم البحيصي. مكتبات علم البيانات بالبايثون اشهر 5 بايثونات. *بايثونات.* [متصل] 7 4, 2022. [تاريخ الاقتباس: 29 5, 2022.]
https://pythonat.com/articles/%d9%85%d9%83%d8%aa%d8%a8%d8%a7%d8%aa-%d8%b9%d9%84%d9%85-%d8%a7%d9%84%d8%a8%d9%8a%d8%a7%d9%86%d8%a7%d8%aa-%d8%a8%d8%a7%d9%84%d8%a8%d8%a7%d9%8a%d8%ab%d9%88%d9%86-5-./%d9%85%d9%83%d8%aa%d8%a8%d8%a7%d8%aa

31. Scikit_Learn : guide démarrage rapide en Machine Learning avec Python. *Data Transition Numérique* . [En ligne] [Citation : 29 5 2022.] https:// www.data-transitionnumerique.com/scikit-learn-python/.

32. Gbadebo , Bello. Introduction à Jupyter Notebook pour les débutants. *GEEKFLARE.* [En ligne] 12 11 2019. https://geekflare.com/fr/jupyter-notebook-basics/.

33. Cymbalta. *Drugs.com.* [En ligne] 22 5 2022. https://www.drugs.com/search.php?searchterm=Cymbalta&a=1.

34. Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. *In Proceedings of the 2018 International Conference on Digital Health (DH '18).* New York : ACM, 2018.

35. Tokenising into Words and Sentences | What is Tokenization and it’s Definition? *greatlearning.* [En ligne] 29 5 2020. https://www.mygreatlearning.com/blog/tokenization/.

36. *An Interpretation of Lemmatization and Stemming in Natural Language Processing.* Divya Khyani, Siddhartha B S, Niveditha N M,Divya B M. 10, Shanghai : University of Shanghai, 2020, Journal of University of Shanghai for Science and Technology, Vol. 22.

37. Algorithme TF - IDF (principe + implémentation de code Python). [En ligne] 31 12 2021. https://pythonmana.com/2021/12/202112310216439737.html.