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

MEMOIRE

Présenté en vue de l'obtention du diplôme

Master en informatique

Spécialité : Technologies de l'information et de la communication

THEME

Analyse des sentiments et modélisation thématique concernant les maladies chroniques

Présenté par :

ARBOUCHE Massine Issiakhem

ADJOUT Lotfi

Soutenu publiquement le : jj/mm/aaaa

Devant le jury composé de:

Président : ……..

Examinateur : ……

Encadreur : ……

2022/2023

Dédicace

“

À travers ces mots, je tiens à exprimer ma profonde gratitude envers mes estimés grands-parents, mes chers parents dévoués et ma sœur bien-aimée, Alicia. Votre dévouement sans faille à mon égard a été la source de ma persévérance et de ma détermination. C'est grâce à votre présence que j'ai pu parvenir à atteindre ce jalon de ma vie. Vos encouragements, vos sacrifices et votre amour ont forgé l'individu que je suis aujourd'hui. Je dédie cette thèse à vous tous, avec une reconnaissance éternelle pour votre influence bénéfique dans ma vie. Puissions-nous continuer à partager des moments exquis en votre compagnie et à édifier un futur empli de bonheur et de triomphes.

”

Massine

Dédicace

“

Je dédie ce travail à À mes chers parents, votre amour et soutien infaillibles m'ont guidé. Vos sacrifices et votre présence constante ont été inestimables. À mes chers frères et sœurs, vos encouragements constants ont été une source d'inspiration. Votre soutien inconditionnel et votre présence à mes côtés ont été essentiels. À ma famille et à mes amis, votre soutien indéfectible a été d'une valeur inestimable. Votre présence et vos encouragements tout au long de mon parcours ont été précieux. Grâce à vous, je suis arrivé jusqu'ici. Votre soutien et votre implication ont été des moteurs essentiels dans ma réussite.

”

Lotfi

Remerciement

Nous souhaitons tout d'abord exprimer notre profonde gratitude envers notre encadrante Dr. LAIFA Meriem, pour son précieux soutien et son engagement infaillible tout au long de notre mémoire. Nous souhaitons également souligner les efforts remarquables qu'elle a consacrés à l'encadrement de notre projet de fin d'études. Sa disponibilité, sa bienveillance et ses conseils éclairés ont été d'une importance inestimable. Nous lui sommes particulièrement reconnaissants de son souci du détail et de son perfectionnisme exemplaire dans l'exécution de son travail.

Nous souhaitons également exprimer nos sincères remerciements aux membres du jury pour l'honneur qu'ils nous accordent en prenant le temps de lire et d'évaluer ce travail.

Nous tenons également à exprimer notre chaleureuse gratitude envers l'équipe pédagogique et administrative du département d'informatique pour leurs efforts constants en vue de nous offrir une formation de qualité.

Enfin, nous souhaitons exprimer notre profonde gratitude envers toutes les personnes qui ont contribué, de manière directe ou indirecte, à la réalisation de ce travail.

Résumé

L'objectif principal de ce projet consiste à collecter des informations précieuses sur les sentiments associées aux maladies chroniques afin d'améliorer la compréhension et les soins de santé dans ce domaine spécifique. Pour atteindre cet objectif, deux méthodes d'analyse textuelle ont été utilisées : l'analyse des sentiments et la modélisation thématique.

En ce qui concerne l'analyse des sentiments, une analyse approfondie a été réalisée en implémentant plusieurs algorithmes tels que K-Nearest Neighbors (KNN), Support Vector Machines (SVM) et l'algorithme d'arbre de décision en utilisant deux techniques d'extraction de caractéristiques, à savoir Term Frequency-Inverse Document Frequency (TF-IDF) et Bag of Words (BoW), qui ont été combinées respectivement avec des algorithmes de régression logistique et linéaire.

De plus, pour analyser efficacement le langage naturel en arabe, trois modèles pré-entraînés adaptés, ArabBert, DziriBert et Cross-lingual Language Model (XLM), ont été utilisés grâce à l'algorithme de transfert d'apprentissage. Les résultats obtenus mettent en évidence la supériorité des modèles de transfert par apprentissage par rapport aux méthodes classiques, en termes de performance.

En ce qui concerne la modélisation thématique, les algorithmes Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) et (Hierarchical Dirichlet Process) HDP ont été utilisés, afin d'identifier et d'analyser les thèmes clés présents dans les données.

Mots clés : Analyse des sentiments, modélisation thématique, transfert par apprentissage, extraction des caractéristiques, Maladies chroniques.

Abstract

The main objective of this project is to collect valuable information about the sentiments associated with chronic diseases in order to understand and enhance healthcare in this specific field. To achieve this goal, two methods of text analysis were employed: sentiment analysis and topic modeling.

For sentiment analysis, a comprehensive analysis was conducted by implementing several algorithms such as KNN, SVM, and the decision tree algorithm, using two feature extraction techniques, namely TF-IDF and Bag of Words. These techniques were respectively combined with logistic regression and linear regression algorithms.

Moreover, to effectively analyze Arabic natural language, three adapted pre-trained models, ArabBert, DziriBert, and XLM, were employed through transfer learning algorithms. The obtained results highlight the superiority of transfer learning models over conventional methods in terms of performance.

As for topic modeling, the LSA, LDA, and HDP algorithms were utilized to identify and analyze key themes present in the data.

Keywords: Sentiment analysis, Topic modeling, Transfer learning, Feature extraction, Chronic diseases.

ملخص

الهدف الأساسي من هذا المشروع هو جمع معلومات قيمة حول المشاعر المرتبطة بالأمراض المزمنة لتحسين الفهم ورعاية الصحة في هذا المجال المحدد. من أجل تحقيق هذا الهدف، تم استخدام طريقتين لتحليل النص: تحليل المشاعر ونمذجة المواضيع.

فيما يتعلق بتحليل المشاعر، تم إجراء تحليل مفصل من خلال تنفيذ عدة خوارزميات مثل KNN وSVM وخوارزمية الشجرة القرارية باستخدام تقنيتي استخراج الميزات، وهما TF-IDF وBag of Words. تم دمج هذه التقنيات على التوالي مع خوارزميات الانحدار اللوجستي والانحدار الخطي.

بالإضافة إلى ذلك، من أجل معالجة اللغة الطبيعية بشكل فعال في اللغة العربية، تم استخدام ثلاثة نماذج مدربة مسبقًا متكيفة وهي ArabBert وDziriBert وXLM من خلال خوارزمية نقل التعلم. تظهر النتائج المحصل عليها تفوق نماذج نقل التعلم على الأساليب التقليدية من حيث الأداء.

أما بالنسبة لنمذجة المواضيع، تم استخدام خوارزميات LSA وLDA وHDP لتحديد وتحليل المواضيع الرئيسية الموجودة في البيانات.

الكلمات المفتاحية: تحليل المشاعر، نمذجة المواضيع، التعلم النقلي، استخراج السمات، الأمراض المزمنة.

Table des matières

Liste des abréviations

IA Intelligence Artificielle

LDA Latent Dirichlet Allocation

PLSA Probabilistic Latent Semantic Analysis

NNMF Non-Negative Matrix Factorization

LSA Latent Semantic Analysis

ASL Analyse Sémantique Latente

BoW Bag of Words

TF-IDF Term Frequency - Inverse Document Frequency

DVS Décomposition en Valeurs Singulières

API Application Programming Interface

HDP Hierarchical Dirichlet Process

KNN K-Nearest Neighbors

SVM Support Vector Machines

XLM Cross-lingual Language Model

TL Transfer Learning

Liste des figures

Liste des tableaux

Introduction Générale

1.1. E-santé et Analyse de données

Le domaine de la santé est en constante évolution, et les avancées technologiques récentes permettent de collecter, stocker et d'analyser des données de plus en plus importantes. Toutefois, pour pouvoir tirer pleinement parti de ces données, il est nécessaire d'avoir à disposition des outils et des méthodes performantes afin de les exploiter efficacement.

C'est là qu'entre en jeu la e-Santé, Santé 2.0 et l'analyse des données, tous jouant un rôle important dans l'amélioration de la qualité des soins et des résultats en matière de santé. En effet la e-Santé est un domaine émergent qui associe l'informatique médicale, la santé publique et les affaires. Il regroupe l'ensemble des services et des informations de santé fournis ou améliorés par Internet et les technologies associées. La Santé 2.0, quant à elle, utilise les technologies Web 2.0, les nouvelles technologies mobiles et cloud pour améliorer les soins de santé.

Enfin, l'analyse de données est le processus, qui consiste à examiner et à interpréter des données afin d'élaborer des réponses à des questions, qui est dans notre cas de figure, la compréhension des tendances en matière de soins de santé. Ensemble, ces technologies ont le potentiel de révolutionner les soins de santé et d'améliorer les résultats pour les patients.

1.2. Importance de l'IA dans l'analyse des données relatives à la santé

Au cours des dernières décennies, l'intelligence artificielle (IA) est devenue omniprésente dans de nombreux aspects de notre vie quotidienne. Cette technologie a émergé comme une force révolutionnaire dans de nombreux secteurs, offrant une efficacité accrue et des avantages économiques grâce à sa capacité à automatiser des tâches complexes autrefois exclusivement effectuées par les humains.

Grâce à ces avancées, l'analyse des données est également profondément influencée par l'émergence de l'IA, ce qui en fait un outil prometteur pour l'analyse des données liées à la santé. Cette utilisation offre de nouvelles perspectives et opportunités, permettant la découverte de modèles et de relations au sein de vastes ensembles de données complexes en rapport avec la santé.

De plus, les professionnels de la santé peuvent recueillir des informations précieuses pour améliorer la qualité des soins dispensés aux patients atteints de maladies chroniques, en tenant compte des attitudes du public à leur égard. Cela nous conduit à affirmer que l'exploitation des données à travers l'IA contribue à une amélioration globale de la qualité des soins de santé en ligne, les rendant ainsi plus accessibles à un large public.

1.3. Objectif et contribution

Notre attention sera principalement portée sur l'intégration de l'IA dans le domaine médical, en mettant l'accent sur l'analyse des sentiments et la modélisation thématique. Ces approches s'avèrent extrêmement utiles dans notre étude, car elles nous permettent d'examiner les opinions et les attitudes des patients ainsi que du grand public à l'égard des maladies chroniques, tout en identifiant les sujets et les thèmes les plus fréquemment discutés sur les réseaux sociaux.

Plus spécifiquement, notre projet vise à exploiter les données et les commentaires relatifs aux maladies chroniques afin d'acquérir une compréhension approfondie des opinions, des émotions et des thèmes dominants parmi les personnes affectées par ces affections.

1.4. Structure du rapport

Ce rapport de mémoire offre une présentation minutieuse de la méthodologie ayant été employée pour la réalisation de ce projet. Cette présentation se décline en quatre chapitres distincts, qui se présentent comme suit :

- Le **premier** chapitre est consacré à l'exploration des maladies chroniques, ainsi qu'à l'analyse de données textuelles en examinant l'analyse des sentiments et la modélisation thématique, tout en passant en revue les défis qui y sont associés.
- Le **deuxième** chapitre se penche sur les approches et techniques d'analyse des sentiments et de modélisation thématique, en mettant en évidence les avantages et les inconvénients de chacune d'entre elles.
- Le **troisième** chapitre décrit la méthodologie du projet en détaillant les différentes étapes à suivre.

- Le **quatrième** chapitre présente la mise en œuvre de l'analyse de sentiment et de la modélisation thématique sur notre corpus de données, en décrivant l'environnement matériel et logiciel utilisé.
- Le rapport se clôture par une **conclusion générale** dans laquelle il est effectué un récapitulatif exhaustif de l'ensemble du projet, en mettant en évidence ses limites y compris les améliorations envisagées pour l'avenir.

Chapitre 01 : Concepts de base

2.1. Introduction

Dans ce chapitre, nous allons procéder à la présentation et à la définition des maladies chroniques, qui font l'objet du domaine d'étude de notre thèse, bien évidemment nous aborderons les aspects les plus importants de notre sujet en fournissant les informations nécessaires.

Nous allons également explorer deux domaines cruciaux de l'analyse de données textuelles : l'analyse des sentiments et la modélisation thématique. Nous aborderons les concepts fondamentaux de chaque domaine, leur champ d'application et leur pertinence dans divers contextes, tout en examinant les défis qu'ils posent.

2.2. Maladies chroniques

Les maladies chroniques sont des affections de longue durée qui nécessitent une attention médicale régulière. Ces maladies peuvent affecter la capacité des patients à accomplir leurs activités quotidiennes, entraînant une détérioration de leur qualité de vie [1].

Cependant, le monde d'aujourd'hui est confronté à une augmentation significative des maladies chroniques. Cette tendance peut être attribuée à divers facteurs de risques liés au développement des maladies chroniques, qui ont profondément influencé notre mode de vie, à l'image des facteurs environnementaux tels que la pollution et les pratiques agricoles, qui ont un impact direct sur notre régime alimentaire.

Si nous devions citer quelques exemples de maladies chroniques courantes qui sont devenues de plus en plus prévalentes à travers le monde et en Algérie bien particulièrement ces derniers temps, on pourrait mentionner, le cancer, les maladies respiratoires chroniques, le diabète, ainsi que les maladies cardiovasculaires telles que les maladies coronariennes et les accidents vasculaires cérébraux, ainsi que les troubles de santé mentale tels que la dépression, l'anxiété et les troubles bipolaires, qui sont également une cause majeure d'invalidité dans le monde.

Comme on pourrait le constater sur la **Figure 1**, les maladies cardiovasculaires ont le taux de mortalité le plus élevé au monde, par rapport à d'autres maladies chroniques, avec un taux de 48%. Les cancers arrivent en deuxième position et chaque type peut être différencié selon son taux de mortalité. En comparaison, les maladies respiratoires et le diabète représentent respectivement un risque de mortalité de 12% et 3%, ce qui est beaucoup moins élevé que pour les autres maladies chroniques. Cela est dû en grande partie aux traitements médicaux disponibles qui permettent de contrôler leur évolution au fil du temps et de s'accommoder avec [2].

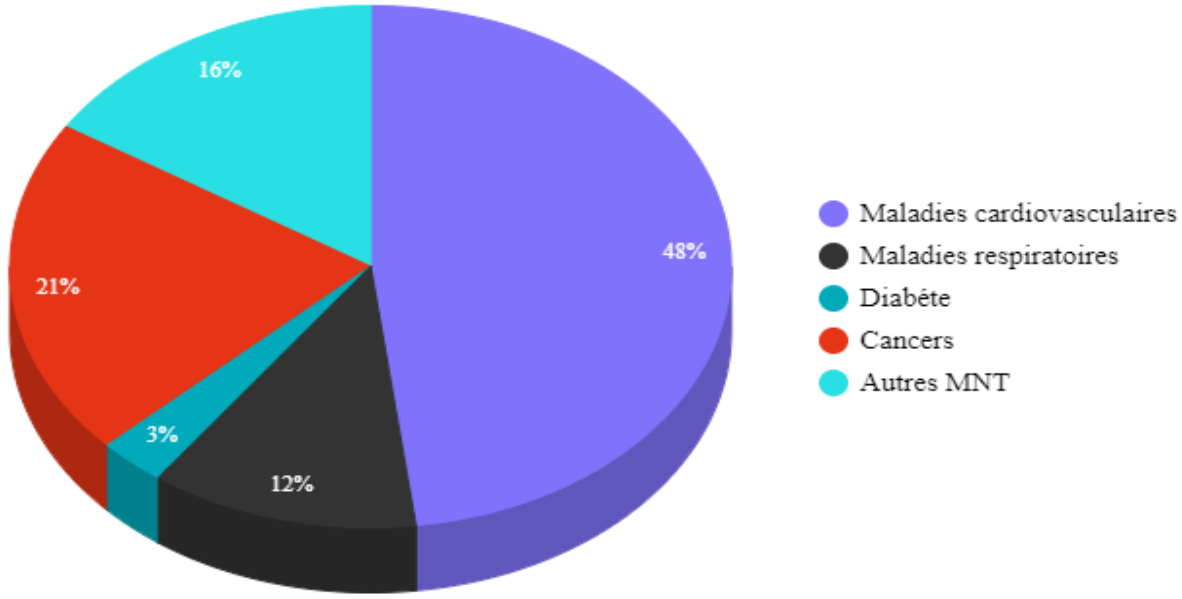

Figure 1. Taux de mortalité pour chaque maladie chronique à travers le monde [2]

2.2.1. Définitions

Nous avons listé dans cette section un ensemble de maladies chroniques, en fournissant leurs définitions les plus courantes telles qu'elles apparaissent dans notre base de données.

- **L'arthrose :** L'arthrose est une maladie dégénérative des articulations dans laquelle le cartilage lisse qui recouvre les surfaces osseuses aux articulations est soit blessé soit usé au fil du temps. Cette usure des articulations peut causer de la douleur, des gonflements et des déformations [3].

- **Asthme :** L'asthme est une maladie des voies respiratoires qui se caractérise par des épisodes récurrents de difficultés respiratoires appelés crises ou dyspnées [4].
- **Cancer :** Le cancer est une maladie qui se manifeste par la croissance anormale de cellules qui se divisent de manière incontrôlable, envahissent les tissus sains et peuvent les détruire [5].
- **Diabète :** Le diabète fait référence à un ensemble de troubles métaboliques qui se caractérisent par la présence d'hyperglycémie non traitée. Ces troubles peuvent être causés par des défauts dans la sécrétion d'insuline, l'action de l'insuline, ou les deux, ainsi que par des perturbations du métabolisme des glucides, des graisses et des protéines [6].
- **Psoriasis :** Le psoriasis est une maladie de la peau papulo squameuse chronique courante qui se caractérise par une inflammation due à une dysfonction du système immunitaire qui provoque une inflammation dans le corps [7].
- **Helicobacter pylori :** Helicobacter pylori est une bactérie fréquente qui peut infecter la muqueuse de l'estomac. Elle est généralement contractée pendant l'enfance et peut entraîner une inflammation de l'estomac, qui peut rester asymptomatique [8].
- **La maladie de Crohn** : la maladie de Crohn est une maladie inflammatoire chronique de l'intestin qui peut résulter d'une interaction complexe entre la susceptibilité génétique, les facteurs environnementaux et la modification du microbiote intestinale, conduisant à des réponses immunitaires innées et adaptatives dys régulées [9].

2.3. L'analyse des sentiments

L'analyse de sentiment est une tâche de traitement du langage naturel ("Natural Language Processing", en anglais), qui consiste à extraire et analyser les opinions, les sentiments, les attitudes et les perceptions des personnes à l'égard de différentes entités telles que des sujets, des produits et des services [10].

La méthode d'analyse de sentiment est principalement composée de trois niveaux d'analyse, comme illustré dans la **figure 2**, qui sont les suivants [10] :

- **Niveau du document :** à ce niveau, le sentiment est extrait à partir de la globalité du document ou du texte, ainsi l'objectif est de le classer comme étant positif, négatif ou neutre.
- **Niveau de la phrase :** s'appuie essentiellement sur le sentiment de chaque phrase, contenue dans le texte, représentée comme une entité unique.
- **Niveau des aspects :** l'analyse est plus profonde que les niveaux précédents, ceci-dit le processus consiste à diviser une phrase, en un ensemble d'entités où, chaque entité renvoie un sentiment.

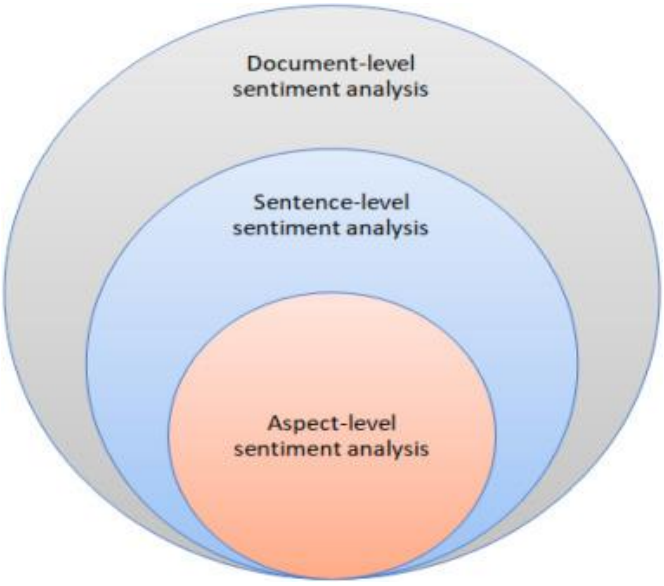

Figure 2. Les 3 principaux niveaux de l'analyse des sentiments [10]

2.3.1. Domaines d'applications

L'analyse de sentiment est largement applicable et indispensable dans de nombreux domaines différents, tels que [10] :

L'informatique décisionnelle : L'analyse des sentiments est largement utilisée dans le domaine de l'informatique décisionnelle en raison de ses nombreux avantages. Par exemple, les entreprises peuvent exploiter les résultats de l'analyse des sentiments pour améliorer leurs produits, étudier les réactions des clients ou élaborer une nouvelle stratégie de marketing.

L'intelligence gouvernementale : En plus des produits et services, les gens émettent également des avis sur divers sujets, tels que la politique, la religion et les problèmes sociaux.

Cependant, L'utilisation de l'analyse des sentiments pour identifier les points de vue sur les décisions politiques ou d'autres problèmes similaires est très utile pour surveiller la réaction éventuelle du public à l'implémentation de certaines mesures.

Les systèmes de recommandation : Un système de recommandation est un algorithme qui propose des contenus pertinents (films, musique ou produits à acheter) aux utilisateurs. Il peut générer un grand nombre de revenus pour certaines industries, ainsi il est possible d'améliorer ces systèmes en appliquant l'analyse des sentiments afin de faire des recommandations plus pertinentes.

Le domaine médical : Les acteurs de la santé s'intéressent de plus en plus à l'utilisation de l'analyse des sentiments dans le domaine médical. Cela leur permet d'obtenir des données sur une variété de sujets, comme les maladies, les effets secondaires des médicaments, les épidémies et les humeurs des patients, et d'utiliser ces informations pour améliorer leurs services de santé.

2.3.2. Défis de l'analyse des sentiments

Bien que l'analyse des sentiments soit une technique utile pour comprendre les opinions et les attitudes des individus, elle doit faire face à plusieurs défis qui peuvent rendre sa mise en œuvre difficile et moins fiable. Nous allons passer en revue les principaux défis auxquels l'analyse des sentiments est confrontée à l'image de [10] :

La détection du sarcasme : Selon le dictionnaire anglais Macmillan, le sarcasme se réfère à l'action de dire ou d'écrire le contraire de ce que l'on veut dire ou de parler d'une manière visant à ridiculiser ou exprimer sa colère. Cependant, cela pose un problème pour l'analyse des sentiments lorsque quelqu'un exprime quelque chose de positif tout en voulant dire le contraire, ou vice versa, ce qui complique la tâche de l'analyse des sentiments. Les expressions sarcastiques sont couramment utilisées dans notre vie quotidienne, de sorte que la détection du sarcasme suscite un intérêt croissant pour surmonter les faux sentiments en identifiant automatiquement les expressions sarcastiques dans un texte donné.

Le traitement de la négation : Le traitement des mots de négation tels que "ne, pas, ni," etc. est très important pour l'analyse des sentiments car ils peuvent inverser le sens de la phrase et la

polarité d'un texte donné. Par exemple, la phrase "Ce film est bon.", est classée comme une phrase positive, alors que "Ce film n'est pas bon", devrait être classée comme une phrase négative. Malheureusement, dans certaines approches, les mots de négation sont supprimés parce qu'ils sont inclus dans des listes de mots d'arrêt ou ignorés implicitement parce qu'ils ont une valeur de sentiment neutre dans un lexique qui n'a pas d'impact sur la polarité finale. Cependant, il n'est pas facile de traiter cette tâche en inversant la polarité, car les mots de négation peuvent être trouvés dans une phrase sans influencer le sentiment du texte.

La détection des spams : La détection des spams joue un rôle très important dans l'analyse des sentiments. Les avis en ligne ayant un impact sur les décisions d'achat des consommateurs, les spams et les faux avis peuvent nuire à la réputation des marques et fausser artificiellement la perception des utilisateurs sur les produits, services, entreprises ou autres entités. Le développement d'un système de détection des spams, capable de repérer les faux avis dans un vaste ensemble d'opinions, est une tâche très complexe car il n'y a pas de différence apparente entre les avis.

La désambiguïsation du sens des mots : Un mot peut avoir différentes significations et en fonction du contexte et du domaine utilisé, le sens de ce mot peut être différent pour chaque situation. La désambiguïsation du sens des mots vise à déterminer quel sens d'un mot est utilisé dans une phrase. Par exemple, le mot "courbe" peut avoir une connotation positive s'il est utilisé en relation avec une télévision, mais peut prendre un sens négatif s'il est associé à un téléphone mobile. Par conséquent, l'identification du sens d'un mot au sein d'une phrase représente un véritable défi.

Les langues à faible niveau de ressources : La plupart des recherches en analyse des sentiments ont été axées sur la langue anglaise et d'autres langues qui disposent d'une quantité satisfaisante de ressources linguistiques. Les méthodes d'apprentissage supervisé sont les plus couramment utilisées pour cette tâche, mais ces approches nécessitent des ressources linguistiques qui peuvent être difficiles à obtenir pour les langues moins populaires. Pour contourner ce problème, diverses méthodes peuvent être employées, telles que la construction de ressources linguistiques à partir de zéro, l'utilisation de méthodes non supervisées, semi-supervisées et

d'apprentissage par transfert, mais n'empêche qu'il persiste parmi les obstacles majeurs à surmonter pour l'analyse des sentiments.

2.4. La modélisation thématique

2.4.1. Définition

La modélisation thématique est une méthode qui fait appel à une série d'algorithmes pour révéler et annoter la structure thématique d'une collection de documents. Elle repose sur le modèle de l'espace vectoriel, qui joue un rôle essentiel dans diverses techniques avancées de recherche d'informations et de modélisation de thèmes [11].

2.4.2. Domaines d'applications

Les progrès technologiques récents ont considérablement élargi les possibilités d'utilisation de la modélisation thématique, en améliorant sa précision et ses capacités dans divers domaines nécessitant une analyse de grandes quantités de données et la recherche d'informations pertinentes tout en prenant en compte la structure des données. Les domaines d'application de cette méthode sont très variés et incluent notamment [11] :

La bio-informatique : La modélisation thématique se révèle être une méthode pertinente pour l'analyse de données en bioinformatique, surpassant les méthodes traditionnelles comme la classification et le regroupement. Cette approche permet une meilleure interprétation des informations biologiques, tout en offrant de nouvelles perspectives pour l'analyse de l'expression génique et la compréhension de l'impact des médicaments sur les voies cellulaires. Ces avancées ouvrent la voie à la mise en place de thérapies ciblées plus efficaces.

L'analyse des réseaux sociaux : Il existe diverses approches de modélisation thématique pour l'analyse des données provenant des réseaux sociaux. Parmi celles-ci, on retrouve l'utilisation de modèles de sujets pour analyser les relations entre les utilisateurs, l'analyse des sentiments des utilisateurs, l'analyse des sujets abordés dans les communiqués de presse politiques, ainsi que l'intégration des attributs tels que les hashtags et le temps dans les modèles de sujets. Ces méthodes

se révèlent particulièrement utiles pour l'analyse des réseaux sociaux dans les situations où seules les données de liaison sont disponibles.

Le génie logiciel : Le développement de l'industrie du logiciel a produit une grande quantité de données non structurées provenant de diverses sources telles que le code source, la documentation, les cas de test et les référentiels de bogues. L'utilisation de la modélisation thématique sur ces données permet de découvrir des informations utiles pour soutenir diverses tâches d'ingénierie logicielle telles que la compréhension du programme et la récupération de liens de traçabilité.

La recherche scientifique : La modélisation thématique est une technique qui peut être appliquée en recherche scientifique pour diverses applications. Elle permet de détecter les tendances et orientations de la recherche, de repérer les domaines émergents, de mettre en évidence les collaborations et les réseaux de chercheurs, et d'évaluer l'efficacité des politiques et programmes de financement de la recherche.

2.4.3. Défis de la modélisation thématique

La modélisation thématique est une technique qui présente plusieurs défis. Il est important de tenir compte de ces obstacles lors de l'utilisation de cette méthode. Parmi les défis rencontrés, on cite [11] :

La visualisation : La visualisation peut être considérée comme un obstacle pour la modélisation thématique car elle ne fournit qu'une compréhension superficielle des sujets traités dans les documents. Bien que les visualisations, telles que les nuages de mots et les outils de réduction de dimension, puissent aider à donner un aperçu global des sujets dans un modèle thématique, ils ne permettent pas une analyse approfondie de chaque sujet individuel et de sa relation avec les autres sujets. Par conséquent, les visualisations peuvent être trompeuses si elles ne sont pas correctement interprétées, conduisant à des conclusions erronées sur les sujets abordés dans les documents.

Interprétation de la modélisation thématique : Comprendre les modèles thématiques peut être difficile en raison de leur nature abstraite et complexe, ce qui rend leur interprétation peu

intuitive. Même si la vraisemblance d'échantillonnage est la mesure la plus couramment utilisée pour évaluer la qualité des modèles de sujets, cette approche a été critiquée pour son manque de facilité d'interprétation des modèles probabilistes. En dépit de cela, l'évaluation de la qualité des modèles reste un enjeu majeur pour garantir leur interprétabilité.

Modélisation de thèmes efficaces en termes de mémoire (Memory Efficient Topic Modeling) : Les modèles thématiques sont souvent confrontés à une complexité de calcul qui restreint leur usage dans les applications à grande échelle et en temps réel. Bien qu'il y ait eu des efforts pour améliorer l'estimation du temps de formation de ces modèles, peu d'attention a été portée au problème crucial de l'inférence de la distribution de sujets, étant donné les modèles préexistants.

Stabilité : L'analyse de stabilité en modélisation thématique a montré que les résultats obtenus avec les algorithmes traditionnels sont instables, que le modèle appliqué aux mêmes documents d'entrée ou mis à jour avec de nouveaux documents. Peu d'efforts ont été consacrés à la résolution de ce problème, qui est essentiel pour produire un modèle thématique définitif, stable et précis.

2.5. Conclusion

En conséquence, bien que les maladies chroniques continuent de représenter un enjeu majeur pour la santé publique à l'échelle mondiale, l'analyse des sentiments et la modélisation thématique se sont avérées être des outils précieux pour la recherche scientifique. Toutefois, il reste des défis à relever pour parvenir à une modélisation précise.

Chapitre 02 : Techniques de l’analyse des sentiments et de la modélisation thématique

3.1. Introduction

Avec la prolifération massive de l'information textuelle sur le web, il est devenu crucial de développer des méthodes et des techniques pour analyser cette diversité de contenu. Cependant, dans ce chapitre, notre étude portera sur les approches de l'analyse des sentiments et de la modélisation thématique, en évaluant les avantages et les inconvénients de chaque méthode. Nous examinerons également les méthodes non supervisées et semi-supervisées utilisées dans notre projet de recherche.

3.2. Les approches et techniques de l’analyse des sentiments

Il est possible de distinguer plusieurs approches, qui concernent l’analyse des sentiments, toutefois nous nous limiterons à mentionner les plus connues, chacune ayant ses propres techniques d'implémentation. Voici quelques exemples [10] :

3.2.1. L’approche d’Apprentissage Automatique (Machine Learning Approach)

Les méthodes d'apprentissage automatique sont capables de créer des modèles adaptés à des domaines spécifiques à partir de textes, ce qui peut améliorer considérablement les performances de classification. Pour y parvenir, trois techniques principales peuvent être utilisées : **l'apprentissage supervisé**, **l'apprentissage non supervisé** et **l'apprentissage semi-supervisé.**

Les avantages de l’approche d'Apprentissage Automatique

- L'approche semi-supervisée est utile pour les ensembles de données non étiquetées qui incluent quelques exemples étiquetés.
- L'approche supervisée est efficace lorsque la tâche de classification a un ensemble spécifique de classes.

- L'approche non supervisée peut être utilisée lorsque la détermination de l'ensemble de classes est difficile en raison de l'absence de données étiquetées.

Les inconvénients de l'approche d'Apprentissage Automatique

- Ces approches nécessitent souvent de grands ensembles de données d'entraînement pour obtenir de bonnes performances de classification.
- Un classificateur formé sur un ensemble de données spécifique ne fonctionne pas aussi bien pour un autre domaine.

3.2.2. L'approche basée sur le lexique (Lexicon Based Approach)

L'approche basée sur le lexique (également appelée approche basée sur la connaissance) est une méthode couramment utilisée en analyse de sentiment et nécessite une ressource lexicale appelée lexique d'opinion (une liste prédéfinie de mots) qui associe des mots à leur orientation sémantique en tant que mots négatifs ou positifs à l'aide de scores.

Les avantages de l'approche basée sur le lexique

- L'approche basée sur le lexique peut être utilisée pour analyser le sentiment dans des domaines où les données d'entraînement sont limitées ou indisponibles.
- L'approche basée sur le lexique est pratique pour l'analyse de sentiment au niveau de la phrase et des caractéristiques.

Les inconvénients de l'approche basée sur le lexique

- Cette approche est dépendante du domaine, car les mots peuvent avoir plusieurs significations et sens. Par conséquent, un mot positif dans un domaine spécifique peut ne pas l'être dans un autre domaine.
- Si un grand ensemble de données d'entraînement est fourni, la performance de cette approche est inférieure à celle de l'approche d'apprentissage automatique.

3.2.3. L’approche hybride (Hybrid Approach)

L'approche hybride combine à la fois des approches lexicales et d'apprentissage automatique, dans le but d'hériter d'une grande précision de l'apprentissage automatique et de la stabilité de l'approche basée sur le lexique.

Les avantages de l’approche hybride

- Combine la rapidité de l'analyse lexicale avec la flexibilité de l'apprentissage automatique pour mieux gérer l'ambiguïté et intégrer le contexte des mots de sentiment.
- L’approche peut améliorer la précision, la stabilité et l'efficacité des modèles de classification.

Les inconvénients de l’approche hybride

- La performance de l'approche hybride dépend de la qualité des lexiques et des algorithmes utilisés.
- La construction et l'utilisation des lexiques peuvent être complexes et coûteuses en temps et en ressources.
- Peu de modèles utilisent l'approche hybride pour l'analyse de sentiment.

3.2.4. L’approche basée sur les aspects (Aspect-based approach)

L'analyse des sentiments basée sur les aspects est une tâche d'analyse fine, qui vise à prédire les polarités de sentiment de certains aspects donnés ou des termes cibles dans les textes. Les aspects peuvent être des attributs, des caractéristiques ou des traits de la cible.

Les avantages de l'approche basée sur les aspects

- L’approche basée sur les aspects consiste en deux étapes distinctes, ce qui permet de traiter chaque aspect individuellement pour une analyse plus précise.
- Des approches non supervisées peuvent être utilisées pour extraire les aspects, ce qui réduit la nécessité de disposer de grandes quantités de données d'entraînement.

- L'analyse de sentiment basée sur les aspects permet une analyse fine des sentiments pour des termes spécifiques tels que des produits ou services.

Les inconvénients de l'approche basée sur les aspects

- L'analyse de sentiment basée sur les aspects peut être complexe et nécessiter des algorithmes sophistiqués pour extraire les aspects pertinents.
- Il peut être difficile de couvrir tous les aspects d'un produit ou service, ce qui peut limiter la précision de l'analyse.

3.2.5. L'approche de l'apprentissage par transfert (Transfer Learning approach)

L'apprentissage par transfert est une technique d'apprentissage automatique qui utilise la similarité des données, la distribution des données et la tâche du modèle pour appliquer les connaissances déjà acquises dans un domaine à un autre.

Les avantages de l'approche de l'apprentissage par transfert

- Cette approche permet de gagner du temps en évitant de devoir entraîner un algorithme de zéro pour chaque domaine.
- Pour une tâche d'analyse de sentiment, cette approche est souvent appliquée pour transférer la capacité acquise de classifier les sentiments d'un domaine à un autre.

Les inconvénients de l'approche de l'apprentissage par transfert

- Limitations de la généralisation des modèles pré-entraînés à des tâches ou des domaines très différents de ceux sur lesquels ils ont été formés.
- La méthode ne fonctionne bien que si les données sources sont similaires aux données cibles.

La **figure 3** ci-dessous récapitule les différentes approches de l'analyse des sentiments abordées.

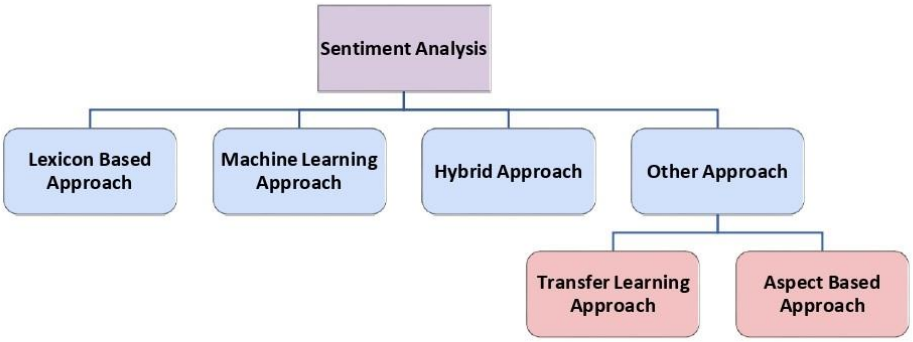

Figure 3. Les différentes approches de l'analyse des sentiments

3.3. Les approches et techniques de la modélisation thématique

La modélisation thématique peut être abordée à travers deux approches principales : les méthodes probabilistes et les méthodes non probabilistes, comme illustré dans la **figure 4**.

Les méthodes probabilistes, comme PLSA (Probabilistic Latent Semantic Analysis) et LDA (Latent Dirichlet Allocation), se basent sur des modèles de probabilité pour identifier les thèmes. Ces méthodes sont principalement utilisées pour des modèles non supervisés, mais peuvent également être appliquées à des configurations supervisées ou semi-supervisées. D'un autre côté, les approches non probabilistes, telles que LSA (Latent Semantic Analysis), NNMF (Non-Negative Matrix Factorization), utilisent des méthodes algébriques pour extraire les thèmes des données textuelles, et sont souvent utilisées pour l'analyse de grands corpus de données textuelles, car elles sont généralement plus rapides et moins complexes que les approches probabilistes [11].

Figure 4. Classification hiérarchique des approches de la modélisation thématique [11]

3.3.1. LDA : Latent Dirichlet Allocation

La Latent Dirichlet Allocation (LDA) est une méthode basée sur le théorème de définition qui permet de capturer la structure statistique significative entre les documents et à l'intérieur de ceux-ci via une distribution mixte. L'hypothèse sous-jacente est que les documents sont composés de plusieurs sujets, qui sont définis comme une distribution de mots.

Un corpus est associé à un nombre prédéfini de sujets k, et chaque document dans le corpus contient ces sujets avec une proportion différente. Le but du modèle de thématisation est d'apprendre ces sujets à partir des données ou du corpus.

Dans un ensemble de données, la distribution de k sujets doit être apprise par inférence statistique. L'algorithme définit un processus génératif comme une distribution conjointe de probabilité sur les variables observées et cachées [11].

Le processus d'apprentissage de la distribution des sujets est décrit à travers une notation en plaque donnée dans la **figure 5.**

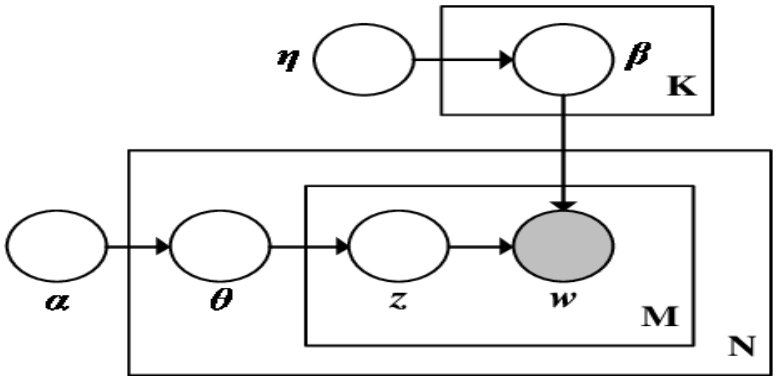

Figure 5. La notation en plaque de Latent Dirichlet Allocation (LDA) [11]

Où pour chaque document :

- (a) : Distribution sur les sujets θd ~ Dir(α), où Dir(.) est un tirage d'une distribution uniforme de Dirichlet avec un paramètre d'échelle α.
- (b) : Pour chaque mot dans le document : 1**)** Tirer un sujet spécifique zd,n~ multi(θd)
  2) Tirer un mot wd,n~βz,d,n.

3.3.2. PLSA : Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (PLSA) est une méthode de réduction de dimension en exploration de texte basée sur le sac de mots (“Bag of Words”, en anglais), qui a été développée par Hoffman pour détecter la co-occurrence sémantique des termes à l'aide d'un cadre probabiliste dans un corpus. Le premier modèle statistique pour révéler la co-occurrence sémantique dans la matrice document-terme d'un corpus était le modèle Aspect. Il est basé sur le concept que chaque mot est généré à partir d'un seul sujet et que différents mots dans un document peuvent être générés à partir de différents sujets.

Comme indiqué dans la **figure 6**, Chaque document est représenté sous la forme d'une liste de proportions de mélange pour ces composants de mélange et est ainsi réduit à une distribution de probabilité sur un ensemble fixe de sujets [11].

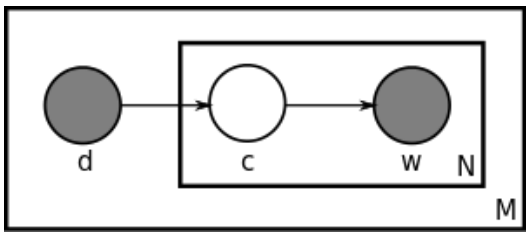

Figure 6. Processus de fonctionnement du modèle PLSA [12]

- d : La variable d'indexation de document.
- c : Désigne le sujet du mot qui est choisi à partir de la distribution de sujets du document.
- w : Est tiré de la distribution de mots du sujet auquel il appartient.

3.3.3. NNMF : Non-Negative Matrix Factorization

Cette technique innovante de réduction de dimensionnalité permet de résoudre le problème de la présence de nombres négatifs dans les données en imposant des contraintes de non-négativité sur le modèle de données.

Elle repose sur le concept de NNMF, qui a été utilisé pour la première fois par Pattero en 1994 pour des données environnementales. L'objectif de NNMF est de remédier aux composantes négatives dans les modèles de données, car celles-ci peuvent contredire la réalité physique dans de nombreuses applications [11].

La **figure 7** présente une illustration de la structure du modèle.

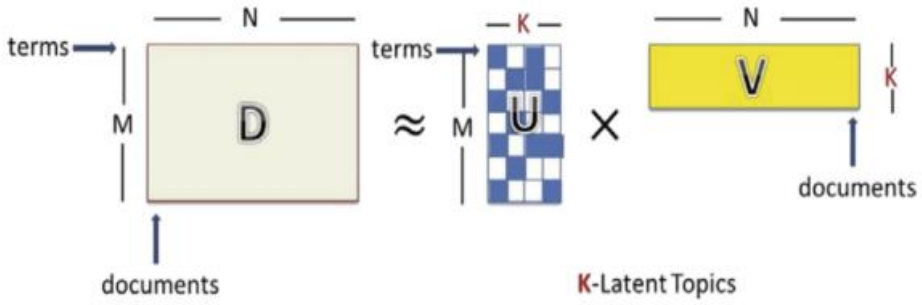

Figure 7. Structure du modèle NNMF [13]

Où :

- D : Le résultat de la factorisation des deux matrices U et V.
- U : Correspond à K axes de coordonnées.
- V : Correspond à N points dans un nouvel espace sémantique.

3.3.4. LSA : Latent Semantic Analysis

La méthode de l'Analyse Sémantique Latente (ASL) repose sur l'algèbre de la décomposition en valeurs singulières (DVS) et permet de représenter l'espace sémantique des documents en rapprochant les relations sémantiques et les usages contextuels. La fondation théorique de LSA est basée sur des hypothèses distributionnelles, qui affirment que les termes ayant une signification similaire apparaissent également très proches dans leur utilisation contextuelle. Toutes les relations sémantiques entre les textes sont directement inférées à partir du corpus de texte donné. En utilisant une représentation vectorielle du texte, LSA calcule la similarité entre les textes et trouve des mots similaires dans le texte pour organiser sémantiquement le texte en clusters sémantiques.

Les domaines d'application de LSA sont nombreux, notamment la notation automatique des essais, l'enseignement intelligent, la recherche d'information, l'analyse de réseaux sociaux et la synthèse de texte [11].

3.3.5. Avantages et inconvénients de la modélisation thématique

Dans le **tableau 1** ci-dessous, vous trouverez une synthèse détaillée des différents avantages et inconvénients des approches de la modélisation thématique [11].

Tableau 1. Les différents avantages et inconvénients des approches de la modélisation thématique

Méthode	**Avantages**	**Inconvénients**
LSA	➔ Permet aux termes de sens similaires d'être très proches dans leur utilisation contextuelle. ➔ Permet de projeter l'espace de grande dimension en basse dimension avec DVS. ➔ Montre la corrélation des termes dans l'espace sémantique avec différents rangs.	➔ Pourrait avoir besoin d'un grand corpus de textes pour être efficace. ➔ Pourrait être coûteuse en termes de temps de traitement et de stockage de données. ➔ Basée sur une méthode algébrique et peut ne pas prendre en compte tous les aspects du langage humain.
NNMF	➔ Permet une représentation de données limitée à des vecteurs non négatifs. ➔ Prend en charge les contraintes de parcimonie dans de nombreuses applications. ➔ Pourrait être utilisée dans de nombreuses applications telles que l'identification de cancers à partir de données d'expression génétique moléculaire, la reconnaissance de motifs, le traitement d'image.	➔ Limitée à des vecteurs non négatifs, ce qui peut ne pas convenir à certaines applications. ➔ Nécessite des données de grande dimensionnalité, ce qui peut augmenter la complexité et les temps de calcul. ➔ Pourrait être sensible à la qualité et à la représentativité des données utilisées, ce qui peut affecter la qualité des résultats obtenus.
PLSA	➔ Modélise chaque document comme une liste de proportions de mélanges pour les différents sujets dans le corpus. ➔ Permet de représenter chaque document par une distribution de probabilité sur un ensemble fixe de sujets. ➔ Ne repose pas sur des hypothèses et offre une plus grande flexibilité dans la découverte des modèles dans les données textuelles, Contrairement à d'autres techniques.	➔ Suppose que les termes sont indépendants les uns des autres, ce qui peut ne pas être vrai dans certains contextes. ➔ Nécessite un grand corpus de texte pour fonctionner efficacement, ce qui peut être difficile ou coûteux à obtenir. ➔ Est basée sur le modèle "bag of words" (BoW), qui peut ne pas capturer toutes les nuances et les relations de sens entre les mots dans un texte.
LDA	➔ L'approche est flexible car elle suppose que les documents proviennent de plusieurs sujets. ➔ LDA est basé sur un modèle de	➔ Il est nécessaire de définir manuellement le nombre de sujets à générer, ce qui peut être difficile et subjectif.

	variables cachées, utilisé depuis des décennies en apprentissage automatique.	➔ Pourrait avoir des difficultés à gérer des sujets complexes qui sont difficiles à résumer en un petit nombre de mots-clés.

3.3.6. Les méthodes non supervisées et semi-supervisées

La classification de sentiments est un domaine important de l'apprentissage automatique, mais qui peut être difficile en raison du manque de données étiquetées ou de classes spécifiques. Pour surmonter ces défis, les méthodes non supervisées et semi-supervisées offrent une alternative en utilisant des données non étiquetées pour découvrir des structures et des motifs. Ces techniques sont particulièrement utiles dans les cas où les données sont difficiles ou coûteuses à étiqueter, ou lorsque les schémas et les relations entre les données sont inconnus [10].

3.3.6.1. Les méthodes non supervisées

Les méthodes non supervisées en apprentissage automatique permettent de classifier des données sans données étiquetées. Elles utilisent les propriétés statistiques des documents pour regrouper les données en différents groupes ou clusters. Dans l'analyse de sentiment, ces méthodes reposent principalement sur des techniques de clustering qui permettent de classifier les données en fonction de leur similarité, sans toutefois préciser exactement quel sentiment est représenté par chaque groupe. Cependant, on se contentera de mentionner les deux méthodes de clustering les plus réputées, qui peuvent être utilisées pour cette tâche, qui sont comme suit [10] :

Les méthodes hiérarchiques (hierarchical methods) : Les méthodes hiérarchiques sont des techniques d'analyse de données qui créent une décomposition hiérarchique d'un ensemble de données, représentée par des clusters imbriqués (groupes qui ont des sous-groupes) organisés sous forme d'arbre.

Les méthodes de partitionnement (clustering methods) : Les méthodes de partitionnement visent à diviser les données en un ensemble de clusters non chevauchants où chaque élément est assigné à un seul cluster. Cette partition est basée sur un critère de similarité qui est généralement la distance euclidienne entre les éléments. Les données à l'intérieur d'un

cluster ont une distance très courte les unes des autres tandis qu'elles ont la plus grande distance par rapport aux données des autres clusters.

3.6.2. Les méthodes semi-supervisées

Les méthodes d'apprentissage semi-supervisé sont mises en œuvre lorsque l'obtention de données étiquetées est difficile. Elles diffèrent des approches non supervisées car elles utilisent un petit ensemble de données d'entraînement initiales étiquetées pour guider la procédure d'apprentissage des caractéristiques.

De ce fait, l'apprentissage semi-supervisé se situe entre les approches supervisées et non supervisées, et englobe plusieurs méthodes, telles que [10] :

L'approche générative (Generative Approach) : Cette approche suppose que les données de différentes catégories suivent des distributions différentes et que les paramètres de chaque distribution peuvent être estimés s'il y a au moins une donnée étiquetée par catégorie.

En d'autres termes, un modèle génératif définit des distributions sur les entrées et utilise la règle de Bayes pour prédire l'étiquette (classe) d'une entrée de test après avoir entraîné ce modèle pour chaque classe.

L'approche de co-entraînement (Co-training approach) : Cet algorithme suppose que les données peuvent être représentées à l'aide de deux vues indépendantes, chaque vue contenant des informations sur chaque donnée.

En co-entraînement, deux classificateurs distincts sont formés pour s'enseigner mutuellement en se basant sur les informations partagées entre eux pendant le processus d'apprentissage. Chaque classificateur est entraîné sur un ensemble de caractéristiques différent correspondant aux deux vues des données.

Le processus d'entraînement est itératif et à chaque itération, le co-entraînement met à jour le jeu de données en ajoutant les instances les plus classifiées avec confiance de chaque classificateur aux données étiquetées. Le processus s'arrête lorsque toutes les données non étiquetées ont été utilisées ou qu'un nombre spécifique d'itérations a été atteint.

L’approche d’Auto Apprentissage (Self-training approach) : Cette approche se divise en deux étapes distinctes. La première étape consiste à entraîner un classificateur à partir d'un ensemble réduit de données étiquetées. Dans la seconde étape, le classificateur ainsi entraîné est utilisé pour classer des données non étiquetées, permettant ainsi d'ajouter les exemples les plus fiables à l'ensemble initial de données d'entraînement, en tant que nouvelles données étiquetées.

Cette dernière étape est ensuite répétée de manière itérative, avec l'inclusion de nouvelles données étiquetées. Le modèle final obtenu est ensuite évalué à l'aide d'un ensemble de données de test

L'approche basée sur les graphes (Graph-based approach) : Dans cette approche, une architecture de graphe est utilisée pour représenter les données. Les sommets illustrent les instances (par exemple, des phrases) dans le graphe tandis que les arêtes décrivent la similarité entre les instances. Les instances fortement connectées ont tendance à appartenir à la même classe. En raison de l'utilisation répandue de cette approche par de nombreuses études, son efficacité a été prouvée dans de nombreuses tâches de NLP, notamment l'analyse des sentiments.

L'approche d'apprentissage multi-vues (Multi-view learning) : Cette approche prend en considération plusieurs points de vue pour traiter le problème et les performances globales sont obtenues en utilisant l'accord entre eux. Chaque classificateur sera entraîné sur une seule vue, puis ces classificateurs seront utilisés pour étiqueter les échantillons non étiquetés qui seront ajoutés à l'ensemble d'entraînement s'ils sont classés avec une grande fiabilité. Cette technique est généralement appliquée aux problèmes avec plusieurs ensembles de caractéristiques différents.

La **figure 8** fournit un aperçu des différentes méthodes non supervisées et semi-supervisées.

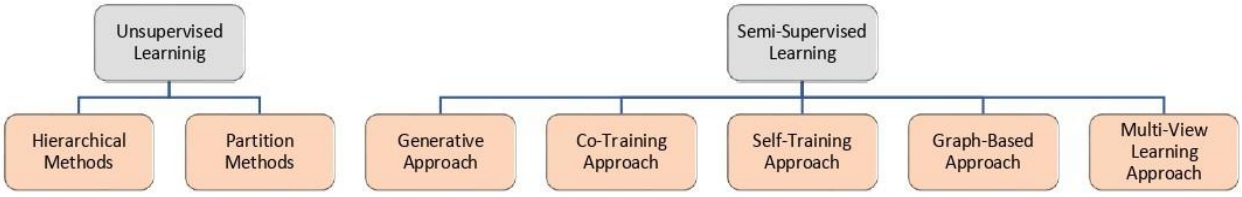

Figure 8. Diverses méthodes non supervisées et semi-supervisées

3.4. Conclusion

En conclusion de ce chapitre, nous avons présenté un aperçu des approches et techniques couramment utilisées dans deux domaines importants de la fouille de texte : l'analyse des sentiments et la modélisation thématique.

Nous avons ainsi exploré les méthodes non supervisées et semi-supervisées, qui sont souvent utilisées dans la fouille de texte pour traiter des ensembles de données étiquetés ou non étiquetés, et cela dans le but d'améliorer considérablement la performance et la compréhension des données textuelles complexes.

En dernier lieu, une bonne compréhension des approches et des techniques de la fouille de texte est essentielle pour sélectionner la méthode la plus appropriée en fonction des données et des objectifs de l'analyse.

Chapitre 03 : Méthodologie

4.1. Introduction

Ce chapitre propose une approche complète pour mener une analyse de sentiments sur les maladies chroniques à partir de données collectées sur les réseaux sociaux. Le but est d'acquérir une connaissance approfondie et précise des opinions et des sentiments exprimés sur les réseaux sociaux à propos des maladies chroniques.

4.2. Description des étapes du projet

Afin d’effectuer une analyse de sentiments approfondie, il est nécessaire de suivre un processus bien défini, tel qu'illustré dans la **figure 9**. Tout d'abord, les données sont importées et soumises à un prétraitement au moyen d'un ensemble d'étapes visant à faciliter leur exploitation. Ensuite, une série de méthodes de modélisation thématique et d'analyse de sentiments est appliquée. Cette approche permet une exploration approfondie de la structure des données, en extrayant les caractéristiques pertinentes des avis et des ressentis exprimés sur les réseaux sociaux concernant les maladies chroniques.

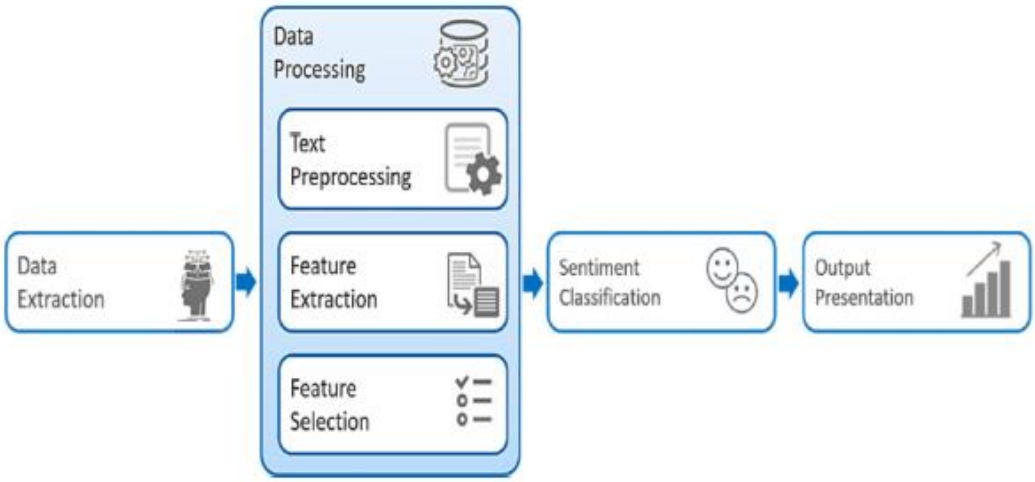

Figure 9. Le processus générique d'analyse des sentiments [10]

4.2.1. Collecte et extraction de données

Pour conduire une analyse de sentiments efficace, il est primordial de se doter de sources de données adéquates et représentatives. Dans cette optique, notre étude se focalise sur les commentaires en ligne portant sur les maladies chroniques, collectés à partir de la plateforme Facebook, qui est sans conteste l'une des plus vastes plateformes de réseaux sociaux à travers le monde.

Notre démarche consiste à extraire les commentaires provenant d'un ensemble de pages et de groupes traitant des maladies chroniques, étant donné que les commentaires sur Facebook fournissent une mine d'informations précieuses car ils permettent de rassembler les avis, les vécus et les ressentis des individus directement concernés par les maladies chroniques, ainsi que de leur entourage.

Dans le cadre de notre thèse, Nous avons décidé d'utiliser un algorithme d'extraction de commentaires à partir d'une clé API Facebook, tel qu'indiqué dans la **figure 10**. Cette méthode de collecte de données implique l'utilisation de l'API de recherche, qui repose sur des requêtes de recherche pour obtenir des textes contenant des mots-clés spécifiques, ou l'API de flux, qui permet de capturer des données textuelles en temps réel filtrées selon des critères tels que les mots-clés et la localisation géographique [10].

Algorithme : Algorithme d'extraction de commentaires à partir d'une clé API Facebook

```
DEBUT
IMPORTER la bibliothèque Facebook API
INITIALISER la clé API = votre_clé_API
CREER une instance de l'API Facebook
RECUPERER les commentaires d'un post
POUR chaque commentaire dans commentaires FAIRE
    AFFICHER les commentaires
FIN POUR
FIN
```

Figure 10. Algorithme d'extraction de commentaires à partir d'une clé API Facebook

4.2.2. Défis liés à l'extraction des données et leurs qualités

Comme évoqué précédemment, les données ont été extraites exclusivement à partir de la plateforme Facebook. Cette décision a été motivée par les défis d'extraction de données qui peuvent se poser en raison du manque de données disponibles sur d'autres réseaux sociaux.

L'extraction de données relatives aux maladies chroniques à partir d'autres plateformes de réseaux sociaux que Facebook est rendue difficile en raison de plusieurs facteurs. Les Algériens ont tendance à préférer Facebook pour exprimer leurs opinions sur ces maladies, ce qui réduit la quantité de données disponibles sur les autres plateformes. De plus, le manque de participation des professionnels de la santé et des associations de patients contribue également à la rareté des données. La qualité des données est un autre défi majeur, car les commentaires peuvent comporter des erreurs d'écriture, des fautes d'orthographe, être mal formulés ou imprécis, ce qui complique leur extraction et leur analyse.

4.2.3. Nettoyage de données

Le nettoyage de données (ou "cleaning data" en anglais) revêt une importance cruciale lors du traitement de données et diverses méthodes et pratiques sont employées pour y parvenir. Ci-dessous, nous présentons quelques-unes des méthodes couramment utilisées pour le nettoyage de données :

Identification et suppression des doublons : Les données collectées à partir de différentes sources ou saisies de manière erronée peuvent générer des doublons, ce qui peut impacter la qualité des données. Toutefois, ces doublons peuvent être identifiés en comparant les enregistrements de données et en éliminant les duplications afin de garantir l'intégrité des données.

Correction des erreurs de saisie : Les erreurs de saisie de données peuvent avoir différentes origines, qu'il s'agisse d'erreurs humaines ou de problèmes liés au logiciel utilisé. Pour corriger ces erreurs, il peut être nécessaire d'effectuer une vérification manuelle des données ou d'utiliser des algorithmes de détection d'erreurs pour identifier et corriger les problèmes.

Imputation des données manquantes : Les valeurs manquantes dans les données peuvent être dues à des erreurs de saisie ou à des informations non collectées. Pour combler ces données manquantes, différentes techniques peuvent être utilisées, telles que l'imputation de données manquantes à l'aide de méthodes statistiques visant à remplacer les valeurs manquantes.

Normalisation des données : Afin d'assurer la cohérence et la comparabilité des données, une étape de normalisation peut être nécessaire. Cette opération peut inclure la conversion des données en une échelle commune ou l'application de techniques de normalisation pour éliminer tout biais présent dans les données.

Élimination des valeurs aberrantes : Les valeurs aberrantes peuvent altérer les résultats des analyses et des modélisations de données. Pour y remédier, il peut être nécessaire de détecter et d'éliminer ces valeurs. Cette opération peut être effectuée en utilisant des seuils pour identifier les valeurs aberrantes ou en appliquant des techniques statistiques pour détecter les valeurs extrêmes.

4.2.4. Prétraitement de données

Le prétraitement de données représente une étape fondamentale dans la chaîne de traitement des données avant leur utilisation dans une analyse ou un modèle. Cette phase englobe plusieurs opérations, y compris l'acquisition, la transformation et la sélection des données pertinentes. Elle permet de corriger les erreurs et les anomalies, de standardiser les données et de les préparer pour une utilisation optimale dans les analyses et les modèles.

Dans ce qui suit, nous détaillons les fonctions de prétraitement qui ont été appliquées à nos textes.

La traduction en arabe : L'objectif est de permettre une meilleure compréhension et une analyse précise des sentiments exprimés dans les textes arabe ou en dialecte algerien.

Suppression des mentions : Les utilisateurs mentionnés ont été retirés du contenu des commentaires car ils ne fournissent aucune information supplémentaire sur les sentiments des utilisateurs.

Suppression des emojis et des chiffres : Cette étude étant axée sur le texte, tous les emojis et chiffres ont été supprimés.

Suppression des caractères spéciaux : Des caractères tels que ~, %, *, !, +, ", {] ont également été supprimés.

Suppression des diacritiques : Les diacritiques arabes comprennent des indicateurs de consonnes et des marques de voyelles connues sous le nom de Tashkil. Tous les diacritiques ont été supprimés. Par exemple, "إِنَّ السَّعِيدَ لَمَن جُنِّبَ الفِتَن" (signifiant : Heureux est celui qui évite la tentation) devient "إن السعيد لمن جنب الفتن."

Suppression de la ponctuation : Les marques de ponctuation arabes (par exemple "،", "؟" ,"؛") et anglaises (par exemple ";", ",", "?") ont été utilisées dans les tweets extraits. Elles ont toutes été supprimées.

Suppression des caractères répétés : Parfois, les lettres sont répétées pour un effet d'exagération ou d'accentuation. Toutes les lettres arabes répétées ont été supprimées. Par exemple, "الداااااااار" qui signifie une maison, devient "الدار".

Suppression des mots vides : Les mots fréquemment utilisés qui ne transmettent pas de sens fort en arabe (par exemple "اما", "فاذا", "في", "متى", "لماذا") et en dialecte algerien (par exemple " راني", "هذوك", "وعلاش", "تاع", "ماشي", "لانو", "راكم", "هذو"") ont été supprimés.

Normalisation : Différentes formes de lettres arabes ont été transformées en une seule forme. Par exemple, la première lettre de l'alphabet arabe "أ" a différentes formes telles que "ا" ,"آ", "إ". Pour normaliser le texte, toutes les lettres Alef ont été transformées en "ا".

Tokenisation : Chaque commentaire a été divisé en un ensemble de jetons (c'est-à-dire des mots séparés par des caractères non alphabétiques).

4.2.5. Les annotations

Pour assurer la qualité de nos données et éviter toute forme de biais liée à l'annotation automatique, nous avons privilégié une approche d'annotation manuelle pour les données textuelles collectées dans le cadre de notre projet. En effet, l'annotation manuelle nous a permis d'obtenir des données plus précises et plus détaillées, reflétant ainsi de manière plus exacte la complexité des opinions et des sentiments exprimés dans les textes. Grâce à cette étape d'annotation manuelle, nous avons considérablement amélioré la qualité de notre base de données et renforcé la fiabilité de nos modèles d'analyse de sentiment.

De ce fait, il est possible de constater dans la **figure 11** ci-dessous que notre base de données comporte 1919 commentaires neutres, 788 commentaires positifs et 143 commentaires négatifs.

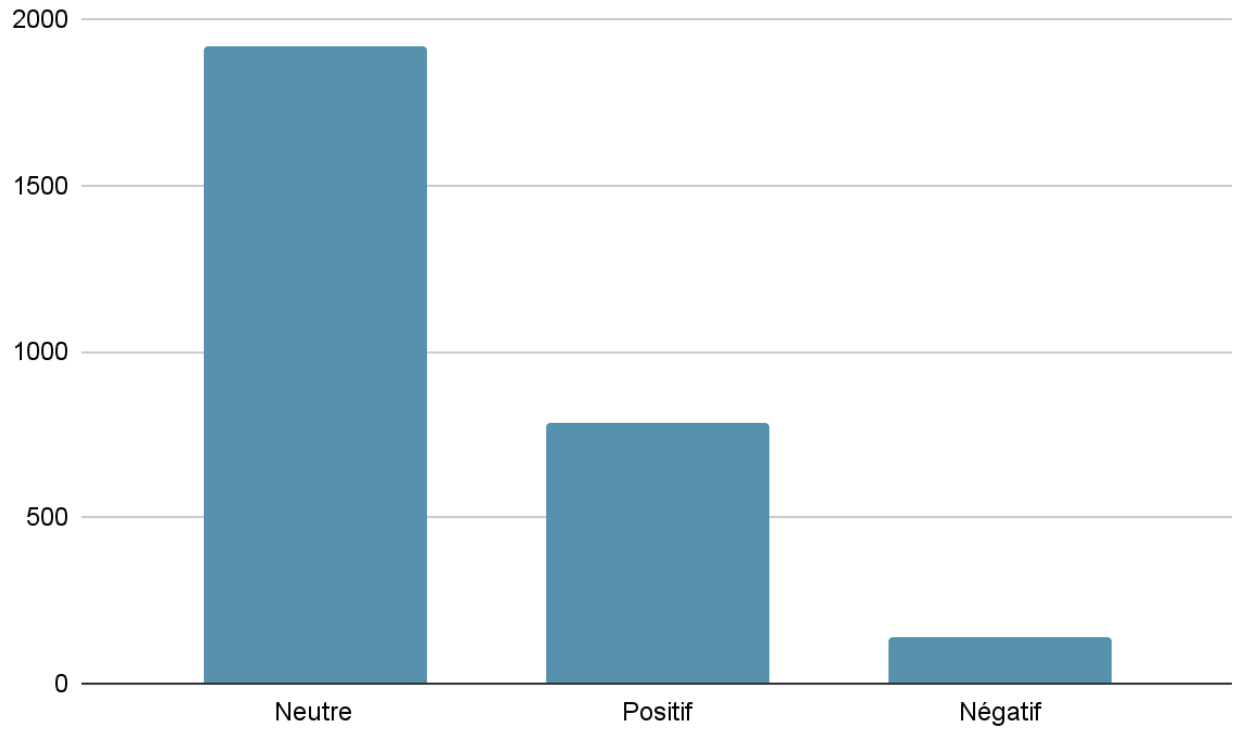

Figure 11. Nombre de commentaires selon le sentiment

4.2.6. L’extraction des caractéristiques

Lorsqu'il s'agit de traiter des données brutes en vue de l'analyse ou de l'entraînement de modèles, il est souvent nécessaire d'extraire des caractéristiques significatives et informatives. C’est précisément l'objectif de l'extraction de caractéristiques, qui utilise des techniques couramment utilisées en traitement du langage naturel telles que les modèles Bag of Words (BoW) et TF-IDF. Ces techniques permettent de transformer les données en informations exploitables, facilitant ainsi l'analyse et l'interprétation des résultats.

Bag of Words

Le modèle Bag of Words (BoW) est l'une des techniques les plus simples et les plus courantes pour transformer le texte en une représentation numérique. Cependant, il présente

l'inconvénient de perdre l'information syntaxique du texte, car il ne prend pas en compte l'ordre des mots, la structure de la phrase ou la construction grammaticale, et ne considère que l'occurrence d'un mot. Par exemple, en prenant les phrases suivantes :

S1: "La caméra de ce téléphone est géniale".

S2: "Je veux ce téléphone ; tout est dans la caméra. Je l'adore".

Tout d'abord, le modèle BoW crée un vocabulaire de tous les mots uniques présents dans le document, puis encode chaque phrase (S1 et S2 dans ce cas) sous forme d'un vecteur de longueur fixe, correspondant à la taille du vocabulaire de mots connus, où la valeur de chaque position dans le vecteur représente le compte ou la fréquence de chaque mot dans l'ensemble d'entraînement [10].

Term Frequency-Inverse Document Frequency

TF-IDF est un schéma de pondération de termes couramment utilisé pour représenter des documents textuels sous forme de vecteurs. Ce processus est utilisé à des fins de classification, de clustering, de visualisation, de récupération, etc. Pour ce faire, on considère un ensemble de tous les termes (T) apparaissant dans le corpus de documents à étudier, puis chaque document (d_i) est représenté par un vecteur de valeurs réelles à n dimensions ($x_i = (x_{i1},\ldots, x_{in})$) avec une composante pour chaque terme possible de T.

Le poids x_{ij} correspondant au terme t_j dans le document d_i est généralement le produit de trois parties : une dépendant de la présence ou de la fréquence de t_j dans d_i, une dépendant de la présence de t_j dans le corpus dans son ensemble, et une partie de normalisation dépendant de d_j. Le poids TF-IDF le plus courant est défini par $x_{ij} = TF_i \cdot IDF_j \cdot (\sum_j (TF_{ij} \cdot IDF_j^2))^{-1/2}$, où TF_{ij} est la fréquence du terme (c'est-à-dire le nombre d'occurrences) de t_j dans d_i, et IDF_j est la fréquence inverse du document, qui est le logarithme du nombre total de documents divisé par le nombre de documents contenant t_j. [14].

4.2.7. La modélisation thématique

Nous avons appliqué trois algorithmes différents, à savoir LSA, LDA et HDP. Chaque algorithme est ensuite brièvement expliqué ci-dessous à travers des figures, en décrivant sa méthode de fonctionnement respective.

4.2.7.1. LSA

LSA repose sur la réduction de la dimensionnalité des données textuelles pour découvrir les relations sémantiques entre les mots. Cette méthode de traitement du langage naturel permet d'effectuer des tâches telles que la recherche d'information et la catégorisation de documents [11].

Algorithme 1 : Algorithme LSA

DEBUT
Input : documents
DÉFINIR le nombre de sujets
CRÉER une matrice terme-document
APPLIQUER une pondération tf-idf à la matrice terme-document
CALCULER la décomposition en valeurs singulières (SVD) de la matrice terme-document
RÉDUIRE la dimension de la matrice SVD en ne gardant que les k premières valeurs singulières
UTILISER la matrice SVD réduite pour représenter les documents et les termes dans l'espace des sujets
FIN

Figure 12. Algorithme LSA

4.2.7.2. LDA

LDA repose sur un modèle statistique appelé distribution de Dirichlet pour découvrir les thèmes latents dans un ensemble de documents. Cette méthode de traitement du langage naturel considère chaque document comme une combinaison de thèmes, et chaque thème comme une distribution de mots. En utilisant des techniques probabilistes, LDA infère les distributions de thèmes dans les documents et les distributions de mots dans les thèmes [11].

Algorithme 2 : Algorithme LDA

```
DEBUT
Input : documents
DÉFINIR le nombre de sujets
DÉFINIR le nombre d'itérations
DÉFINIR alpha
DÉFINIR beta
DÉFINIR les affectations de sujets pour chaque mot dans chaque document au hasard
POUR chaque document d dans les documents FAIRE
  POUR chaque mot w dans d FAIRE
    POUR chaque sujet t FAIRE
      CALCULER p(t|d) * p(w|t)
    FIN POUR
    NORMALISER les probabilités
    CHOISIR un nouveau sujet en fonction des probabilités
  FIN POUR
FIN POUR
FIN
```

Figure 13. Algorithme LDA

4.2.7.3. HDP

HDP (Hierarchical Dirichlet Process) est une méthode statistique non paramétrique utilisée pour regrouper des données groupées. Elle repose sur l'utilisation de processus de Dirichlet pour chaque groupe de données, où tous les groupes partagent une distribution de base tirée d'un processus de Dirichlet. Cette approche permet aux groupes de partager des caractéristiques communes et facilite la découverte de structures latentes dans les données [15].

Algorithme 3 : Algorithme HDP

DÉBUT
Input : Corpus, α, γ
Échantillonner la distribution de thème global G ~ GEM(γ)
POUR chaque document dans le corpus **FAIRE**
  Échantillonner les proportions de thème θ ~ Dirichlet (α * G)
    POUR chaque mot dans le document **FAIRE**
      Échantillonner une attribution de thème local z ~ Multinomial(θ)
      Échantillonner un mot w ~ Multinomial(β_z), où β_z est la distribution de thème-mot pour le thème local z
    FIN POUR
FIN POUR
Échantillonner les proportions de thème pour chaque thème global k ~ Beta (1, γ)
Échantillonner la distribution de thème-mot pour chaque thème global k ~ Dirichlet(β_0)
FIN

Figure 14. Algorithme HDP

4.2.8. L'analyse des sentiments

Nous avons mis en place trois algorithmes à savoir KNN (K-Nearest Neighbors), SVM (Support Vector Machines), et l'algorithme d'arbre de décision en utilisant deux techniques distinctes d'extraction de caractéristiques en l'occurrence TF-IDF et Bag of Words, En plus de cela, nous avons implémenté deux algorithmes spécifiques pour chaque technique d'extraction de caractéristiques, y compris l'algorithme de régression logistique en utilisant TF-IDF et l'algorithme de régression linéaire en s'appuyant sur Bag of Words. Finalement, nous avons utilisé l'algorithme de transfer learning (TL) en appliquant trois modèles pré-entraînés spécifiquement adaptés au traitement du langage naturel en arabe à savoir : ArabBert, DziriBert et XLM.

Une brève explication est fournie pour chacun des algorithmes mentionnés, illustrant leur fonctionnement spécifique, comme le montrent les figures ci-dessous.

4.2.8.1. KNN

L'algorithme KNN (“Algorithme des k-plus proches voisins”, en français) est une méthode d'apprentissage supervisé utilisée pour la classification et la régression. Elle repose sur la sélection des k exemples d'entraînement les plus proches de l'entrée et renvoie une appartenance de classe ou une valeur de propriété pour l'objet [16].

Algorithme 4 : Algorithme KNN

```
DEBUT
Input : données d'entraînement, données de test, nombre de voisins k
POUR chaque point de données de test FAIRE
   CALCULER la distance entre le point de données de test et chaque point de données
d'entraînement
   TROUVER les k points de données d'entraînement les plus proches du point de données de test
   ATTRIBUER l'étiquette majoritaire parmi les k voisins au point de données de test
FIN POUR
CALCULER la précision du modèle sur les données de test
FIN
```

Figure 15. Algorithme KNN

4.2.8.2. SVM

L'algorithme SVM (Support Vector Machine), est un algorithme supervisé, qui est particulièrement adapté aux ensembles de données de taille réduite mais complexes. Il peut être utilisé pour résoudre des problèmes de régression et de classification, mais il est généralement plus efficace pour les tâches de classification [17].

Algorithme 5 : Algorithme SVM

DEBUT
Input : données d'entraînement
DÉFINIR les paramètres C et gamma
CONVERTIR les données d'entraînement en un format approprié pour l'algorithme SVM
ENTRAÎNER un modèle SVM avec les données d'entraînement et les paramètres C et gamma
UTILISER le modèle entraîné pour prédire les étiquettes des données de test
CALCULER la précision du modèle sur les données de test
FIN

Figure 16. Algorithme SVM

4.2.8.3. Algorithme de l'arbre de décision

Un arbre de décision est un modèle de classification structuré en arborescence, facile à comprendre même pour les utilisateurs non experts, et qui peut être efficacement induit à partir des données. L'induction d'arbres de décision est l'une des techniques les plus anciennes et les plus populaires pour apprendre des modèles discriminatoires, qui a été développée indépendamment dans les domaines statistiques [18].

Algorithme 6 : Algorithme de l'arbre de décision

DEBUT
Input : données d'entraînement, données de test, critère de division
CRÉER un arbre de décision à partir des données d'entraînement en utilisant le critère de division
POUR chaque point de données de test **FAIRE**
PARCOURIR l'arbre de décision en utilisant les valeurs des caractéristiques du point de données de test
ATTRIBUER l'étiquette prédite au point de données de test en fonction du nœud final atteint
FIN POUR
CALCULER la précision du modèle sur les données de test
FIN

Figure 17. Algorithme de l'arbre de décision

4.2.8.4. Algorithme de régression logistique

La régression logistique est une technique d'analyse de régression qui permet de prédire le résultat d'une variable dépendante catégorielle basée sur une ou plusieurs variables prédictives [19].

Algorithme 7 : Algorithme de régression logitsique

```
DEBUT
  Input : données d'entraînement, données de test, taux d'apprentissage, nombre d'itérations
  INITIALISER les poids à des valeurs aléatoires
  POUR chaque itération FAIRE
    CALCULER les prédictions pour les données d'entraînement en utilisant la fonction sigmoïde
    CALCULER l'erreur entre les prédictions et les étiquettes réelles
    METTRE À JOUR les poids en utilisant la descente de gradient
  FIN POUR
  CALCULER les prédictions pour les données de test en utilisant la fonction sigmoïde
  CALCULER la précision du modèle sur les données de test
FIN
```

Figure 18. Algorithme de régression logistique (TF-IDF)

4.2.8.5. Algorithme de régression linéaire

La régression linéaire est une méthode statistique pour modéliser la relation entre une variable dépendante et une ou plusieurs variables indépendantes [19].

Algorithme 8 : Algorithme de régression lineaire

DEBUT

Input : données d'entraînement, données de test, taux d'apprentissage, nombre d'itérations

INITIALISER les poids à des valeurs aléatoires

POUR chaque itération **FAIRE**

CALCULER les prédictions pour les données d'entraînement en utilisant l'équation de la droite

CALCULER l'erreur entre les prédictions et les étiquettes réelles

METTRE À JOUR les poids en utilisant la descente de gradient

FIN POUR

CALCULER les prédictions pour les données de test en utilisant l'équation de la droite

CALCULER la précision du modèle sur les données de test

FIN

Figure 19. Algorithme de régression linéaire (BoW)

4.2.8.6. Algorithme Transfer Learning

L'apprentissage par transfer est une approche de l'apprentissage automatique qui capitalise sur la similitude des données, leur distribution et la tâche du modèle afin de transférer les connaissances préalablement acquises d'un domaine à un autre [10].

DziriBert

Algorithme 9: Algorithme TransferLearning (DziriBert)

```
DEBUT
Input: source data, target data
DÉFINIR le modèle pré-entraîné DziriBERT
DÉFINIR les couches à geler
DÉFINIR les couches à entraîner
POUR chaque couche dans le modèle FAIRE
  SI la couche est dans les couches à geler ALORS
    GELER les poids de la couche
  FIN SI
FIN POUR
ENTRAÎNER le modèle sur les données source
ENTRAÎNER les couches à entraîner sur les données cible
FIN
```

Figure 20. Algorithme Transfer Learning en utilisant DziriBert

ArabBert

Algorithme 10: Algorithme TransferLearning (ArabBert)

```
DEBUT
Input: source data, target data
DÉFINIR le modèle pré-entraîné ArabBERT
DÉFINIR les couches à geler
DÉFINIR les couches à entraîner
POUR chaque couche dans le modèle FAIRE
  SI la couche est dans les couches à geler ALORS
    GELER les poids de la couche
  FIN SI
FIN POUR
ENTRAÎNER le modèle sur les données source
ENTRAÎNER les couches à entraîner sur les données cible
FIN
```

Figure 21. Algorithme Transfer Learning en utilisant ArabBert

XLM

Algorithme 11: Algorithme TransferLearning (XLM)

```
DEBUT
    Input : données d'entraînement, données de test, taux d'apprentissage, nombre d'itérations
    CHARGER le modèle pré-entraîné XLM
    INITIALISER les poids du modèle à des valeurs aléatoires
    POUR chaque itération FAIRE
        CALCULER les prédictions pour les données d'entraînement en utilisant le modèle XLM
        CALCULER l'erreur entre les prédictions et les étiquettes réelles
        METTRE À JOUR les poids du modèle en utilisant la descente de gradient
    FIN POUR
    CALCULER les prédictions pour les données de test en utilisant le modèle XLM
    CALCULER la précision du modèle sur les données de test
FIN
```

Figure 22. Algorithme Transfer Learning en utilisant XLM

4.3. Conclusion

Ce chapitre de méthodologie s'achève sur une revue des différentes méthodes utilisées pour la collecte, le nettoyage et le prétraitement des données. Nous avons également expérimenté diverses techniques d'extraction de caractéristiques, notamment TF-IDF et Bag of Words, afin d'améliorer la classification des textes. Enfin, plusieurs approches algorithmiques ont été développées pour l'analyse des sentiments et la modélisation thématique.

Dans le chapitre suivant, nous présenterons les résultats de nos algorithmes en détail pour mieux comprendre l'impact de nos choix méthodologiques sur les résultats de notre étude.

Chapitre 04 : Résultats et discussion

5.1. Introduction

Ce chapitre se concentrera sur la mise en œuvre pratique des différentes méthodologies sur notre ensemble de données. Les résultats obtenus seront présentés de manière visuelle à l'aide de tableaux et de figures, accompagnés d'une explication détaillée.

5.2. Environnement de travail

5.2.1 Matériel utilisé

Les caractéristiques du matériel utilisé sont présentées dans le **tableau 2** ci-dessous.

Tableau 2. Caractéristiques du matériel utilisé

Caractéristiques	**Poste de travail N°01**	**Poste de travail N°02**
PC	DELL	HP
Système d'exploitation	Windows 10 Professionnel	Windows 10 Professionnel
Processeur	Intel(R) Core(TM) i3 - 7020U CPU @ 2.3 GHz	Intel(R) Core(TM) i7 - 6600U CPU @ 2.60 GHz 2.81 GHz
RAM	4,00 GO	16,00 GO
Type de système	SE 64 bits	SE 64 bits

5.2.2 Logiciel utilisé

Dans le cadre de l'implémentation de notre projet nous avons fait usage du logiciel Python (version 3.11.0) qui est un langage interprété de haut niveau, orienté objet et doté d'une sémantique dynamique. Les structures de données intégrées de haut niveau, associées à un typage et à une

liaison dynamique, en font un choix intéressant pour le développement et l'implémentation de la thèse de notre projet. La syntaxe simple et facile à apprendre de Python privilégie la lisibilité et permet de réduire les coûts de maintenance du programme [20].

5.2.3 Les principaux packages Python utilisés

- **Scikit-Learn :** Scikit-learn est une bibliothèque open source d'analyse de données, et la référence en matière d'apprentissage automatique dans l'écosystème Python. Cette bibliothèque offre une grande variété d'algorithmes pour l'aide à la prise de décision en utilisant des données. Parmi les principales fonctionnalités de Scikit-learn, on peut citer : Les méthodes de prise de décision algorithmiques, comprenant : La classification, La régression et Le clustering [21].
- **NumPy :** NumPy est une bibliothèque Python open source utilisée pour travailler avec des tableaux. Elle propose également des fonctions pour travailler dans le domaine de l'algèbre linéaire, de la transformée de Fourier et des matrices [22].
- **Gensim :** Gensim est une bibliothèque open source pour la modélisation de sujets non supervisée, l'indexation de documents, la récupération par similarité et d'autres fonctionnalités de traitement du langage naturel, en utilisant des techniques modernes d'apprentissage automatique statistique [23].
- **PyTorch :** PyTorch est une bibliothèque d'apprentissage automatique open source utilisée pour créer des réseaux de neurones profonds et est écrite dans le langage de script Lua. Elle est l'une des plateformes privilégiées pour la recherche en apprentissage profond [24].
- **TQDM :** TQDM est une bibliothèque Python utilisée pour créer des barres de progression. Elle tire son nom du mot arabe "taqaddum" qui signifie "progrès" [25].

5.3. Analyse exploratoire de données

5.3.1. Caractéristiques du jeu de données

Les spécifications du jeu de données final que nous avons employé sont exposées ci-dessous dans le **tableau 3**.

Tableau 3. Caractéristiques de la base de données

Nombre de colonnes	11
Nombre de commentaires avant nettoyage	3562
Nombre de commentaires arabes avant nettoyage	2093
Nombre de commentaires non-arabes avant nettoyage	1469
Nombre de commentaires après nettoyage et traduction	2981

5.3.2. WordCloud des données

Le wordcloud est une représentation graphique de mots, dont la dimension de chaque mot est proportionnelle à sa fréquence d'apparition dans un texte ou un groupe de textes. Cet outil est souvent utilisé pour fournir rapidement un aperçu des thèmes ou des sujets dominants dans un texte ou un ensemble de textes. Les mots les plus utilisés sont affichés en plus grande taille, tandis que ceux moins fréquents sont représentés en plus petite taille.

La **figure 23** présentée ci-dessus suit cette même logique en mettant en évidence les mots les plus récurrents.

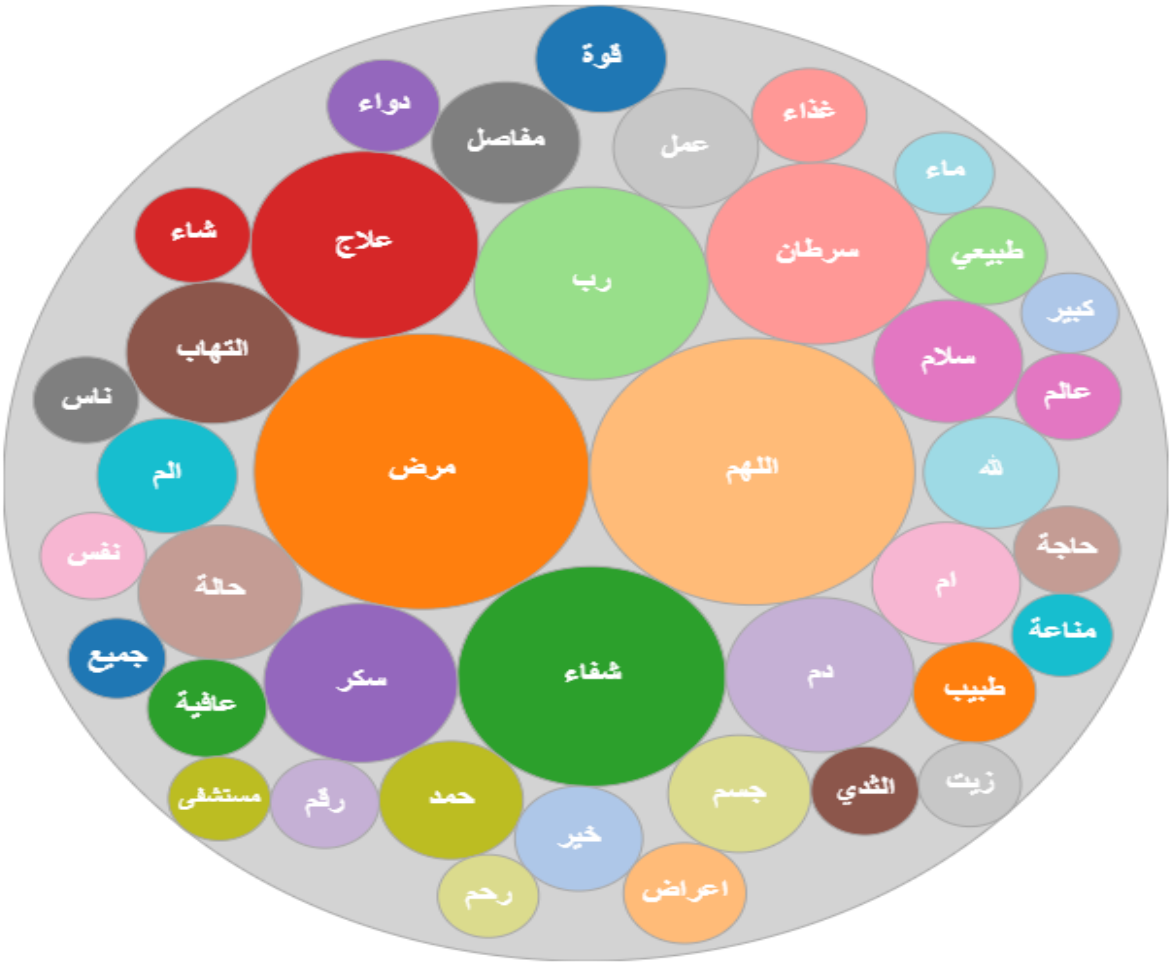

Figure 23. Les mots les plus couramment utilisés dans la base de données

5.3.3. Fréquences des bigrammes et trigrammes

L'exploration des fréquences des bigrammes et trigrammes dans un ensemble de textes vise à examiner la récurrence des séquences constituées de deux (bigrammes) ou trois (trigrammes) mots consécutifs. Dans le cadre de cette analyse, il est procédé au décompte des occurrences de chaque bigramme et trigrammes présents dans le corpus. Les séquences de mots les plus prévalentes sont ensuite identifiées et représentées sous forme de graphiques à barres, en fonction de leur fréquence.

Les résultats des bigrammes présentés dans la **figure 24** révèlent que les séquences de mots "مرض سكر" et "التهاب مفاصل" ont été relevées respectivement 19 et 17 fois dans notre base de données. Par ailleurs, l'expression "حمد الله" a été identifiée 13 fois.

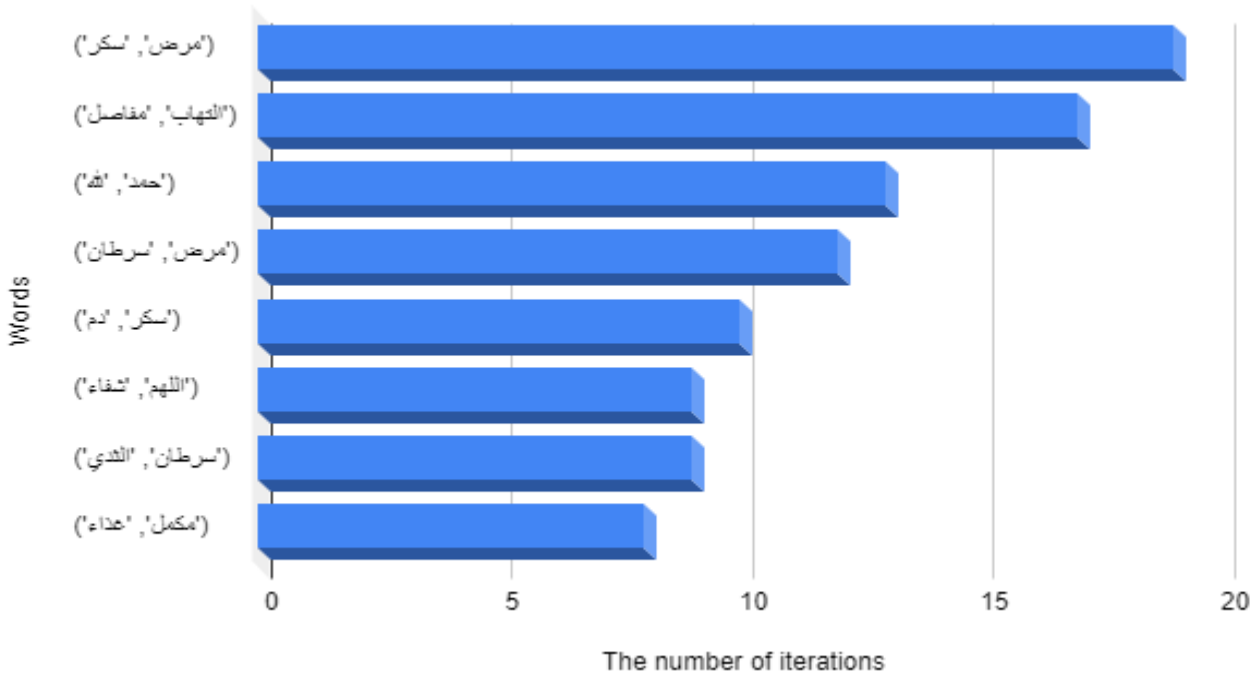

Figure 24. Les bigrammes les plus fréquemment employés dans la base de données

Lors de l'analyse des trigrammes, les résultats obtenus à partir des données analysées dans la **figure 25** révèlent que les séquences de mots "اصاب التهاب مفاصل" et "علاج التهاب مفاصل" ont été identifiées 5 fois, tandis que les séquences "مرض سكر نوع" et "التهاب مفاصل روماتيزم" ont été relevées 4 fois.

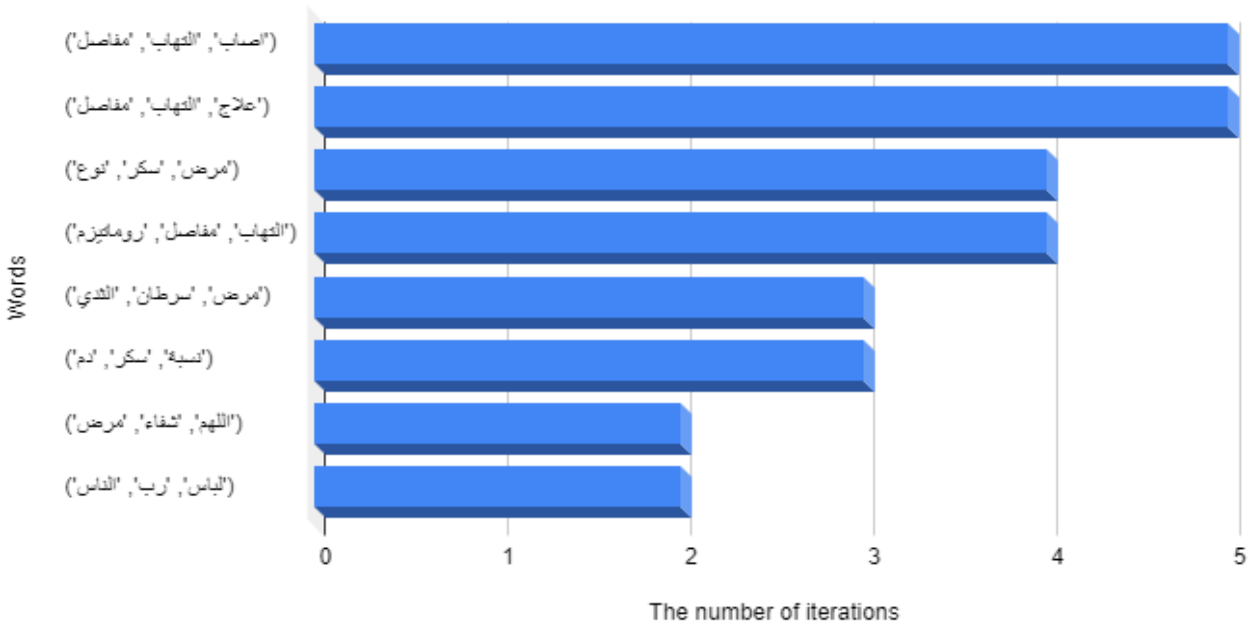

Figure 25. Les trigrammes les plus fréquemment employés dans la base de données

5.4. Prétraitement de données

Le **tableau 4** affiché ci-dessus fournit une explication détaillée de chaque étape en l'illustrant avec un exemple concret.

Tableau 4. Les étapes du prétraitement des données

Étape	**Avant prétraitement**	**Après prétraitement**
Nettoyage	" مَريضُ السُكَّرِ يَعاني مِن ارْتِفاعِ نِسْبَةِ السُّكَّرِ فِي الدَّم"	"مريض السكر يعاني من ارتفاع نسبة السكر في الدم"
Normalisation	"أنا أعاني من ارتفاع ضغط الدم والكوليسترول، أنا خائف!! ☹ "	"انا اعاني من ارتفاع ضغط الدم والكوليسترول أنا خائف "
Suppression des mots vides	"يجب على مرضى الضغط المرتفع	["مرضى", "الضغط", "المرتفع",

	الابتعاد عن الأطعمة الغنية بالملح"	"الابتعاد", "الأطعمة", "الغنية", "الملح"]
Stemming	"أمراض القلب والشرايين"	[" مرض","قلب", "شريان"]
Lemmatisation	"مرض السكري يسبب أعراضاً مثل زيادة العطش والجوع"	""مرض السكر يسبب أعراض مثل زيادة العطش والجوع"
Tokenisation	"أعاني مرض السكري مرض القلب"	["أعاني", "مرض", "السكري", "مرض", "القلب"]

La **figure 26** ci-dessous illustre un texte avant et après avoir été soumis au processus de prétraitement

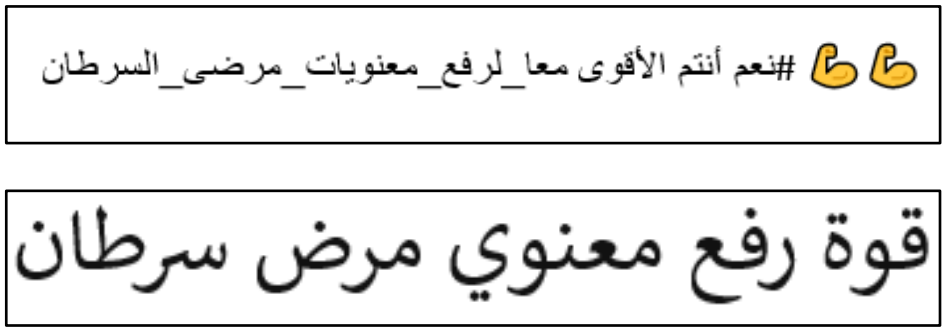

Figure 26. Un texte avant et après le prétraitement

5.5. Résultats de l'implémentation

5.5.1 La modélisation thématique

Dans le cadre de notre étude, nous avons réalisé une modélisation thématique de notre corpus de données en identifiant 5 thèmes pour chaque algorithme, puis avons généré des tableaux présentant les résultats de chaque algorithme, ainsi que les scores de cohérence correspondants aux modèles de chaque algorithme.

Par ailleurs, les figures ci-dessous qui suivent les tableaux illustrent les performances de chaque modèle en fonction du nombre de thèmes identifiés. L'objectif de cette démarche est de

déterminer le nombre optimal de thèmes nécessaires pour obtenir une représentation optimale des thèmes dans notre corpus de données.

Modèle LSA

Tableau 5. Modélisation thématique avec 5 thèmes en utilisant LSA

Topic 00		Topic 01		Topic 02		Topic 03		Topic 04	
Mot-clé	**Dist**	**Mot-clé**	**Dist**	**Mot-clé**	**Dist**	**Mot-clé**	**Dist**	**Mot-clé**	**Dist**
مرض	0.532	اللهم	-0.585	اللهم	-0.494	سكر	-0.512	مرض	0.482
علاج	0.309	التهاب	0.348	التهاب	-0.339	سرطان	0.429	سكر	-0.342
سرطان	0.261	مفاصل	0.292	سرطان	0.327	مرض	-0.294	جسم	-0.226
التهاب	0.233	شفاء	-0.227	مفاصل	-0.302	علاج	0.275	مفاصل	0.208
سكر	0.182	سرطان	-0.209	مرض	0.257	دم	-0.215	دم	-0.199
دم	0.179	رب	-0.165	زيت	-0.174	حلق	0.178	سرطان	-0.160
مفاصل	0.177	زيت	0.160	الثدي	0.140	التهاب	0.168	ثمر	-0.141
اللهم	0.168	حمد	-0.120	شر	-0.128	الثدي	0.167	زيت	-0.135
شفاء	0.136	شر	-0.114	سكر	0.120	مفاصل	0.160	ساعد	-0.132
جسم	0.125	مرض	-0.099	حمد	-0.114	انسولين	-0.112	التهاب	0.128

Modèle LDA

Tableau 6. Modélisation thématique avec 5 thèmes en utilisant LDA

Topic 00		Topic 01		Topic 02		Topic 03		Topic 04	
Mot-clé	**Dist**	**Mot-clé**	**Dist**	**Mot-clé**	**Dist**	**Mot-clé**	**Dist**	**Mot-clé**	**Dist**
اللهم	0.018	سرطان	0.009	اللهم	0.044	مرض	0.013	مرض	0.027
تبرع	0.014	التهاب	0.006	مرض	0.014	اللهم	0.012	التهاب	0.012
رقم	0.011	مرض	0.005	شفاء	0.012	التهاب	0.011	مفاصل	0.012
سلام	0.011	رقم	0.004	سرطان	0.011	علاج	0.011	سكر	0.011
هاتف	0.009	عمل	0.004	زمر	0.009	مفاصل	0.009	علاج	0.010
دم	0.008	طفل	0.004	علاج	0.007	زيت	0.008	دم	0.007
مستشفى	0.007	مستشفى	0.004	دم	0.007	دم	0.008	سرطان	0.006
رب	0.006	علاج	0.004	تبرع	0.006	ام	0.007	الم	0.005
مرض	0.006	حاجة	0.004	حمد	0.006	شفاء	0.007	حالة	0.005
حاجة	0.005	حمد	0.004	جزاك	0.006	جسم	0.006	رب	0.005

Modèle HDP

Tableau 7. Modélisation thématique avec 5 thèmes en utilisant HDP

Topic 00		Topic 01		Topic 02		Topic 03		Topic 04	
Mot-clé	**Dist**	**Mot-clé**	**Dist**	**Mot-clé**	**Dist**	**Mot-clé**	**Dist**	**Mot-clé**	**Dist**
حالة	0.009	وهران	0.014	الم	0.007	مضاد	0.005	ام	0.003
تبرع	0.003	صفايح	0.005	أصاب	0.002	شفاء	0.001	انسان	0.001
اللهم	0.003	حاجة	0.005	الثدي	0.002	زنجبيل	0.001	عمل	0.001
جسم	0.003	تبرع	0.008	دم	0.003	اللهم	0.002	علاج	0.001
ام	0.004	رقم	0.008	سكر	0.003	سرطان	0.002	جسم	0.001
التهاب	0.005	هاتف	0.008	عمل	0.003	علاج	0.002	شفاء	0.001
مفاصل	0.005	مستشفى	0.008	علاج	0.003	مفاصل	0.002	رب	0.002
علاج	0.005	مرض	0.009	مفاصل	0.005	زيت	0.003	سكر	0.002
دم	0.005	دم	0.011	التهاب	0.005	مرض	0.003	مرض	0.002
مرض	0.006	اللهم	0.012	مرض	0.005	التهاب	0.003	اللهم	0.003

Le **tableau 8** ci-dessous montre les scores de cohérence des modèles thématiques LDA, LSA et HDP :

Tableau 8. La performance des modèles de modélisation thématique avec 5 thèmes

Modèles	**Cohérence**
Modèle LDA	0.3147
Modèle LSA	0.4515
Modèle HDP	**0.6623**

Les figures ci-dessous présentent des diagrammes qui illustrent les performances de chaque modèle utilisé en fonction du nombre de thèmes identifiés. Cependant, les graphiques révèlent une instabilité oscillante des performances des modèles LDA et LSA, tandis que HDP présente une variation d'oscillation subtile en comparaison. HDP se démarque ainsi en obtenant des résultats globalement élevés en termes de performance, contrairement à LSA et LDA qui montrent des performances comparativement inférieures.

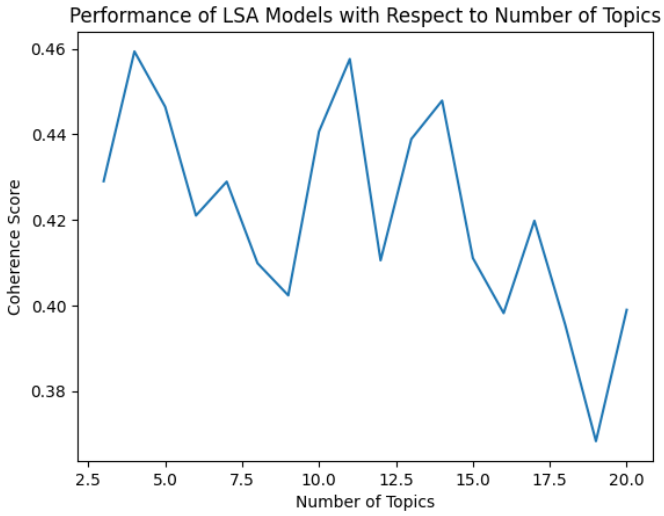

Figure 27. Performance des modèles LSA en fonction du nombre de thèmes identifiés

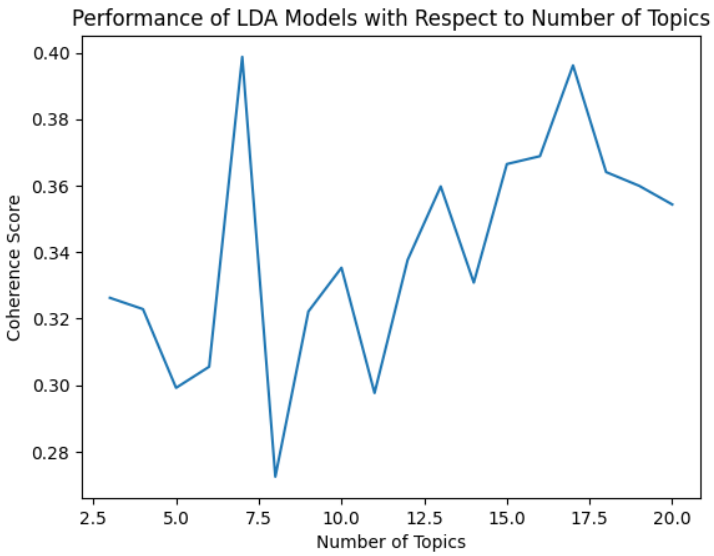

Figure 28. Performance des modèles LDA en fonction du nombre de thèmes identifiés

Figure 29. Performance des modèles HDP en fonction du nombre de thèmes identifiés

5.5.2 L'analyse des sentiments

Nous avons évalué les performances de nos modèles d'analyse de sentiment à l'aide de trois méthodes de classification différentes. La première méthode repose sur la représentation TF-IDF,

tandis que la deuxième méthode utilise la représentation bag-of-words. Enfin, la troisième méthode consiste à employer le transfer learning.

Afin d'évaluer la précision de chaque méthode de classification, l'opération a été réalisée sur notre corpus de données et avons enregistré les résultats d'exactitude pour chacune des méthodes. Ces résultats ont été consignés dans des tableaux illustrés ci-dessous pour chaque méthode employée.

Tableau 9. Les résultats d'exactitude des classificateurs avec TF-IDF

Classificateur	Exactitude	Positif			Neutre			Négatif		
		Précision	Rappel	F1-score	Précision	Rappel	F1-score	Précision	Rappel	F1-score
SVM	0.82	0.80	0.68	0.74	0.83	0.92	0.87	0.58	0.32	0.41
KNN	0.75	0.61	0.86	0.72	0.89	0.73	0.80	0.41	0.32	0.36
LR	**0.83**	0.86	0.65	0.74	0.83	0.95	0.88	0.78	0.32	0.45
DT	0.77	0.70	0.70	0.70	0.82	0.84	0.83	0.33	0.23	0.27

Tableau 10. Les résultats d'exactitude des classificateurs avec BoW

Classificateur	Exactitude	Positif			Neutre			Négatif		
		Précision	Rappel	F1-score	Précision	Rappel	F1-score	Précision	Rappel	F1-score
SVM	0.79	0.69	0.77	0.73	0.86	0.84	0.85	0.58	0.32	0.41
KNN	0.74	0.57	0.91	0.70	0.92	0.70	0.79	0.73	0.36	0.48
LR	**0.81**	0.76	0.71	0.73	0.85	0.90	0.87	0.67	0.36	0.47

DT	0.76	0.66	0.66	0.66	0.82	0.84	0.83	0.60	0.27	0.37

Tableau 11. Les résultats d'exactitude des classificateurs avec le transfer Learning

Classificateur	**Exactitude**	**Positif**			**Neutre**			**Négatif**		
		Précision	**Rappel**	**F1-score**	**Précision**	**Rappel**	**F1-score**	**Précision**	**Rappel**	**F1-score**
DziriBert	**0.85**	0.82	0.79	0.80	0.88	0.91	0.89	0.56	0.41	0.47
XLM	**0.85**	0.80	0.81	0.80	0.89	0.90	0.90	0.50	0.36	0.42
ArabBert	0.84	0.80	0.77	0.78	0.87	0.90	0.89	0.57	0.36	0.44

5.6. Discussion

Comme évoqué précédemment, notre thèse a pour objectif ultime de mieux comprendre les opinions, les émotions et les thèmes prédominants chez les personnes atteintes de maladies chroniques, en utilisant des commentaires liés à ces maladies, et cela à travers l'application d'un ensemble de méthodes d'analyse de données et de traitement du langage naturel.

Nous avons également accordé une importance particulière à l'analyse exploratoire des données textuelles provenant de patients atteints de maladies chroniques. L'objectif principal de cette analyse était de mieux comprendre les thèmes et les sentiments prédominants dans les commentaires des patients. Pour ce faire, nous avons utilisé des techniques de traitement du langage naturel, telles que la détection de bigrammes et de trigrammes. Cependant en examinant les bigrammes et les trigrammes les plus fréquents dans les commentaires des patients, nous avons pu identifier des thèmes clés tel que "علاج التهاب مفاصل", "مرض سكر". Ce qui indique que ces sujets reviennent régulièrement dans les commentaires des patients.

En ce qui concerne les résultats de la modélisation thématique, il est relevé que les modèles LSA et LDA ont obtenu des scores de cohérence de 0,31 et 0,45 respectivement pour 5 thèmes identifiés. Cependant, le modèle HDP obtient une cohérence plus élevée, avec un score qui se situe approximativement autour de 0,66. Quel que soit le nombre de thèmes identifiés, les résultats suggèrent que le modèle HDP est plus efficace pour extraire des thèmes cohérents à partir des données fournies, contrairement aux modèles LSA et LDA qui nécessitent une spécification minutieuse du nombre de thèmes. Cela est dû éventuellement au fait que les modèles LDA et LSA ne tiennent pas compte de la distance entre les mots clés dans un thème, ce qui peut entraîner une cohérence moins élevée. En revanche, le modèle HDP peut intégrer cette information de distance, ce qui lui permet de capturer les relations de proximité entre les termes et d'extraire des thèmes plus cohérents. De plus, les modèles LDA et LSA peuvent avoir des difficultés à capturer les relations complexes et non linéaires entre les mots clés dans un thème, ce qui peut limiter leur capacité à extraire des thèmes cohérents. Le modèle HDP, grâce à son approche probabiliste plus avancée, peut modéliser de manière plus précise ces relations, ce qui se traduit par une cohérence plus élevée dans les thèmes extraits.

Pour notre transition vers l'analyse de sentiments, nous avons opté pour l'utilisation de données étiquetées. Cette décision découle de leur rôle crucial dans l'entraînement de modèles d'apprentissage automatique, tels que les algorithmes de classification. Ces modèles requièrent des exemples d'entrée accompagnés de leurs étiquettes correspondantes afin d'apprendre à généraliser et à réaliser des prédictions précises sur de nouvelles données non étiquetées. Les données étiquetées jouent également un rôle central dans l'évaluation des performances des modèles d'analyse de sentiments. En comparant les prédictions des modèles aux étiquettes de référence, nous sommes en mesure de mesurer l'exactitude, la précision, le rappel et d'autres métriques afin d'évaluer la qualité du modèle de manière approfondie.

Les résultats de l'analyse des sentiments en utilisant les méthodes d’extraction classiques TF-IDF et BOW révèlent des performances globales satisfaisantes pour les classificateurs SVM, KNN, LR et de l'arbre de décision, avec des valeurs d'exactitude allant de 0,74 à 0,83. Ces résultats indiquent que les méthodes d'extraction de caractéristiques utilisées sont efficaces pour capturer les informations discriminantes des données. Notamment, le classificateur KNN se démarque avec

un rappel de 0,91 en utilisant la méthode BOW et de 0,86 en utilisant TF-IDF pour la classe positive, montrant ainsi sa capacité à rappeler davantage de cas positifs que les autres classificateurs. D'autre part, la régression logistique (LR) affiche une valeur d'exactitude de 0,83, ce qui en fait le classificateur le plus précis en termes de performance globale. Elle démontre également de meilleures performances pour la classe positive, avec une précision de 0,86 et un score F1 de 0,74.

En ce qui concerne la classe neutre, les classificateurs se distinguent avec des valeurs de précision supérieures à 0,8, atteignant même 0,92, tel que KNN utilisant la méthode BOW, tandis que les classificateurs SVM et régression logistique (LR) affichent des rappels supérieurs à 0,9. Ces résultats démontrent l'efficacité de ces méthodes de classification pour bien classifier les exemples neutres. Cependant, en ce qui concerne la classe négative, tous les classificateurs montrent des performances relativement faibles, avec des précisions inférieures à 0,6 lors de l'utilisation de la méthode TF-IDF, à l'exception du modèle régression logistique (LR) qui atteint une précision de 0,78. De plus, les rappels sont inférieurs à 0,4 et les scores F1 sont inférieurs à 0,5. Ces résultats indiquent que les classificateurs utilisant les méthodes d'extraction classiques rencontrent des difficultés à bien classifier les cas négatifs.

En gros, ces observations mettent en évidence que le classificateur de la régression logistique (LR) se distingue par son efficacité par rapport aux autres méthodes classiques, en termes d’exactitude et de performances relatives pour les classes positive et négative.

Les résultats obtenus avec le transfer learning révèlent que les classificateurs DziriBert, ArabBert et XLM affichent des performances globalement similaires, avec des valeurs d'exactitude supérieures à 0,80. Cela indique que les trois modèles sont capables de classifier les données avec une précision élevée. En ce qui concerne la classe positive, les trois classificateurs présentent des résultats comparables, avec des précisions supérieures à 0,80. De plus, les rappels et les scores F1 de chaque classificateur convergent autour d'une valeur unique, démontrant ainsi leur cohérence et leur performance équivalente. Cela signifie que cette constance renforce la fiabilité de ces modèles et leur capacité à fournir des prédictions précises et fiables pour chaque classe. Ce qui nous amène à dire que les trois modèles sont efficaces pour détecter les cas positifs et rappellent un pourcentage élevé de ces cas.

Concernant la classe neutre, les trois classificateurs affichent des performances très élevées par rapport à la classe positive, avec des précisions, des rappels et des scores F1 atteignant une valeur de 0,91. Ce qui signifie que les classificateurs sont capables de classifier impeccablement les exemples neutres et d'obtenir des résultats précis. En revanche, pour la classe négative, comme pour les méthodes classiques, les performances des trois classificateurs sont relativement faibles, avec des précisions et des scores F1 inférieurs à 0,60 et des rappels inférieurs à 0,5. Cela indique que les trois modèles ont des difficultés à bien classer les cas négatifs et ne parviennent pas à obtenir des résultats aussi précis que pour les autres classes.

Sur la base de ces observations, il est difficile de déterminer avec certitude quelle méthode, entre DziriBert, ArabBert et XLM, est la plus performante. Bien que les trois modèles présentent des performances similaires dans l'ensemble, il est important de noter que le modèle DziriBert se distingue légèrement avec des résultats supérieurs dans la classe négative par rapport à ArabBert et XLM. Cela suggère que le modèle DziriBert peut être considéré comme plus performant dans ce contexte spécifique.

Enfin, Si l'on devait choisir la meilleure méthode parmi les méthodes du transfer learning et les méthodes classiques, en l’occurence DziriBert et la régression logistique (LR), les résultats suggèrent que le modèle DziriBert surpasse la régression logistique en termes de performances globales. Le modèle du transfert learning affiche une précision élevée avec une valeur d'exactitude de 0,85, tandis que la régression logistique présente des performances légèrement inférieures, avec une valeur d'exactitude de 0,83. De plus, le modèle DziriBert obtient des résultats comparables voire supérieurs à la régression logistique pour les classes positive et neutre. Cependant, bien qu'il puisse être observé que LR obtient de meilleurs résultats que DziriBert en termes de précision dans la classe négative, il convient de noter que les deux méthodes rencontrent des difficultés lorsqu'il s'agit de classifier cette classe spécifique. Ainsi, si l'on devait opter pour la méthode la plus efficace, sur la base des comparaisons effectuées, le modèle DziriBert semble être le choix le plus prometteur en termes de performances globales pour l'analyse de sentiments.

5.7. Conclusion

En conclusion, nous avons exposé les différentes étapes de mise en œuvre de notre projet de fin d'études, en commençant par la présentation de notre environnement de travail, suivie d'une analyse exploratoire des données. Nous avons par la suite détaillé les étapes de prétraitement des données où chaque étape étant illustrée d'exemples. Les résultats de l'analyse de sentiment et de la modélisation thématique ont été présentés en utilisant des tableaux et des figures, et ont été accompagnés d'explications détaillées. Enfin, une discussion a été menée pour évaluer les résultats obtenus tout au long de ce chapitre.

Conclusion générale

En conclusion, ce projet a pour objectif principal de collecter des informations précieuses sur les émotions liées aux maladies chroniques afin de favoriser une meilleure compréhension des soins de santé et contribuer à leur amélioration et cela en faisant usage de deux méthodes d'analyse textuelle, à savoir l'analyse des sentiments et la modélisation thématique.

L'analyse des sentiments a été réalisée en utilisant plusieurs algorithmes et techniques d'extraction de caractéristiques. Les résultats obtenus ont démontré l'efficacité des modèles de transfert d'apprentissage, notamment ArabBert, DziriBert et XLM, pour l'analyse du langage naturel en arabe. Ces modèles ont surpassé les méthodes classiques, améliorant ainsi les performances de l'analyse des sentiments dans le contexte des maladies chroniques.

La modélisation thématique a joué un rôle crucial en utilisant divers algorithmes tels que LDA, LSA et HDP. Grâce à l'application de ces techniques, des informations pertinentes et significatives ont été extraites concernant les thèmes liés aux maladies chroniques. Ces découvertes ouvrent des perspectives pour des améliorations futures et une prise en compte plus efficace des besoins émotionnels des patients souffrant de maladies chroniques, dans le but d'améliorer les soins de santé dans ce domaine particulier.

Il convient ainsi de souligner que ce projet présente certaines limites, cependant les résultats obtenus dépendent de la qualité et de la quantité des données disponibles. De plus, l'utilisation de modèles pré-entraînés adaptés à des contextes spécifiques peut être limitée par leur disponibilité et leur adaptabilité aux différents dialectes de la langue arabe.

Pour les améliorations futures, il est recommandé de se pencher sur l’utilisation des jeux de données plus vastes et variés afin d'améliorer la représentativité des analyses. Il convient également d'explorer des approches avancées en matière d'apprentissage automatique, notamment l'utilisation de réseaux de neurones profonds, pour une analyse plus précise des émotions associées aux maladies chroniques. De plus, l'intégration de données provenant de sources supplémentaires, telles que les dossiers médicaux électroniques, pourrait fournir une perspective plus complète des expériences vécues par les patients atteints de maladies chroniques.

Les références

[1]: R. A. Goodman, S. F. Posner, E. S. Huang, A. K. Parekh, et H. K. Koh, "Defining and measuring chronic conditions: imperatives for research, policy, program, and practice," Prev Chronic Dis, vol. 10, p. E66, 2013.

[2]: A. Ly, J. Afr, "Cancers et autres maladies non transmissibles : vers une approche intégrée de santé publique - Scientific Figure on ResearchGate," Cancer, vol. 4, pp. 137-139, 2012.

[3]: B. M. Rothschild, Ed., "Principles of Osteoarthritis - Its Definition, Character, Derivation and Modality-Related Recognition", IntechOpen, 2012.

[4]: National Asthma Education and Prevention Program, "Expert Panel Report 3: Guidelines for the Diagnosis and Management of Asthma," Bethesda, MD, USA: National Heart, Lung, and Blood Institute, 2007.

[5]: F. Pezzella, M. Tavassoli, et D. J. Kerr, Eds., Oxford Textbook of Cancer Biology, Oxford University Press, 2019.

[6]: World Health Organization, “Classification of diabetes mellitus,” World Health Organization, Geneva, 2019

[7]: C. E. M. Griffiths, A. W. Armstrong, J. E. Gudjonsson, et J. N. W. N. Barker, "Psoriasis," VOLUME 397, ISSUE 10281, pp. 1301-1315, April 03, 2021.

[8]: H. L. T. Mobley, G. L. Mendz, et S. L. Hazell, Eds., Helicobacter pylori: Physiology and Genetics, Wiley Online Books, 2001.

[9]: J. Torres, S. Mehandru, J. F. Colombel, et L. Peyrin-Biroulet, "Crohn's disease," The Lancet, vol. 389, no. 10080, pp. 1741-1755, 2017.

[10]: M. Birjali, M. Kasri, et A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges, and trends," Knowledge-Based Systems, vol. 226, p. 107134, 2021.

[11]: P. Kherwa et P. Bansal, "Topic modeling: a comprehensive review," EAI Endorsed Transactions on Scalable Information Systems, vol. 7, no. 24, 2020.

[12]: Wikipedia, “Probabilistic latent semantic analysis,”.

https://en.wikipedia.org/wiki/Probabilistic_latent_semantic_analysis

[13]: R. Albalawi, T. H. Yeap, et M. Benyoucef, "Using topic modeling methods for short-text data: A comparative analysis," Frontiers in Artificial Intelligence, vol. 3, p. 42, 2020.

[14]: C. Sammut et G. I. Webb, Eds., Encyclopedia of Machine Learning, Springer, 2011.

[15]: Y. W. Teh and M. I. Jordan, "Hierarchical Bayesian Nonparametric Models with Applications," in Bayesian Nonparametrics, Cambridge University Press, pp. 158-207, 2010.

[16]: Wikipedia, “k-nearest neighbors algorithm,”.

https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[17]: A. Saini, "Support Vector Machine (SVM): A Complete guide for beginners,", 2021.

[18]: J. Fürnkranz, T. Kliegr and H. Paulheim, “On cognitive preferences and the plausibility of rule-based models,” Machine Learning, vol. 109, pp. 853-898, 2020.

[19]: S. Chatterjee et J. S. Simonoff, "Handbook of Regression Analysis," John Wiley & Sons, Inc., 2013.

[20]: Python.org, “What is Python? Executive Summary,”.

https://www.python.org/doc/essays/blurb/

[21]: ActiveState, “What is Scikit-Learn in Python?,” August 5, 2022.

https://www.activestate.com/resources/quick-reads/what-is-scikit-learn-in-python

[22]: W3Schools, “Introduction to NumPy,”.

https://www.w3schools.com/python/numpy/numpy_intro.asp

[23]: Wikipedia, “Gensim,”. https://en.wikipedia.org/wiki/Gensim

[24]: K. Yasar and S. Lewis, “What is PyTorch?,” TechTarget.

https://www.techtarget.com/searchenterpriseai/definition/PyTorch

[25]: R. Shah, “TQDM: Python Library to Monitor Code Progress (Updated 2023),” Analytics Vidhya, May 22, 2021. https://www.analyticsvidhya.com/blog/2021/05/how-to-use-progress-bars-in-python