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

FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

DEPARTEMENT D'INFORMATIQUE

N°: ……

DOMAINE : MATHEMATIQUES ET INFORMATIQUE

FILIERE : INFORMATIQUE

OPTION : SYSTEME D'INFORMATION ET GENIE LOGICIEL

MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de Master Académique

Spécialité : Systèmes d'Informations Avancés

Par : HADJI Nour ELhouda

LAICHI Nour ELhouda

SUJET

L'Analyse Automatique des sentiments dans les textes Arabes

Soutenu publiquement en : juin 2022 devant le jury composé de :

……	**Université de M'sila**	**Président**
Mm. HELASSA Madiha	**Université de M'sila**	**Rapporteur**
……	**Université de M'sila**	**Examinateur**

Promotion : 2021/2022

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

FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

DEPARTEMENT D'INFORMATIQUE

N°: ……

DOMAINE : MATHEMATIQUES ET INFORMATIQUE

FILIERE : INFORMATIQUE

OPTION : SYSTEMES D'INFORMATION ET GENIE LOGICEL

MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de Master Académique

Spécialité : Systèmes d'Informations Avancés

Par : HADJI Nour ELhouda

LAICHI Nour ELhouda

SUJET

L'Analyse Automatique des sentiments dans les textes Arabes

Soutenu publiquement en : juin 2022 devant le jury composé de :

……	**Université de M'sila**	**Président**
Mm. HELASSA Madiha	**Université de M'sila**	**Rapporteur**
……	**Université de M'sila**	**Examinateur**

Promotion : 2021/2022

Dédicace

Dieu tout Puissant merci pour le pouvoir et le courage que vous nous avez donné pour compléter ce travail.

Avec mes sentiments de gratitudes les plus profonds, que Je dédie :

Je remercie Dieu Tout-Puissant pour la force et le courage qu'il nous a donnés pour mener à bien ce travail.

Avec mes sentiments de gratitudes les plus profonds, que Je dédie :

A mes très chers parents :

Ma très chère mère

Tu m'as donné la vie, la tendresse et le courage pour réussir, tout ce que je peux t'offrir ne pourra exprimer l'amour et la reconnaissance que je te porte.

A mon très cher père

Ce travail est le fruit de tes sacrifices que tu as consentis pour mon éducation et ma formation. Aucune dédicace ne saurait exprimer mes sentiments, que dieu te préserve et te procure santé et longue vie.

Je tiens à remercier profondément À tous les membres de ma famille

A tous mes amis que j'ai vécus avec eux des beaux moments à l'université et À tous les professeurs qui ont été la raison de notre succès

Et à chaque département de la Faculté de Mathématiques et Informatique Média ... Et toute la promotion 2022, Université Mohamed Boudiaf, M'SILA

Remerciements

Nous tenons tout d'abord à remercier Dieu le tout puissant et miséricordieux, qui nous a donné la force et la patience d'accomplir ce Modeste travail, qu'il nous soit permis de témoigner de notre profonde et sincère gratitude envers tous ceux qui ont contribué de loin ou de près.

En second lieu, nous tenons à remercier notre encadreur Mm:

HALASSA MADIHA

Pour avoir répondu positivement à notre demande de d'orientation, de patience et de confiance. Sa rigueur scientifique et ses remarques ont été utiles pour la qualité de ce travail.

Nous présentons également notre gratitude à tous les professeurs, chefs de travaux et assistants de l'université de Mohamed Boudiaf-M'SILA- en général, et singulièrement ceux de la faculté de Mathématique et informatique pour leur dévouement.

Ainsi, nous remercions nos parents:

HADJI MEFTAH & LAICHI KAMAL

Pour leur soutien tant moral, spirituel et matériel.

Merci à toutes

Table des matières

Liste des Figures

Liste des tableaux

Liste des Équation

INTRODUCTION

GENERALE

INTRODUCTION GENERALE

L’internet est devenu un outil indispensable, tant dans le domaine professionnel que dans la vie quotidienne, notamment avec le développement rapide et la popularité des réseaux sociaux. La popularité des médias sociaux est liée au besoin d’information, qui devient de plus en plus important dans notre société. En général, les gens s’expriment et aiment aussi voir les réactions et les interactions des autres, comme les opinions. Cependant, la grande qualité d’informations générées sur ces médias sociaux nécessite des outils adéquats pour les traiter et les analyser. Il est fondamentales à presque toutes les activités humaines et ont une influence majeure sur nos actions et nos croyances. Notre perception des réalités et des choix que nous faisons dépend en grande partie de la façon dont les autres voient et évaluent le monde. Cela est vrai non seulement pour les personnes, mais aussi pour les organisations et les entreprises. Les méthodes traditionnelles d’analyse de cette taille sont certainement inutiles.

Par conséquent, l’analyse des sentiments (opinion mining) est une science spécialisée dans l’identification automatique de ces textes et de les distinguer de positif ou négatif. Il est un domaine émergent du TALN qui vise à analyser les commentaires des utilisateurs sur les réseaux sociaux afin de prendre des décisions dans différents domaines : politique, marketing, santé, éducation…

L'analyse des sentiments fait partie de l'exploration de textes qui vise à identifier les opinions, les sentiments et les attitudes présents dans un texte ou un ensemble de textes. La quantité de données disponibles sur le Web augmente de façon exponentielle. Cependant, ces données sont pour la plupart décrites dans un format non structuré et ne peuvent donc être traitées que par la machine. Par conséquent, les techniques d'exploration de graphes et de traitement du langage naturel (PNL) peuvent aider à extraire des connaissances et des opinions à partir de la grande quantité d'informations présentes sur le Web.

L'analyse des sentiments peut améliorer les capacités des systèmes de gestion de la relation client et de recommandation, par exemple en aidant à découvrir les caractéristiques qui intéressent particulièrement les clients ou en excluant les éléments qui ont reçu des critiques défavorables des publicités.

Problématique d’analyse des sentiments représentés dans Les méthodes d'analyse des sentiments permettent de classer le texte comme positif, négatif. La classification des sentiments pour les textes arabes n'a pas reçu autant d'attention que l'anglais en raison des raisons pour lesquelles, premièrement, l'arabe manque d'outils et de ressources pour extraire les

sentiments arabes du texte et, deuxièmement, l'unicité de la langue arabe parce qu'elle est exceptionnelle, donc le mot peut avoir plusieurs sens. Et riche en composition qui peut complètement changer la question, et aussi un mot peut renvoyer à plusieurs sens différents selon la phrase .Il peut être le nom d'une personne ou un adjectif et aussi le mot peut avoir diverses formes en insérant des accessoires (suffixe ou préfixe).

Pour résoudre ce problème nous allons réaliser plusieurs modèles capables d'analyser et d'extraire l'opinion d'un commentaire (Positif/Négatif). Afin d'atteindre cet objectif nous suivrons la mémoire comme suit:

Dans le premier chapitre, nous avons présenté les définitions et la terminologie utilisées dans ce mémoire. Nous avons vu dans l'ordre, le processus de l'analyse de sentiments, les domaines, les types et les tâches de l'analyse de sentiments et finalement la présentation de la langue arabe et les défis rencontrés au cours de cette analyse.

Dans le deuxième chapitre, l'état de l'art résume les approches les plus prometteuses pour classer les sentiments. Ces approches se répartissent en deux catégories. La première catégorie concerne les méthodes d'apprentissage en profondeur qui classent les documents d'une base de données d'apprentissage.

La deuxième catégorie concerne les méthodes de classifications basées sur le lexique qui sont basées sur le lexique des sentiments, une collection de termes de sentiments connus et précompilés.

Dans le troisième chapitre, nous présentons et identifions les moyens que nous avons utilisés dans notre recherche, et la manière dont nous avons adopté la solution au problème.

Ce mémoire se termine par une conclusion qui, en revenant sur les grandes thématiques qui nous aurons guidés tout au long de cette lecture, sur l'ensemble des techniques d'analyse des sentiments apportées et finalement ses limites. Cette conclusion donne également l'occasion d'exprimer les perspectives de nos travaux de recherche.

CHAPITER 1 :

ANALYSE DES SENTIMENT

CHAPITER 1 : ANALYSE DES SENTIMENTS

1.1 Introduction

Le traitement automatique du langage naturel fait partie des domaines qui ont rencontré une large diffusion car il a fallu un grand défi à travers le traitement du langage qui permet la communication avec les ordinateurs sans avoir recours à des langages de programmation, ce qui a permis à nombreuses personnes d'exprimer leurs opinions sur diverses plateformes de médias sociaux. Dans ce contexte, nous trouvons l'analyse des sentiments, qui est l'un des plus actifs champs de ce domaine, qui s'intéresse particulièrement à l'analyse des opinions. Dans ce chapitre, nous expliquerons les notions l'analyse des sentiments, les tâches et les niveaux, aussi les méthodes d'analyse sentiments, Nous étudierons également l'analyse des sentiments en langue arabe et les différentes complications dans ce domaine

1.2 Traitement Automatique de la langue naturelle (TALN)

Le traitement automatique du langage naturel ou NLP est la science qui combine le langage et un certain nombre de domaines de l'informatique, tels que : l'apprentissage automatique, l'apprentissage en profondeur et les réseaux de neurones artificiels.

Le traitement du langage naturel tente de rendre la machine capable de comprendre et de générer le langage humain, qu'il s'agisse d'un langage écrit ou d'un langage audible. Le traitement du langage naturel est également l'un des domaines les plus importants et les plus difficiles de l'intelligence artificielle, car il est fondamental pour améliorer les machines et les dispositifs d'intelligence artificielle**.** [1]

1.3 Champs de recherche et applications de TALN

Le domaine du traitement des langues naturelles comprend un grand nombre de disciplines de recherche variés en termes d'objectif ou des méthodes de traitement, ainsi que la forme d'information. Ce domaine est divisé en trois axes principaux : Sémantique, Extraction d'informations et Syntaxe, dont chacun comprend nombreux domaines, nous mentionnons certains dans la figure suivante [2]

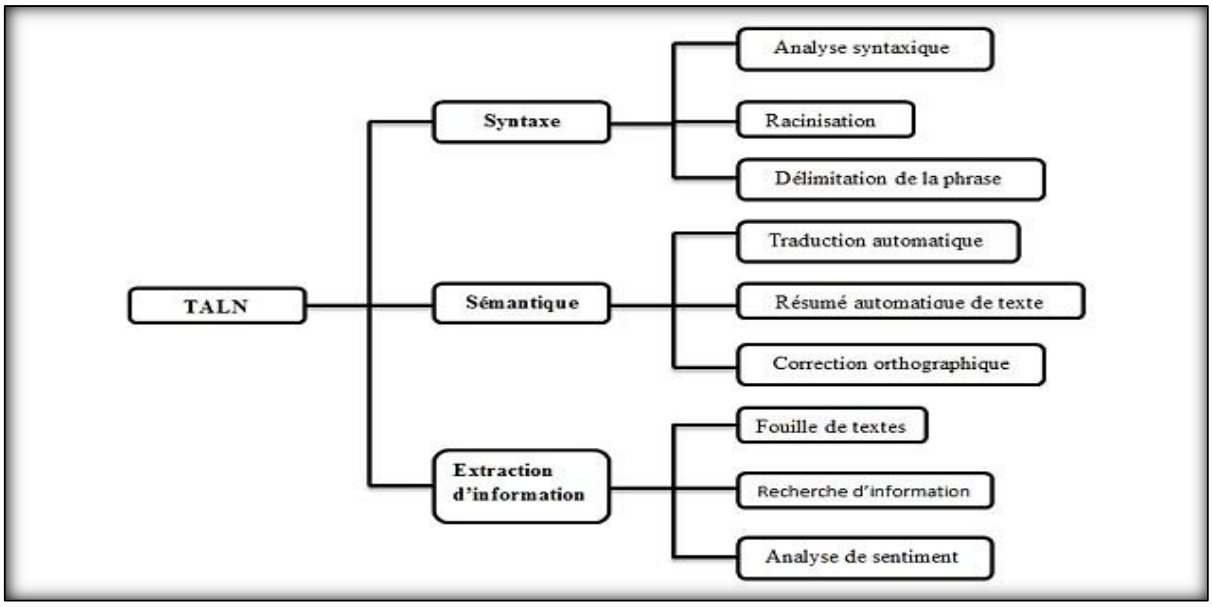


Figure 1. 1 Applications de TALN

1.4 Définition d'analyse de sentiment

Dans la littérature, l'analyse des sentiments (sentiment analysis) est également appelée opinion Mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, appraisal extraction, est un domaine de recherche qui consiste à analyser les sensations, les attitudes et les émotions des individus vis-à-vis des entités telles que les produits, les services et les organisations économique. L'analyse des sentiments est l'un des domaines de recherche les plus actifs en traitement automatique de langage naturel, Machine Learning, statistiques et linguistique depuis le début de l'année 2000. Les origines de l'analyse des sentiments se réfère aux des sciences de la psychologie, la sociologie et de l'anthropologie, qui se concentrent sur les émotions humaines.

L'analyse des sentiments consiste à construire des outils automatiques capables d'extraire des informations subjectives de textes en langage naturel, de manière à créer des connaissances structurées et exploitables pouvant être utilisées par un système d'aide à la décision ou un décideur. [3]

Figure 1. 2 Schéma présenté les sentiments

1.5 Les niveaux de l'analyse des sentiments

De nombreux articles ont présenté différentes approches pour l'analyse des sentiments en arabe, où ils ont traité le problème de la classification des sentiments de différentes manières. En général, l'analyse des sentiments peut être effectuée à trois niveaux différents du texte :

1.5.1 Analyse des sentiments au niveau de la phrase :

A ce niveau, chaque phrase peut avoir une opinion différente, donc la polarité sera calculée pour chaque phrase où chaque phrase est considérée comme une unité séparée.

1.5.2 Analyse des sentiments au niveau des aspects :

Le sentiment et les opinions des phrases individuelles sont déterminées sur un aspect particulier, où les caractéristiques du produit sont identifiées et extraites des données sources.

1.5.3 Analyse des sentiments au niveau du document :

Il s'agit du niveau le plus courant. Comme son nom l'indique, il suppose que l'auteur du document a une opinion sur un objet principal exprimé dans l'ensemble du document, où l'opinion négative est exprimée par une valeur négative et l'opinion positive par une valeur positive. Ainsi, à ce niveau, nous pouvons aller au-delà du problème de la détermination des limites des phrases qui font face au niveau précédent. Mais le niveau du document présente des défis particuliers, notamment le fait de contenir l'article pour plus d'une opinion, ainsi que les sentiments inverses dans le même article, où l'opinion opposée peut invalider l'opinion principale [4].

Il existe également d'autres niveaux d'analyse des sentiments comme :

- Analyse des sentiments au niveau de conception
- Analyse des sentiments au niveau d'utilisateur

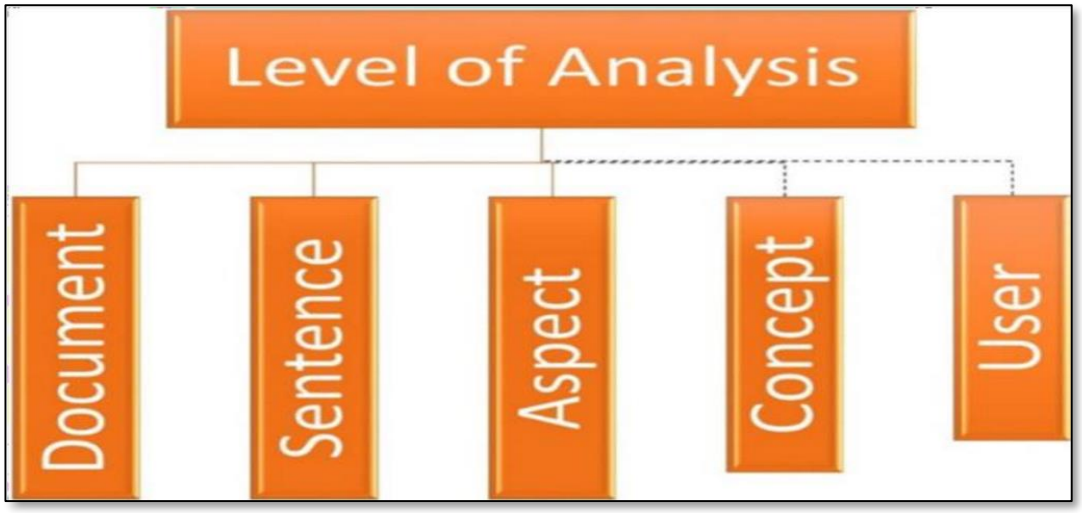


Figure 1. 3 Niveaux d'analyse des sentiments

1.6 Domaine d'analyse de sentiments

L'importance de l'analyse des sentiments est présente dans plusieurs domaines ainsi plusieurs applications ont vu le jour dans ce contexte. Nous mentionnons brièvement quelques applications ci-dessous:

- Politique

Aujourd'hui, les acteurs politiques ont suivi la tendance de l'analyse des sentiments, car avant de déclarer une nouvelle loi, les politiciens tentent de recueillir l'opinion des utilisateurs de médias sociaux sur cette loi. Il est hautement stratégique de connaître également l'opinion des internautes sur un politicien lors d'une élection présidentielle

- .Économie

Avant d'acheter un produit, la majorité des clients demandent conseil sur un produit ou un service donné et sont même disposés à payer plus pour un produit dont l'opinion est plus favorable qu'un autre, ce qui peut augmenter les ventes. Grâce à l'analyse des sentiments, les entreprises peuvent connaître l'opinion des clients sur leurs produits ou leurs services. Dans une perspective d'amélioration de leurs produits et d'augmentation de leurs ventes et revenus.

- Éducation

L'analyse des sentiments peut être utilisée pour extraire des informations utiles sur la méthodologie d'enseignement d'un enseignant et également sur le programme du cours. Il identifie le degré d'apprentissage des étudiants, comprend leurs besoins, prévoit leurs

performances et apporte des changements effectifs dans le style. Les résultats de l'analyse des sentiments aident les enseignants et les établissements à prendre des mesures correctives**.** [2]

1.7 Types d'analyse des sentiments

Il existe de nombreux types d'analyses de sentiments allant des systèmes qui se concentrent sur la classification de la polarité (positif, négatif, neutre) aux systèmes qui détectent des émotions (en colère, heureux, triste, etc.) ou identifient des intentions (par exemple, intéressé, pas intéressé). [2]

1.7.1 Analyse fine des sentiments

Au lieu de parler de phrases positives, négatives ou neutres, nous considérons les catégories suivantes :

- Très positive
- Positive
- Neutre
- Négative
- Très négative

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

1.7.2 Détection d'émotion (Emotion détection)

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

1.7.3 Analyse de sentiments à base d'aspects

Au lieu de classer le sentiment général d'un texte en positif ou en négatif, l'analyse de sentiments à base d'aspects permet d'analyser le texte afin d'identifier différents aspects et de déterminer le sentiment correspondant pour chacun. Les résultats sont plus détaillés, intéressants et précis car l'analyse à base d'aspects examine de manière précise les informations contenues dans un texte. [2]

1.8 Les tâches d'analyse des sentiments

Il existe différentes tâches dans l'analyse des sentiments :

- Catégorisation des sentiments.
- Identification de sujet d'opinion.
- détection de l'opinion

1.9 Challenges et solutions

Étant donné que l'analyse des sentiments dépend fortement de la morphologie de la langue analysée, notre objectif dans la présente section est de donner une brève description de l'arabe et de détailler les caractéristiques linguistiques qui font de cette langue l'une des variétés les plus difficiles pour les chercheurs en analyse des sentiments.

L'arabe est l'une des six langues officielles de Les Nations Unies. C'est la langue officielle de 27 pays et est parlée par plus de 422 millions de personnes dans le monde arabe. Sur le web, l'arabe se classe au quatrième rang des langues les plus utilisées et celle qui connaît la croissance la plus rapide au cours des cinq dernières années avec un taux de croissance de 6091,9 % du nombre d'internautes [5].

L'arabe a trois variétés principales : l'arabe classique ; qui est la langue du Coran (Saint de l'Islam Livre); Arabe standard moderne (MSA) et dialectique Arabe. MSA la variété de langue arabe la plus éloquente utilisée à l'écrit et dans la plupart des discours formels.

L'arabe dialectique ou familier fait référence à toutes les variétés orales parlées dans la communication quotidienne. Celles-ci varient d'un pays arabe à l'autre et d'une région d'un même pays à l'autre [5]**.**

1.9.1 Orthographe arabe

Contrairement aux langues latines, l'arabe s'écrit de droite à gauche et se distingue par l'absence de majuscules ou minuscules. Son alphabet comprend 28 lettres : 25 consonnes et seulement 3 voyelles. En plus de ces segments vocaux, l'écriture arabe utilise des signes diacritiques comme voyelles courtes. Ceux-ci sont placés au-dessus ou au-dessous des lettres pour fournir la prononciation correcte et clarifier le sens du mot. La majorité des textes MSA sont écrits sans voyelles courtes. Il en est ainsi parce que les locuteurs compétents n'ont pas besoin de signes diacritiques pour comprendre un texte donné. Cependant, les signes diacritiques sont souvent utilisés dans les livres pour enfants ainsi que dans les livres pour les apprenants en arabe. L'absence de signes diacritiques dans la majorité des textes pose un

problème d'ambiguïté lexicale qui remet en question systèmes. Par Exemple : (شعر) :cheveu (شعر) :poésie(شعر) :signifie

1.9.2 Morphologie arabe

La langue arabe a une morphologie très complexe et riche dans laquelle un mot peut véhiculer des informations importantes. En tant que jeton délimité par un espace, un mot dans L'arabe révèle plusieurs aspects morphologiques : dérivation, flexion et agglutination.

1.9.3 Morphologie dérivationnelle

La morphologie dérivationnelle est le mécanisme de création d'un nouveau mot basé sur un mot existant avec une partie du discours éventuellement différente, par ex. en anglais, l'adjectif "weekly" est dérivé du nom "week".

Comme d'autres langues sémitiques, la morphologie arabe consiste en une représentation racine et motif. Tous Les mots arabes sont basés sur une "racine", qui est une séquence de consonnes contenant le sens de base du mot. Les voyelles et les consonnes non fondamentales sont ajoutées selon des modèles spécifiques pour racine.

Créer une variété de mots liés. Par exemple, le trois lettres "ktb" est une racine qui signifie, (كتب) "Ecrire". S'il est mis dans le modèle "1a2a3a" (kataba,) où les chiffres correspondent aux lettres racines. En ajoutant la voyelle longue ("a:" après la première lettre, nous obtenons un nouveau motif "la:2a3a" "correspond".Qui signifie) (كاتب) (ka:taba), et un nouveau verbe Table 1 listes quelques mots dérivés de la racine "ktb" avec leurs significations [5]**.**

1.9.4 Morphologie flexionnelle

La morphologie flexionnelle définit la variation d'un mot pour décrire le même sens dans différentes catégories grammaticales (par exemple en anglais : écrire, écrit, écrit). L'ensemble de ces formes de mots fléchies est appelé une classe de lexèmes. Pour représenter le lexème, un lemme, qui est une forme particulière, est classiquement sélectionné.

mots dérivés de la racine	**Prononciation**	**schema**	**signification**
كتب	Kataba	1a2a3a	écrire
كاتب	Ka : taba	1a :2a3a	correspondre
مكتب	Maktab	Ma12a3	bureau
كتب	Kutub	1u2u3	livres
كاتب	Ka : tib	1a :2i3	écrivain

Table 1. 1 : Mots dérivés de la racine "ktb"

En arabe, les mots s'infléchissent en sept catégories : temps (passé et présent), personne (1ère, 2ème et 3ème), nombre (singulier, duel et pluriel), genre (féminin et masculin), cas (nominatif, accusatif et génitif), le mode (indicatif, impératif, subjonctif, et énergétique) et la voix (active et passive). Le tableau 2 présente l'inflexion du verbe « ktb » (écrire) en fonction du temps, de la personne, du nombre et du genre.

1.9.5 Morphologie agglutinante

L'arabe est une langue agglutinante, ce qui signifie que le mot peut être attaché à un ensemble de clitiques (affixes).

Ces clitiques sont répartis en 4 classes (cf. tableau 3) et s'appliquent à une base de mots dans un ordre strict :

CONI + PART + DET + BASE + PRON

L'expression française "et avec son travail", par exemple, correspond à la forme arabe ""(وبعمله), Ce mot peut être divisé en quatre

Préfixe	و	Et
préfixe proclitique	ب	Avec
Racine	عمل	travail
le suffixe	ه	son

Table 1. 2 : Comparaison expression française et langue arabe

1.10 Analyse des sentiments en arabe :

L’unicité de la structure des mots arabes est l'une des principales difficultés auxquelles les chercheurs sont confrontés lorsqu'ils traitent de l'analyse des sentiments arabes. La section suivante examine certains des principaux défis auxquels sont confrontés les efforts visant à mettre en place un système précis pour l'arabe.

1.10.1 Analyse morphologique

L'analyse morphologique est une phase importante d’Analyse des sentiments. Son objectif principal est de décomposer les mots en morphèmes et d'associer chaque morphème à une information morphologique telle que radical, racine, POS (Part Of Speech) et affixe. Comme nous l'avons vu dans la section précédente, l'arabe est une langue morphologiquement complexe. Cette complexité nécessite le développement de systèmes appropriés capables de gérer la tokenisation, la vérification orthographique, la radicalisation, la lemmatisation, la

correspondance de modèles et le marquage des parties du discours. De nos jours, de nombreux analyseurs morphologiques pour l'arabe sont déjà développés ; certains d'entre eux sont disponibles gratuitement tandis que les autres ont un but commercial. Parmi celles citées dans la littérature figurent Analyse et génération morphologiques arabes de

1.10.2 Arabe dialectal

À des fins de communication, les arabophones utilisent généralement l'arabe familier plutôt que le MSA. Il existe environ 30 principaux dialectes arabes qui diffèrent du MSA et les uns des autres sur le plan phonologique, morphologique et lexical [5]. De plus, les dialectes arabes n'ont pas d'orthographe standard et pas d'académies de langues Par conséquent, en utilisant des outils et des ressources conçus pour MSA pour traiter les dialectes arabes génère des performances considérablement faibles. Récemment, des chercheurs ont commencé à développer des analyseurs pour des dialectes spécifiques tels que CALIMA [6] pour le dialecte égyptien. Cependant, ces analyseurs ont encore une faible précision et ne sont conçus que pour des dialectes particuliers. Combler cette lacune dans le traitement de l'arabe améliorera l'efficacité de la recherche d'informations, en particulier pour les données des médias sociaux.

1.10.3 Reconnaissance de l'entité désignée

En arabe, de grandes portions de noms arabes sont associées à des correspondants "سعيد""adjectifs positifs. Par exemple, le prénom qui signifie "heureux". De plus, l'arabe" سعيد"à l'adjectif des noms propres ne prend pas de majuscule comme dans les langues latines, ce qui complique l'identification des entités nommées. Pour cette raison, un système de reconnaissance d'entités nommées est Crucial dans l'analyse des textes arabes et la distinction entre les noms d'entités et les mots sentiment.

1.11 Travaux connexes

Malgré les nombreuses difficultés posées par la langue arabe, ces dernières années, de nombreux défis ont été observés dans le domaine de l'analyse des sentiments concernant la langue. Dans cette section, nous listons quelques travaux:

1.11.1 Aux études arabes

1. Afnan Al-Subaihin [7] et autre. Ils ont fourni un outil d'analyse des sentiments pour le texte arabe général utilisé dans les chats quotidiens et les réseaux sociaux, ont utilisé trois des techniques de classification Naïve Bayes SVM et Maximum Entropie, et l'étude a montré que SVM a fourni les meilleurs résultats pour l'évaluation.
2. H.Benbrahim et I.Berrada [8] présentent une étude de la classification supervisée des sentiments en contexte arabe. Ils ont utilisé deux corps arabes qui diffèrent à bien des égards

Ils utilisent donc trois classificateurs populaires connus pour leurs performances, à savoir Naïve Bayes, Support Vector Machines et k-Nearest Neighbor. Ils ont testé certains paramètres pour identifier ceux qui donnaient les meilleurs résultats. Ces paramètres comprennent le type de noyau, le seuil de fréquence des termes, la pondération des termes et les mots de n grammes. Qui montrent que les machines bayésiennes naïves et à vecteurs de support sont efficaces en termes de concurrence ; pourtant k-nn dépend du corpus, leurs résultats montrent que l'efficacité de la classification peut dépendre de la longueur des documents, de l'uniformité des documents et de la nature des auteurs des documents. Cependant, la taille des ensembles de données n'affecte pas les résultats de la classification.

3. Dans les études doctorales. Alaa Fadl al-Hassan, septembre 2016 [9]. Pour créer un lexique pour l'analyse des dialectes palestiniens de Twitter sur les médias sociaux en utilisant une approche d'apprentissage supervisé, a classé les tweets comme positifs, négatifs, il utilisé des techniques pour traiter le langage naturel. RAPID MINER et deux techniques de classement : Naïve Base (NB) et Support Victor Machine (SVM). L'étude a montré que l'utilisation du lexique de polarisation améliore la précision de la classification des sentiments pour le dialecte palestinien, et que la machine à vecteurs de support fournit la meilleure précision pour le modèle de classification des sentiments.

4. Chercheurs Ghada Al-Wakil, Taha Osman, Thomas Hughes [10]. Application de l'analyse des sentiments aux données textuelles du dialecte saoudien sur le chômage en Arabie saoudite Le principal défi auquel les chercheurs sont confrontés dans l'analyse des sentiments en arabe informel, car le dialecte saoudien n'est pas conforme à la grammaire officielle structurelle de l'arabe classique moderne. Les chercheurs ont utilisé le traitement du langage naturel et des techniques d'apprentissage automatique supervisées. Les caractéristiques émotionnelles sont définies, améliorant encore la précision de la classification des émotions de l'accent saoudien.

1.11.2 Aux Études étrangères

1. Les chercheurs Nafissa Yussupova et Diana Bogdanova ont étudié l'analyse des sentiments du texte russe basée sur des méthodes d'apprentissage automatique [11]. Deux chercheurs décrivent le problème de la classification des sentiments dans les messages texte russes, dont l'un est l'utilisation de diverses terminaisons basées sur les préjugés grammaticaux, le temps et le sexe. Un autre problème courant avec la classification des sentiments dans différentes langues est que différents mots peuvent avoir la même signification (synonymes) et peuvent donc se voir attribuer la même

valeur sentimentale. Par conséquent, en comparant les résultats en russe et en anglais, déterminez comment la réduction affecte la précision de la classification des sentiments (ou autre, avec ou sans terminaisons). Pour évaluer l'effet des synonymes, ils ont utilisé une méthode consistant à combiner des mots faisant référence à la même signification en un seul terme. Pour résoudre ces problèmes, ils utilisent la bibliothèque .lemmatisation et des synonymes. Inversement, classer le sentiment d'un texte anglais sans utiliser la lemmatisation donne de meilleurs résultats. L'étude a également révélé que l'utilisation de synonymes dans le modèle avait un effet positif sur la précision.

2. Centre commun de recherche de la Commission européenne [12]. Une méthode d'analyse des sentiments spécifiquement conçue pour traiter les données Twitter (tweets), en tenant compte de leur structure, de leur longueur et de leur langage spécifique, ses principaux apports sont :
    a) Anticiper le traitement des tweets pour normaliser le langage et diffuser pour un vocabulaire expressif de sentiment.
    b) utiliser un traitement linguistique minimal, ce qui rend la méthode facilement transférable à d'autres langues.
    c) inclure des grammes de niveau supérieur pour identifier les changements dans la polarité de l'émotion exprimée.
    d) utiliser des vecteurs de support sur un ensemble de données factuelles Classification linéaire simple (machines à vecteurs de support **SVM**) pour les applications d'apprentissage supervisé. Il a expliqué que l'utilisation du modèle entraîné créé de la manière ci-dessus améliorait les performances de la classification des sentiments.
3. L'analyse des sentiments des films hollywoodiens a été appliquée à Twitter [13]. Le chercheur Umesh Hodeghatta a analysé les tweets de six films hollywoodiens et s'est renseigné sur les sentiments, les émotions et les opinions exprimés dans neuf endroits différents dans quatre pays différents. Le modèle utilise les méthodes d'apprentissage automatique Naïve Bayes et MaxEnt, en utilisant Python et la bibliothèque d'outils de langage naturel, ils prennent en compte les horodatages, les noms d'utilisateur, les emplacements géographiques et les messages de tweet réels. Ceux-ci ont été testés contre deux Unigrams. Bigram Rated MaxEnt Unigram offre la meilleure précision de 84 %.
4. Professeur Shubham, Joy Joseph, Richa Mehra. Ils décrivent différentes techniques utilisées dans l'exploration de texte et l'analyse des sentiments [14]. Il classe également

l'analyse des sentiments au niveau de la phrase et l'analyse des sentiments au niveau du document. Analyse des sentiments sur Twitter à l'aide de l'apprentissage automatique. La recherche utilise une approche de base de connaissances et une approche d'apprentissage automatique pour analyser le sentiment du texte. Les messages Twitter sont numérisés sur des appareils électroniques tels que des téléphones portables, des ordinateurs portables, etc. En analysant le sentiment dans un domaine donné, l'impact des informations de domaine sur la classification des sentiments peut être déterminé. Introduire une nouvelle tendance pour catégoriser les tweets comme positifs et négatifs sur les produits

1.12 Conclusion :

Dans ce premier chapitre, nous avons parlé des notations générales dans le domaine de l'analyse des sentiments, puis nous avons expliqué les niveaux et les types d'analyse des sentiments et la façon dont ils sont appliqués et utilisés. Nous avons enfin montré les problèmes auxquels nous sommes confrontés dans ce domaine.

Pour achever notre étude bibliographique, nous allons présenter, dans le chapitre qui suit, un l'apprentissage en profondeur et classification

CHAPITER 2 :

L'APPRENTISSAGE EN PROFONDEUR

&

CLASSIFICATION

CHAPITER 2 : L'APPRENTISSAGE EN PROFONDEUR & CLASSIFICATION

2.1 Introduction

Nous nous intéressons dans ce chapitre aux travaux relatifs à traiter des tweets arabes pour analyser et classer les sentiments Le traitement avancé des données et la classification des sentiments sont l'une des étapes les plus importantes d'un cadre d'analyse des sentiments, bien que le traitement avancé des données soit un outil puissant pour traiter des données complexes. Il existe de nombreux types de données qui doivent être déclarées. Les données originales doivent être traitées à l'avance, traitées selon un processus normalisé, et les opinions sont extraites des collections des données pertinentes pour découvrir les opinions. Dans ce chapitre, Nous proposons différentes techniques d'apprentissage en profondeur et de classification des sentiments

2.2 L'apprentissage automatique

L'apprentissage automatique est une classe d'algorithmes qui permettent aux applications logicielles de prédire plus précisément les résultats sans être explicitement programmées. Le principe de base de l'apprentissage automatique est de créer des algorithmes capables de prendre des données d'entrée et d'utiliser une analyse statistique pour prédire la sortie, tout en les mettant à jour à mesure que de nouvelles données deviennent disponibles. Il existe deux principaux types d'apprentissages.

2.3 Types d'apprentissage automatique

Les algorithmes d'apprentissage automatique sont divisés en deux parties principales:

2.3.1 Apprentissage supervisé

L'apprentissage est dit supervisé lorsque les données qui entrent dans le processus sont déjà catégorisées et que les algorithmes doivent s'en servir pour prédire un résultat en vue de pouvoir le faire plus tard lorsque les données ne seront plus Catégorisées.

2.3.2 Apprentissage non supervisé

Apprentissage non supervisé Complexe car le système détectera ici les similitudes dans les données , les recevoir et les organiser [15].

Ce type contient également deux tâches principales : clustering et association.

2.4 Etapes de prétraitaient

La préparation du texte implique le nettoyage des données extraites avant que l'analyse ne soit effectuée. Habituellement, la préparation de texte implique l'identification et l'élimination du contenu non textuel de l'ensemble de données textuelles. En outre, tout autre contenu qui n'est pas jugé pertinent pour le domaine d'étude est également supprimé de l'ensemble de données textuelles, comme par exemple des mots vides ou des mots qui ne sont pas pertinents pour le cours de l'analyse. Pour un système qui donne SA des flux de données, la stratégie de prétraitement est la suivante :

2.4.1 Tokenisation

La tokenisation est le processus qui consiste à décomposer un texte donné en unités appelées tokens. Les tokens peuvent être des mots individuels, des phrases ou même des phrases entières. Au cours du processus de tokenisation, certains caractères comme les signes de ponctuation peuvent être éliminés. Les tokens deviennent généralement l'entrée pour les processus tels que l'analyse syntaxique et l'exploration de texte.

Presque toutes les tâches de traitement du langage naturel utilisent une technique de tokenisation. Ces tokens sont très utiles pour trouver de tels modèles et sont considérés comme une étape de base pour le déracinement et la lemmatisation.

L'étymologie et la lemmatisation génèrent toutes deux la forme de la racine des mots fléchis obtenus par tokenisation [16].

2.4.2 Normalisation

La tâche de normalisation est importante afin de produire des formes de mots cohérentes. La normalisation du texte arabe comprend les étapes suivantes [17] :

- Dépouillement des diacritiques : par exemple, " يّة بِ زَ عَ لْ ا " en " العزبية ".
- Allongement par effeuillage : par exemple, " العزبيـــة " à " الععة ". " العزبية "
- Suppression de "ال" au début des mots :
- Par exemple, "العزبية" sera "عزبية".
- Remplacer la lettre "ة" par "ه".
- Remplacer la lettre "ى" par "ي".
- Remplacer les lettres "آ-إ-أ" par "ا".
- Normalisation des lettres répétées : par exemple, "سعاااادة " en "سعادة".

2.4.3 Lemmatisation et Stemming

Lemmatisation : L'une des étapes les plus difficiles, elle nous montre que l'arabe est une langue inflexible, cela est dû aux difficultés que contient ce langage, qui ont été mentionnées dans le premier chapitre. Il s'agit de trouver la racine du mot.

Le stemming : est une méthode de normalisation des mots dans le traitement automatique du langage naturel. C'est une technique dans laquelle un ensemble de mots dans une phrase sont convertis en une séquence pour raccourcir sa recherche. Dans cette méthode, les mots ayant le même sens mais présentant quelques variations selon le contexte ou la phrase sont normalisés.

2.4.4 Représentation vectoriel

Afin d'exploiter des algorithmes d'apprentissage automatique sur nos données textuelles, il nous faut représenter le texte de nos documents sous la forme d'un vecteur de taille fixe, ceci afin de plonger la donnée dans un espace métrique. Nous utiliserons comme méthode de vectorisation, la matrice « Document x Termes » avec une pondération TF-IDF.

Pour ce faire, le corpus est représenté sous la forme d'une matrice contenant les documents en ligne et les lemmes en colonne. Chaque document correspond à un vecteur où la composante associée à un lemme vaut 0 si le lemme est absent du document, et une valeur pondérée selon l'importance du lemme, si le lemme est présent.

La pondération appliquée à cette matrice est la pondération TF-IDF. Il s'agit de la combinaison du « Terme Fréquence » et de l'« Inverse Document Fréquence », c'est à dire la fréquence du terme dans un document par l'inverse de la fréquence du terme dans le corpus

Cette pondération conduit à considérer les mots fréquents à la fois dans un document et dans le corpus en entier avec une importance diminuée par rapport aux mots fréquents dans un document

Et rare dans le corpus. Un terme rare améliorera ainsi la pertinence lexicale. Avec cette représentation, deux documents seront proches s'ils ont de nombreux termes en commun et deux termes seront proches s'ils sont présents ensembles dans de nombreux documents. La représentation vectorielle effectue les deux opérations ci-dessous :

1) Au cours de la phase, il trouve l'ensemble de mots dans tous les documents, puis compte les occurrences de ces mots dans chaque document. L'ensemble des mots constitue le « vocabulaire », dont la taille est paramétrable selon la fréquence d'apparition minimum d'un terme dans le corpus, le nombre minimum de documents dans lequel un terme doit apparaître ou le nombre maximal de mots retenus.

2) Au cours de la phase, les occurrences d'un mot du « vocabulaire » sont comptabilisées pour chaque ligne du Data Frame en entrée et on obtient en sortie un vecteur creux pour chaque document, qui contient la taille du vocabulaire, l'index du mot dans le vocabulaire, puis le nombre d'occurrence dans le document pour ce mot.

La deuxième étape, pour obtenir notre représentation vectorielle, consiste à calculer la fréquence inverse du document pour chaque terme du corpus en créant une nouvelle instance « IDF » et en appelant la méthode « fit » sur les vecteurs « TF » obtenus précédemment. Nous transformons ensuite les vecteurs « TF » en vecteurs « TF-IDF » grâce à la fonction de transformation d' « IDF ».

On obtient en sortie, pour chaque document, un vecteur creux qui contient la taille du vocabulaire, l'index du lemme dans le vocabulaire, puis le poids TF-IDF pour ce lemme.

2.4.5 Division

Dans cette étape, la représentation obtenue est fournie à l'algorithme de classification ML (Machine Learning) pour construire et apprendre un modèle de classification à partir de tweets étiquetés d'entraînement qui peut prédire l'étiquette de sentiment de nouveaux tweets non étiquetés.

2.5 Approches de classification des sentiments

Les approches de classification des sentiments peuvent être généralement divisées en approche d'apprentissage automatique (Machine Learning), approche basée sur le lexique et approche hybride. L'approche Machine Learning (ML) applique les célèbres algorithmes ML et utilise des fonctionnalités linguistiques qui peuvent être généralement divisées en plusieurs parties, à savoir les méthodes d'apprentissage supervisé et non supervisé. Les méthodes d'apprentissage supervisé font appel à un grand nombre de documents de formation labellisés. Dans le cas où il est difficile de trouver les documents de formation étiquetés, les méthodes non supervisées sont utilisées.

L'approche basée sur le lexique repose sur un lexique des sentiments, une collection de termes de sentiments connus et précompilés. Il est divisé en une approche basée sur un dictionnaire et une approche basée sur un corpus qui utilisent des méthodes statistiques ou sémantiques pour trouver la polarité des sentiments. L'approche basée sur un dictionnaire qui dépend de la recherche de mots de semences d'opinion, puis recherche le dictionnaire de leurs synonymes et antonymes [18]

L'approche basée sur le corpus commence par une liste de départ de mots d'opinion, puis trouve d'autres mots d'opinion dans un grand corpus pour aider à trouver des mots d'opinion

avec des orientations spécifiques au contexte. Les différentes techniques et les algorithmes les plus populaires de classification des sentiments sont illustrées sur la figure suivante :

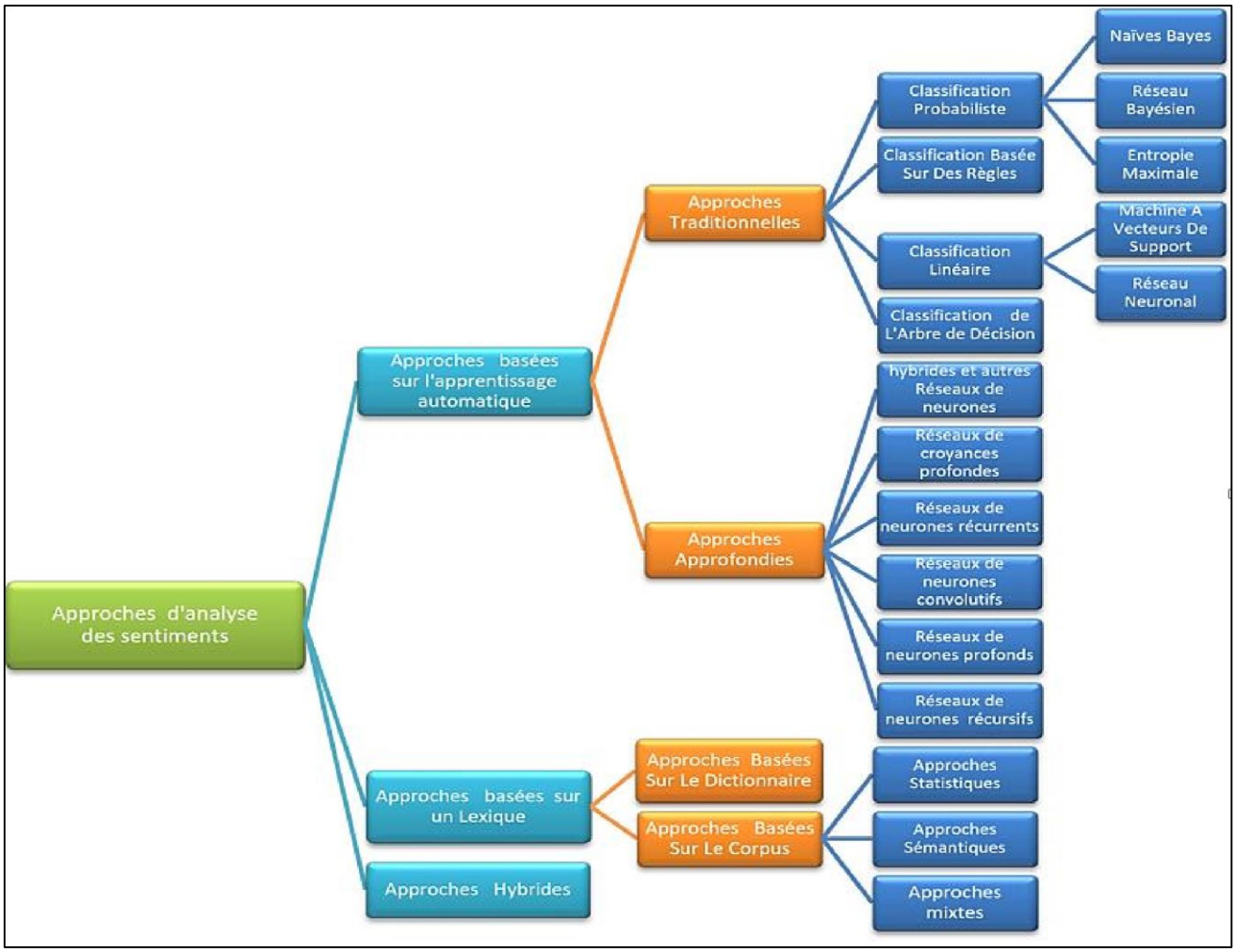


Figure 2. 1*:* Approches d'analyse des sentiments

2.5.1 L'approche basée sur l'apprentissage automatique

L'approche d'apprentissage automatique consiste à donner aux ordinateurs la capacité d'agir sans être programmés. Les programmes informatiques utilisent les données exposées pour détecter des modèles, puis ajuster les actions du programme et prendre des décisions intelligentes. Dans la classification des sentiments, cette approche repose sur l'utilisation de célèbres techniques d'apprentissage automatique sur le texte. La plupart des recherches sur l'analyse des sentiments des tweets arabes ont utilisé des approches ML parce qu'il a été rapporté qu'elles sont plus précises que les approches basées sur les lexiques. [19]

2.5.1.1 L'apprentissage supervisé

Les méthodes d'apprentissage supervisé dépendent de l'existence de documents de formation labellisés. L'apprentissage supervisé est largement utilisé pour construire un système d'Analyse des Sentiments et il peut être catégorisé en deux types : les Méthodes probabilistes et les Méthodes Non- probabilistes. Dans cette section, nous présenterons certaines des méthodes utilisées avec quelques références comme exemples.

2.5.1.2 Méthodes Non-Probabilistes

a. Machines à vecteurs de support

Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais *support-vector machine*, SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination

Les SVM peuvent être utilisés pour résoudre des problèmes de discrimination, c'est-à-dire décider à quelle classe appartient un échantillon, ou de régression, c'est-à-dire prédire la valeur numérique d'une variable. La résolution de ces deux problèmes passe par la construction d'une fonction H qui a un vecteur d'entrée X fait correspondre une sortie **y** :

Y= h(x)

Le cas simple est le cas d'une fonction discriminante linéaire, obtenue par combinaison linéaire du vecteur d'entrée $x=(x_1,\ldots.. xn)^t$, avec un vecteur de poids $w= (w_1,\ldots.. wn)^t$:

$$\mathbf{H(x)=w^t x+w_0}$$

Il est alors décidé que X est de classe 1 si **h(x)>= 0** et de classe -1 sinon. C'est un classifier linéaire.

La frontière de décision **h(x)= 0** est un hyperplan, appelé *hyperplan séparateur*, ou *séparatrice*. Le but d'un algorithme d'apprentissage supervisé est d'apprendre la fonction **h(x)** par le biais d'un ensemble d'apprentissage :

$$\mathbf{\{(x_1,l_1)(\ x_2,l_2),\ldots.,(\ x_k,l_k),\ldots.,(\ x_p,l_p)\}}$$

Où les l_k sont les labels, P est la taille de l'ensemble d'apprentissage, N la dimension des vecteurs d'entrée. Si le problème est linéairement séparable, on doit alors avoir :

$l_k h(x_k) >= 0$ 1=<k=<p autrement dit $\mathbf{(w^t x_k+w_0)>=0}$ 1=<k=<p

Exemple

Imaginons un plan (espace à deux dimensions) dans lequel sont répartis deux groupes de points. Ces points sont associés à un groupe : les points (+) pour $y > x$ et les points (-) pour $y < x$. On peut trouver un séparateur linéaire évident dans cet exemple, la droite d'équation $y = x$. Le problème est dit linéairement séparable.

Pour des problèmes plus compliqués, il n'existe en général pas de séparateur linéaire. Imaginons par exemple un plan dans lequel les points (-) sont regroupés à l'intérieur d'un cercle, avec des points (+) tout autour : aucun séparateur linéaire ne peut correctement séparer les groupes : le problème n'est pas linéairement séparable. Il n'existe pas d'hyperplan séparateur [20].

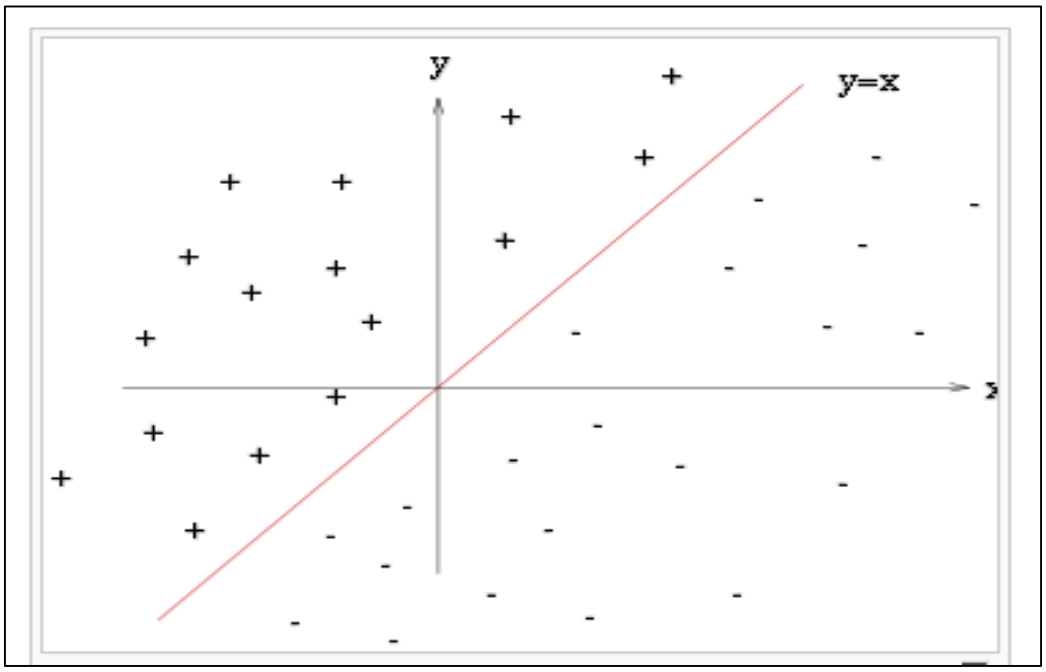


Figure 2. 2 Présentation de problème de discrimination, avec un séparateur linéaire

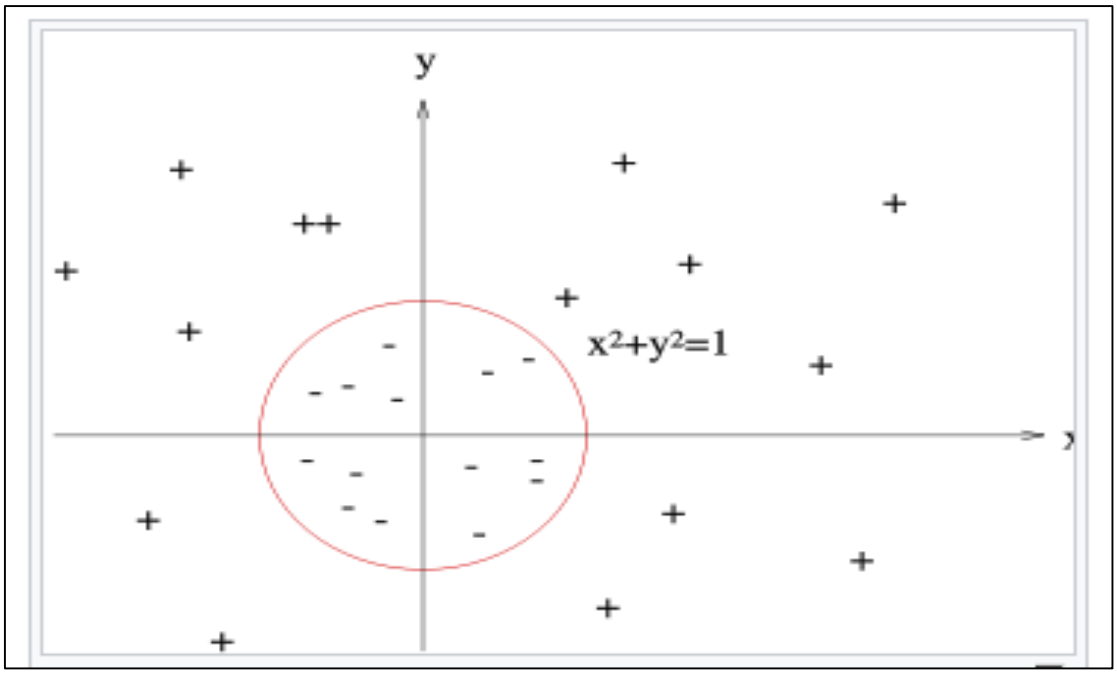


Figure 2. 3 Présentation de problème de discrimination, avec un séparateur non-linéaire

b. Marge Maximale

On se place désormais dans le cas où le problème est linéairement séparable. Même dans ce cas simple, le choix de l'hyperplan séparateur n'est pas évident. Il existe en effet une infinité d'hyperplans séparateurs, dont les performances en apprentissage sont identiques (le risque empirique est le même), mais dont les performances en généralisation peuvent être très différentes. Pour résoudre ce problème, il a été montré, qu'il existe un unique hyperplan optimal, défini comme l'hyperplan qui maximise la marge entre les échantillons et l'hyperplan séparateur.

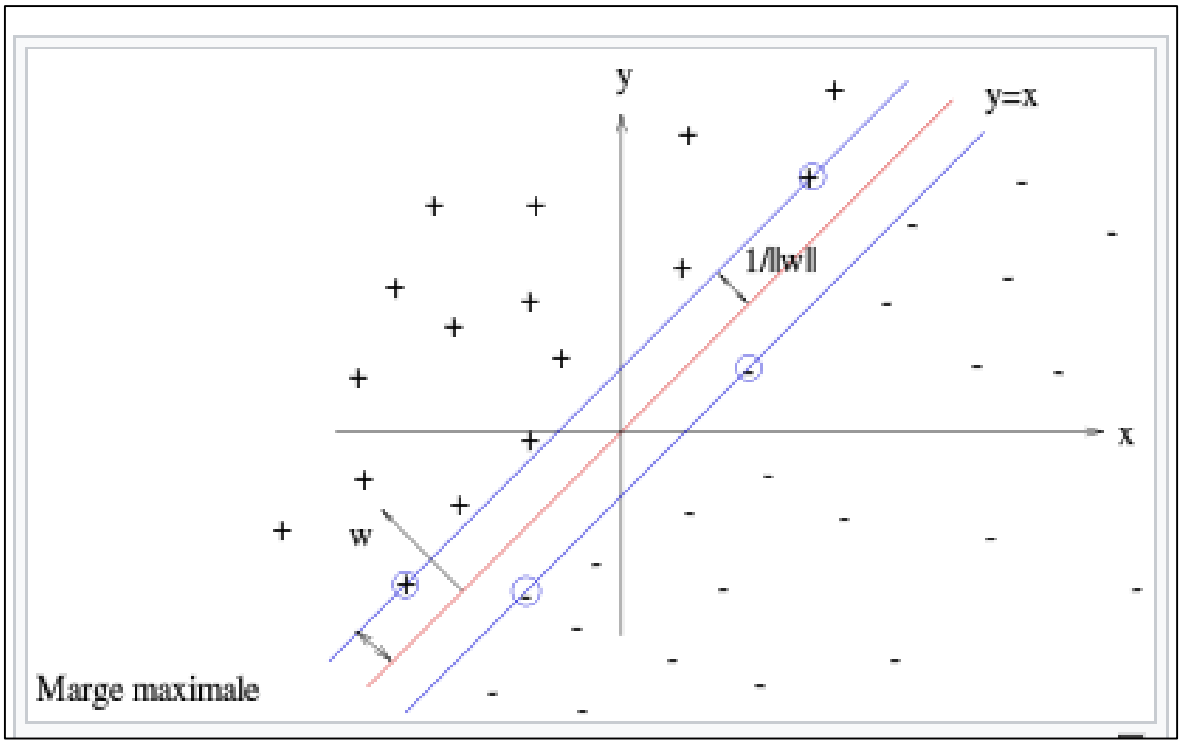


Figure 2. 4 L'hyperplan optimal (en rouge) avec la marge maximale

La marge est la distance entre l'hyperplan et les échantillons les plus proches. Ces derniers sont appelés vecteurs supports. L'hyperplan qui maximise la marge est donné par :

$$\arg\max_{w,w_0} \min_{k} \{||x - x_k|| : x \in \mathbb{R}^N, w^T x + w_0 = 0\} \quad (1)$$

Il s'agit donc de trouver w et w_0 remplissant ces conditions, afin de déterminer l'équation de l'hyperplan séparateur :

$$h(x) = w^T x + w_0 = 0 \quad (2)$$

Avantages de SVM

- Capacité à traiter de grandes dimensionnalités
- Traitement des problèmes non linéaires avec le choix des noyaux
- points supports donne une bonne indication de la complexité du problème traité

Inconvénients de SVM

- Problème lorsque les classes sont bruitées
- Le traitement des problèmes multi-classes reste une question ouverte
- Problème lorsque les classes sont bruitées

c. k-Nearest Neighbors

k-NN voire KNN ou méthode des k plus proches voisins k-NN est un algorithme standard de classification qui repose exclusivement sur le choix de la métrique de classification. Il est « non paramétrique » (seul k doit être fixé) et se base uniquement sur les données d'entraînement. [20]

➢ **Quand utilisons-nous l'algorithme KNN ?**

KNN peut être utilisé pour les problèmes prédictifs de classification et de régression. Cependant, il est plus largement utilisé dans les problèmes de classification dans l'industrie. Pour évaluer une technique, nous examinons généralement 3 aspects importants :

- Sortie facile à interpréter
- Temps de calcul
- Pouvoir prédictif

➢ **Comment fonctionne l'algorithme KNN ?**

Prenons un cas simple pour comprendre cet algorithme. Voici une répartition des cercles rouges (RC) et des carrés verts (GS) :

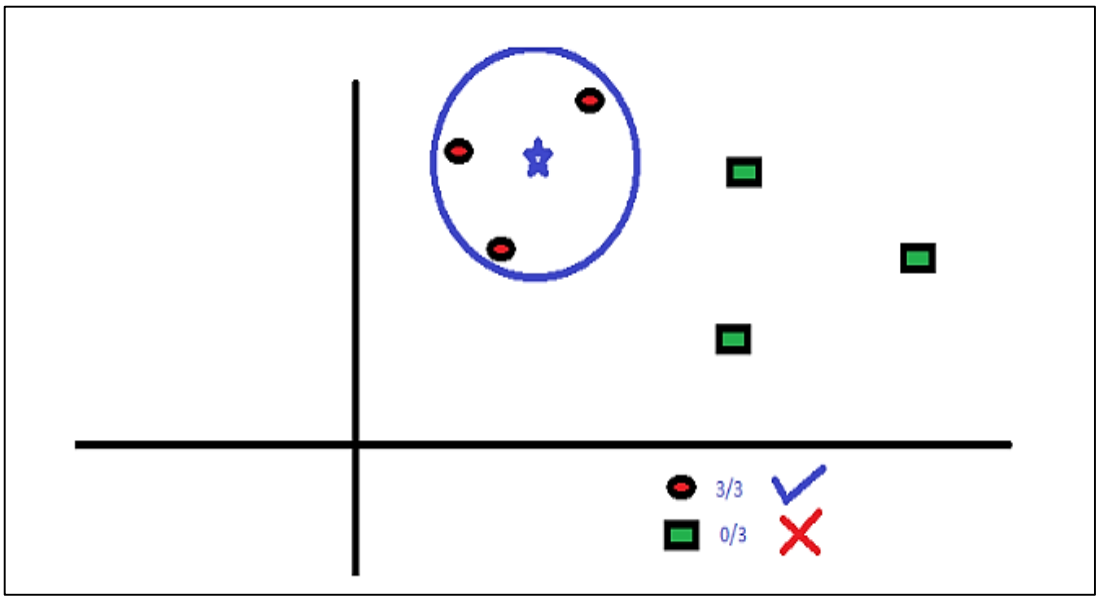


Figure 2. 5 Schéma représenté les cercles rouges (RC) et les carrés verts (GS)

Vous avez l'intention de découvrir la classe de l'étoile bleue (BS). BS peut être RC ou GS et rien d'autre. L'algorithme "K" est KNN est le plus proche voisin duquel nous souhaitons voter. Disons que K = 3. Par conséquent, nous allons maintenant faire un cercle avec BS comme centre

aussi grand que pour enfermer seulement trois points de données sur le plan. Reportez-vous au schéma suivant pour plus de détails :

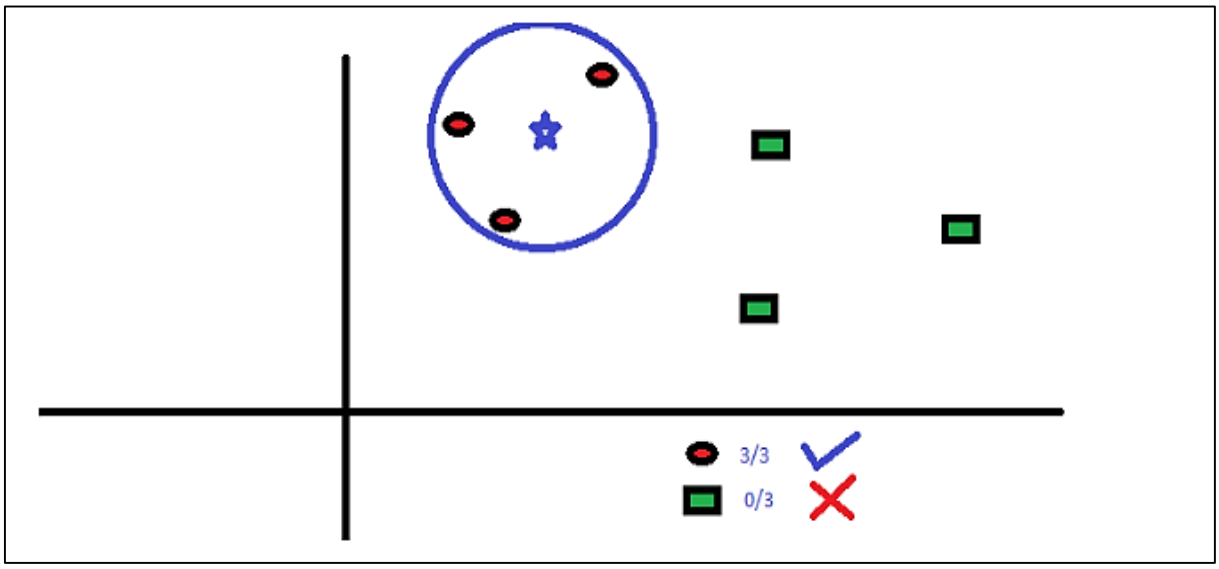


Figure 2. 6 Schéma représenté un cercle avec BS

Les trois points les plus proches de BS sont tous RC. Ainsi, avec un bon niveau de confiance, on peut dire que la BS devrait appartenir à la classe RC. Ici, le choix est devenu très évident puisque les trois votes du voisin le plus proche sont allés à RC. Le choix du paramètre K est très crucial dans cet algorithme. Ensuite, nous comprendrons quels sont les facteurs à considérer pour conclure le meilleur K.

- **Comment choisit-on le facteur K ?**

Essayons d'abord de comprendre quelle est exactement l'influence de K dans l'algorithme. Si nous voyons le dernier exemple, étant donné que toutes les 6 observations d'entraînement restent constantes, avec une valeur K donnée, nous pouvons faire des frontières de chaque classe. Ces limites sépareront RC de GS. De la même manière, essayons de voir l'effet de la valeur "K" sur les frontières de classe. Voici les différentes frontières séparant les deux classes avec différentes valeurs de K.

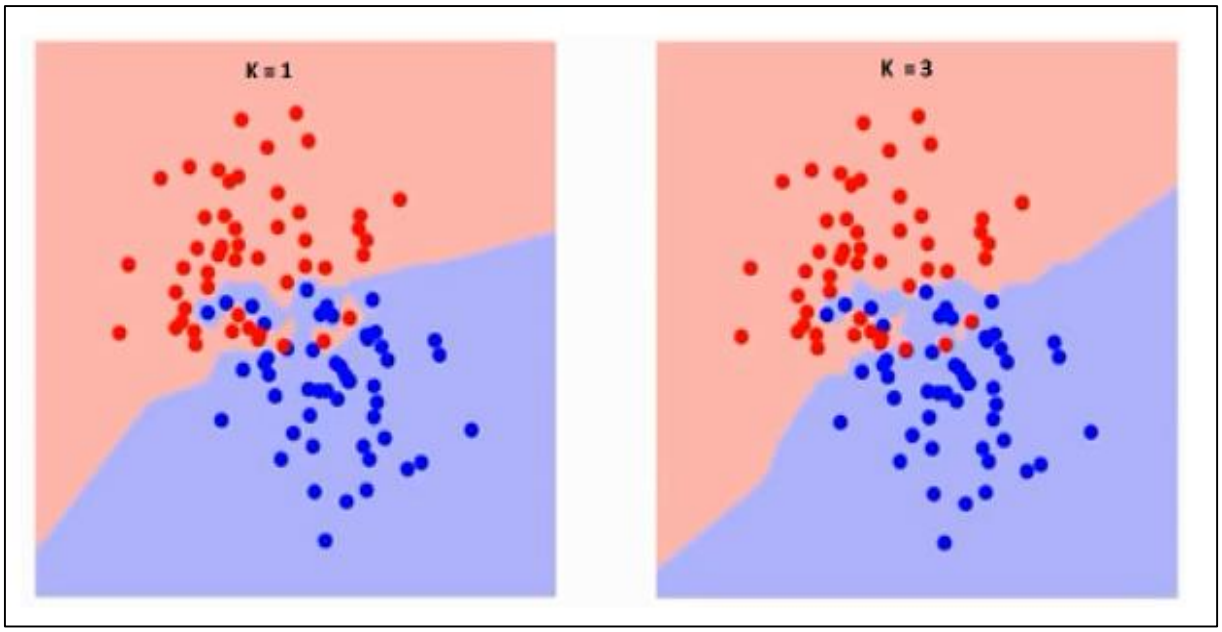


Figure 2. 7 Les différentes frontières séparant les deux classes

Si vous regardez attentivement, vous pouvez voir que la frontière devient plus lisse avec l'augmentation de la valeur de K. Avec K augmentant à l'infini, il devient finalement tout bleu ou tout rouge selon la majorité totale. Le taux d'erreur de formation et le taux d'erreur de validation sont deux paramètres dont nous avons besoin pour accéder à différentes valeurs de K. Voici la courbe du taux d'erreur d'apprentissage avec une valeur variable de K :

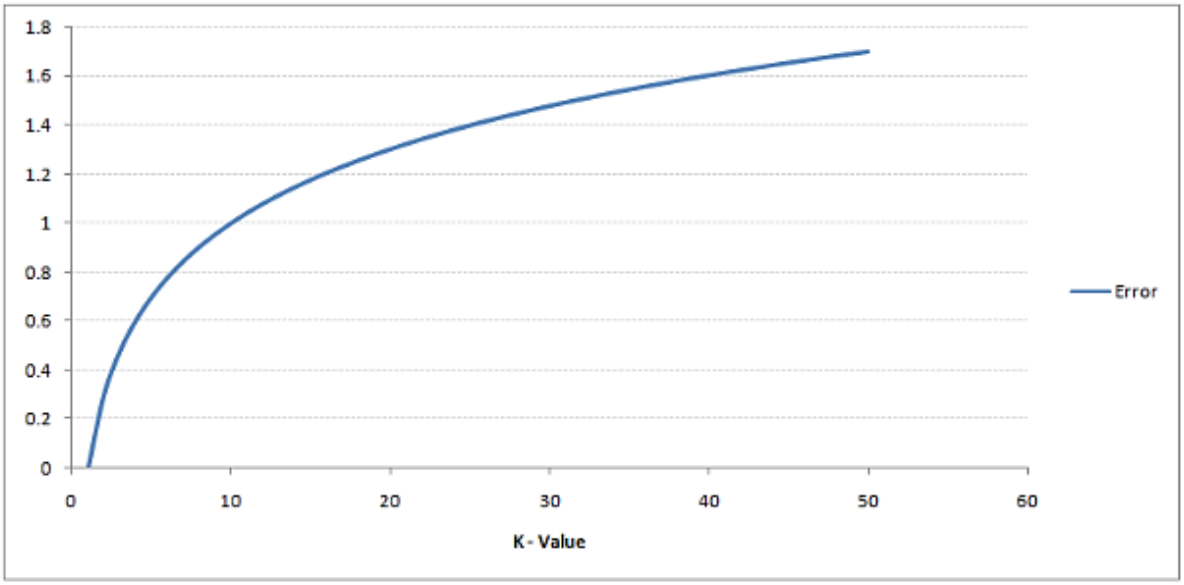


Figure 2. 8 *Taux d'erreur d'apprentissage avec une valeur variable de K*

Comme vous pouvez le voir, le taux d'erreur à K=1 est toujours égal à zéro pour l'échantillon d'apprentissage. En effet, le point le plus proche de tout point de données d'apprentissage est

lui-même. Par conséquent, la prédiction est toujours précise avec K = 1. Si la courbe d'erreur de validation aurait été similaire, notre choix de K aurait été 1. Voici la courbe d'erreur de validation avec une valeur variable de K :

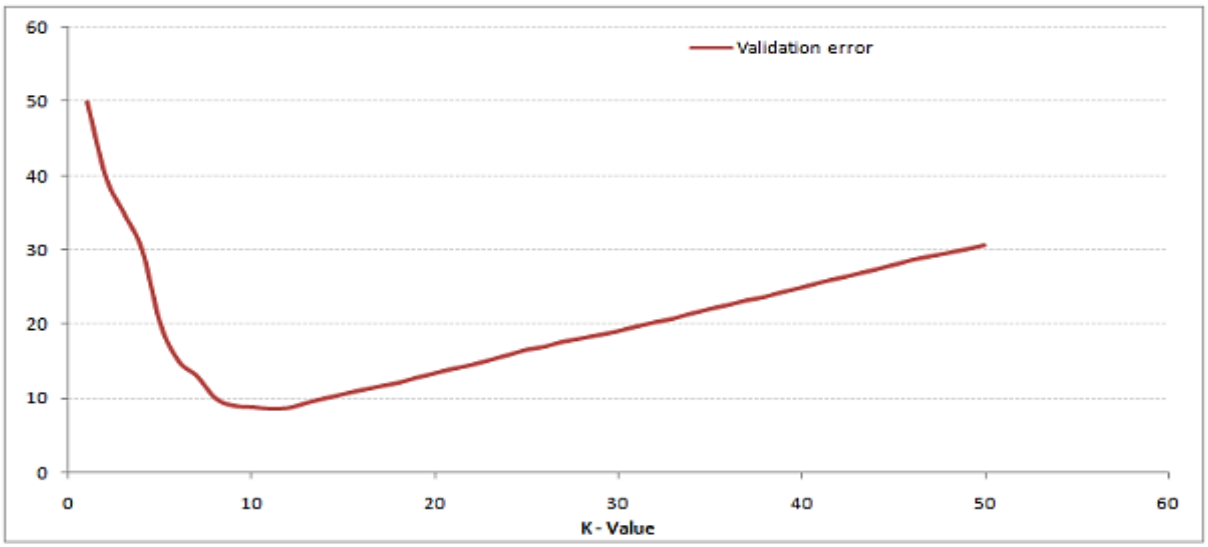


Figure 2. 9 L'erreur de validation avec une valeur variable de K

Cela rend l'histoire plus claire. À K = 1, nous surajoutions les limites. Par conséquent, le taux d'erreur diminue initialement et atteint un minimum. Après le point minimum, il augmente ensuite avec l'augmentation de K. Pour obtenir la valeur optimale de K, vous pouvez séparer la formation et la validation de l'ensemble de données initial. Tracez maintenant la courbe d'erreur de validation pour obtenir la valeur optimale de K. Cette valeur de K doit être utilisée pour toutes les prédictions.

Avantages de K-NN

- Entraînement très rapide.
- Simple et facile à comprendre.
- La méthode des k plus proches voisins n'utilise pas de modèle pour classifier les documents.

Inconvénients de K-NN

- le temps d'exécution qu'elle met pour la classification d'un nouveau cas, car il faut calculer chaque fois la similarité entre les k exemples et le nouveau k, avant de décider quelle classe à choisir.

- Haute complexité de calcul.
- la grande capacité de stockage qu'elle nécessite pour le traitement des Corpus.

2.5.1.3 . Méthodes Probabilistes

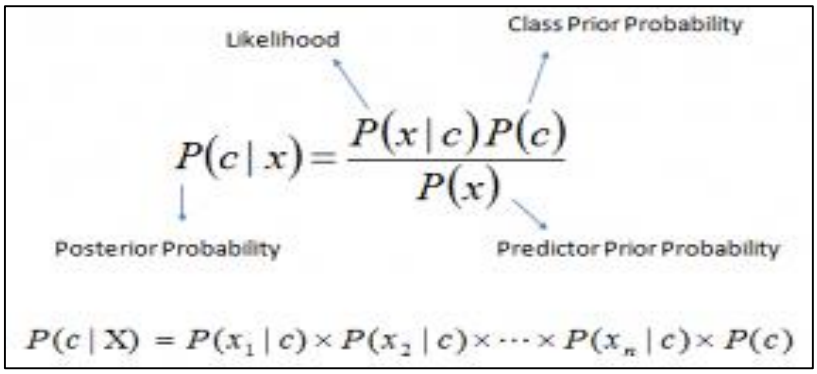
Les classificateurs probabilistes utilisent des modèles de mélange pour la classification. Le modèle de mélange suppose que chaque classe est un composant du mélange. Chaque composant du mélange est un modèle génératif qui fournit la probabilité d'échantillonnage d'un terme particulier pour ce composant. Ces types de classificateurs sont également appelés classificateurs génératifs. Trois des classificateurs probabilistes les plus célèbres sont discutés dans les suivantes sous-sections.

a. Classificateur Bayes naïfs (NB)

Les classificateurs bayésiens sont les classificateurs les plus simples en apprentissage supervisé basés sur le théorème de Bayes. Ils peuvent prédire la classe probabilités d'appartenance, telles que la probabilité qu'un échantillon donné appartient à une classe particulière. Les classificateurs supposent que l'effet d'une valeur d'attribut sur une classe donnée est indépendant des valeurs des autres attributs. Cette hypothèse est appelée indépendance conditionnelle de classe. Il est fait pour simplifier le calcul impliqué et, en ce sens, est considéré comme « naïf ».

Le modèle Naïve Bayes est facile à construire et particulièrement utile pour les très grands ensembles de données. En plus de sa simplicité, Naïve Bayes est connu pour surpasser même les méthodes de classification les plus sophistiquées**.** [21]

Le théorème de Bayes fournit un moyen de calculer la probabilité a posteriori P (c|x) à partir de P(c), P(x) et P (x|c). Regardez l'équation ci-dessous :



(3)

Au-dessus :

- P(c|x) est la probabilité a posteriori de la classe (c, cible) compte tenu du prédicteur (x, attributs).
- P(c) est la probabilité a priori de la classe.
- P(x|c) est la vraisemblance qui est la probabilité du prédicteur pour une classe donnée.
- P(x) est la probabilité a priori du prédicteur.

En langage clair, en utilisant la terminologie de probabilité bayésienne , l'équation ci-dessus peut être écrite comme

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \quad \textbf{(4)}$$

En pratique, on n'a d'intérêt que pour le numérateur de cette fraction, car le dénominateur ne dépend pas de **C** et les valeurs des caractéristiques $\mathbf{X_i}$ sont données, de sorte que le dénominateur est effectivement constant. Le numérateur est équivalent au modèle de probabilité conjointe

Qui peut être réécrite comme suit, en utilisant la règle de la chaîne pour les applications répétées de la définition de la probabilité conditionnelle :

$$\begin{aligned} p(C_k, x_1, \ldots, x_n) &= p(x_1, \ldots, x_n, C_k) \\ &= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2, \ldots, x_n, C_k) \\ &= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2 \mid x_3, \ldots, x_n, C_k)\, p(x_3, \ldots, x_n, C_k) \\ &= \cdots \\ &= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2 \mid x_3, \ldots, x_n, C_k) \cdots p(x_{n-1} \mid x_n, C_k)\, p(x_n \mid C_k)\, p(C_k) \end{aligned}$$

(5)

Maintenant, les hypothèses d'indépendance conditionnelle "naïves" entrent en jeu : supposons que toutes les caractéristiques de **X** sont mutuellement indépendantes , sous réserve de la catégorie. $\mathbf{C_k}$ Sous cette hypothèse,

$$p(x_i \mid x_{i+1}, \ldots, x_n, C_k) = p(x_i \mid C_k)\,. \quad \textbf{(6)}$$

Ainsi, le modèle conjoint peut être exprimé comme

$$\begin{aligned} p(C_k \mid x_1, \ldots, x_n) &\propto p(C_k, x_1, \ldots, x_n) \\ &\propto p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k)\, \cdots \\ &\propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,, \end{aligned} \quad \textbf{(7)}$$

Où **α** désigne la proportionnalité .

Cela signifie que sous les hypothèses d'indépendance ci-dessus, la distribution conditionnelle sur la variable de classe **C** est:

$$p(C_k \mid x_1, \ldots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \quad \textbf{(8)}$$

Où la preuve

$$Z = p(\mathbf{x}) = \sum_{k} p(C_k)\, p(\mathbf{x} \mid C_k) \quad \textbf{(9)}$$

Est un facteur d'échelle dépendant uniquement de x_1,…, xn c'est-à-dire une constante si les valeurs des variables de caractéristique sont connues.

Avantages Naïve Bayes

- Rapide dans la tâche d'entrainement et de classification.
- La facilité et la simplicité de leur implémentation.

Inconvénients Naïve Bayes

- Moins précis que SVM
- ces performances sont limitées quand il s'agit d'une grande quantité de lexiques à traiter

b. Classificateur réseau Bayésien (RB)

Les réseaux Bayésiens constituent un ensemble de méthodes statistiques utilisées pour modéliser des problèmes, extraire de l'information et prendre des décisions. Ils sont un formalisme de raisonnement probabiliste utilisé dans plusieurs domaines tels que l'industrie, la santé, finance et le traitement d'images. L'hypothèse principale du classificateur réseau

Bayésien est l'indépendance des caractéristiques. L'autre hypothèse extrême est de supposer que toutes les fonctionnalités sont entièrement dépendantes. Cela conduit au modèle de réseau bayésien qui est un graphe acyclique dirigé dont les nœuds représentent des variables aléatoires et les arêtes représentent des dépendances conditionnelles. Les réseaux bayésiens ont été largement utilisés dans de nombreuses applications de fouille de texte, comme le filtrage du spam et la récupération d'informations

c. Classificateur d'Entropie Maximale (EM)

Le classificateur d'entropie maximale (appelé classificateur exponentiel conditionnel) convertit les ensembles d'entités étiquetés en vecteurs à l'aide du codage. Ce vecteur codé est ensuite utilisé pour calculer les poids de chaque entité qui peuvent ensuite être combinés pour déterminer l'étiquette la plus probable pour un ensemble d'entités. L'entropie maximale maximise l'entropie définie dans la distribution de probabilité conditionnelle. Il traite de la même manière décrite dans l'algorithme naïf de Bayes.

2.5.2 L'approche basée sur le lexique

L'approche basée sur le lexique dépend du lexique qui contient une collection de mots de sentiment, chaque mot a une valeur de polarité, les mots positifs ont des valeurs supérieures à zéro, les mots négatifs ont des valeurs inférieures à zéro et tout mot qui n'existe pas dans le lexique est pris en compte comme un mot neutre

La tâche de classification des sentiments peut être effectuée sur la base de cette approche en recherchant des mots de sentiment dans un texte ou un document donné, puis en ajoutant des poids ou des balises à ces mots, après en comptant les poids et les balises pour détecter le sentiment général. Une liste de mots de sentiment avec sa valeur de polarité existe dans un lexique de sentiment [18].

Et pour préparer le lexique de sentiment, il existe deux approches : l'approche basée sur le dictionnaire et l'approche basée sur le corpus.

2.5.3 L'approche Hybrides

- Dans cette section, les techniques de classification des sentiments seront abordées, dans lesquelles les auteurs ont utilisé plus d'une technique d'apprentissage automatique et technique basée sur un lexique, connectés les uns aux autres. Cette technique combinatoire pour effectuer la classification est appelée approche hybride. Quelques-uns d'entre eux sont mis en évidence comme ci-dessous:
- Nandi et Agrawal ont proposé une classification hybride des sentiments combinant l'approche du dictionnaire de lexiques avec le résultat du classificateur SVM.

L'approche lexicale est basée sur un dictionnaire de mots, c'est-à-dire un sac de mots à analyser, et fonctionne sur le principe que la polarité d'un document est la somme de la polarité des mots ou des phrases individuelles. Ils ont considéré les tweets Twitter pour la classification. Ils ont collecté des tweets liés à la politique indienne pour les classer [22].

- Desai et Mehta ont proposé un algorithme de classification hybride pour analyser Problèmes et avantages pour les étudiants. Ils combinent des méthodes basées sur la connaissance et l'apprentissage automatique pour traiter les tweets. Ils ont collecté des tweets avec #engineringProblem et le hashtag #engineeringPerks comme ensemble de données pour analyse. Pour effectuer une analyse basée sur les connaissances, un corpus de tous les tweets collectés a été créé. Ensuite, trouvez le dictionnaire avec des valeurs d'opinion plus élevées pour les polarités positives et négatives. Ces mots sont connus comme les mots semences, et tous les synonymes et antonymes possibles pour ces mots semences sont rassemblés pour former un dictionnaire de lexiques. Ce dictionnaire de lexiques est alimenté par l'approche d'apprentissage automatique qui le considère comme une entrée et en fonction de ces entrées, les tweets sont classés. [23].

2.6 Conclusion

Enfin Dans ce chapitre nous avons présenté les deux approches (apprentissage automatique et apprentissage en profondeur) avec les différentes étapes à suivre les méthodes classification que nous avons utilisées pour avoir une bonne prédiction.

CHAPITRE 3

REALISATION

&

EXPERIMENTATION

CHAPITRE 3 : REALISATION & EXPERIMENTATION

3.1 Introduction

Dans ce chapitre, nous commençons de créer notre modèle d'analyse d'opinion. L'objectif principal de la recherche est de concevoir un modèle d'analyse des sentiments pour le texte arabe à l'aide du dictionnaire d'analyse des sentiments, ainsi que d'appliquer des algorithmes d'apprentissage automatique pour analyser les textes arabes en limitant la classification de ces sentiments à positifs, négatifs ou neutre, selon le modèle de classification (deux ou trois).

3.2 Les Outils et Environnement de programmation

3.2.1 Python

Python est un langage de programmation puissant et facile à apprendre. Il dispose de structures de données de haut niveau et permet une approche simple mais efficace de la programmation orientée objet. Parce que sa syntaxe est élégante, que son typage est dynamique et qu'il est interprété, Python est un langage idéal pour l'écriture de scripts et le développement rapide d'applications dans de nombreux domaines et sur la plupart des plateformes [24].

Figure 3. 1 : Logo de python

3.2.2 Anaconda

Anaconda est la plateforme la plus populaire parmi les professionnels de la science des données pour exécuter des implémentations Python et R. Il existe plus de 300 bibliothèques dans le domaine de la science des données. Disposer d'un système de distribution robuste est donc indispensable pour tout professionnel de ce secteur [25].

Figure 3. 2 : Logo d'Anaconda

3.2.3 Jupyter

Jupyter Notebooks est un projet dérivé du projet IPython, qui avait lui-même un projet IPython Notebook. Le nom, Jupyter, vient des principaux langages de programmation qu'il supporte : Julia, Python et R. Jupyter est livré avec le noyau IPython, qui vous permet d'écrire vos programmes en Python, mais il existe actuellement plus de 100 autres noyaux que vous pouvez également utiliser [26].

Figure 3. 3 : Logo de Jupyter.

3.3 Source de données

Un ensemble de mots qu'un classificateur peut utiliser pour évaluer la polarité d'un texte. Le processus de collecte des données nous a pris beaucoup de temps du fait de la difficulté de la langue arabe et de la diversité de son vocabulaire et de ses dialectes. Au final, nous avions deux fichiers, le premier contenant des mots arabes positifs et le second contenant des mots arabes négatifs pour le processus de classement.

- **Première fichier :**

Cette figure présenté le fichier de dictionnaire des mots positif

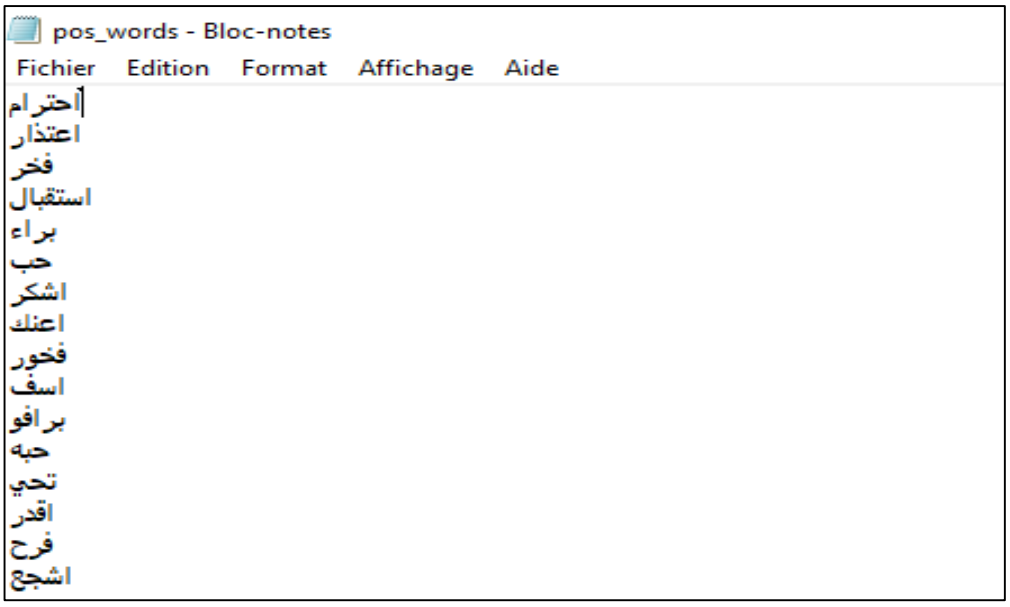


Figure 3. 4 : Exemple d'une partie de dictionnaire (positif).

- **Deuxième fichier :**

Cette figure présente le fichier de dictionnaire des mots négatif

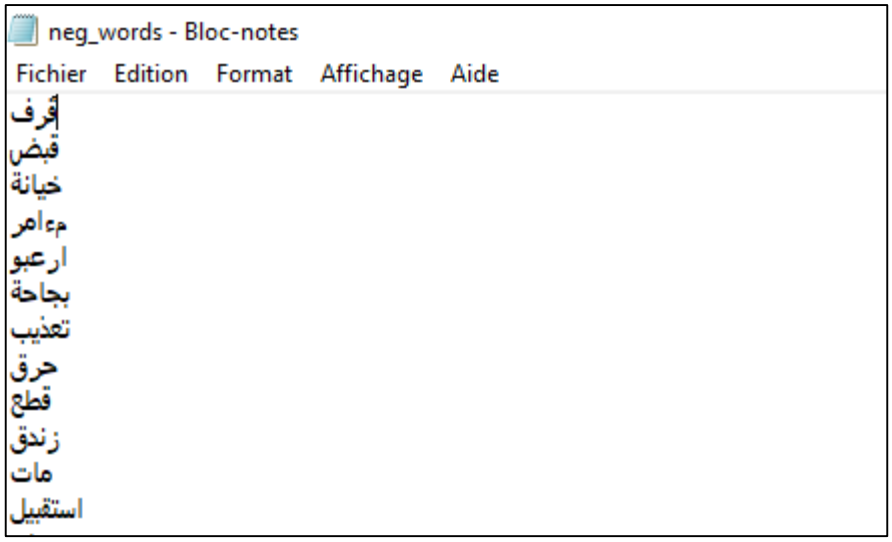


Figure 3. 5 : Exemple d'une partie de dictionnaire (négatif).

3.4 Traitement des données

Maintenant que les données sont prêtes, nous commençons la phase de mise en œuvre

Nous proposons algorithme général de l'analyse suivant :

Algorithme

Importer les bibliothèques

Lire le Dataset

Lire le dictionnaire

Début

Extraire les fonctionnalités

Prétraitaient

Appeler aux classificateurs

Afficher les résultats

Fin

3.4.1 Exemples de codes sources

Dans cette section, nous allons présenter quelques exemples de codes sources. En mentionnant les étapes du traitement.

La Figure 3.6 présente un morceau de code qui permet d'appeler les bibliothèques nécessaires pour compiler notre application

```
import pandas as pd
import numpy as np
import re
import nltk
import csv
import re
import string
from nltk.corpus import stopwords
```

Figure 3. 6 : L'appeler les bibliothèques nécessaires

3.4.2 Prétraitement de texte arabe en Python

Dans chaque tâche de traitement de la langue naturelle, il y a quelques étapes de prétraitement communes que nous devons faire. Je vais expliquer et fournir le code pour les techniques de prétraitement suivantes en python.

3.4.2.1 Nettoyage

• Supprimer des numéros

• Supprimer les signes diacritiques

• Supprimer les caractères remplacés

• Supprimer les mots non arabes…

• Supprimer les ponctuations

Ces images montrent les fonctions en cours de traitement :

```python
#Supprimer les signes diacritique
def remove_diacritics(string):
    regex = re.compile(r'[\u064B\u064C\u064D\u064E\u064F\u0650\u0651\u0652]')
    return re.sub(regex, '', string)

#Supprimer les symboles non arabes
def remove_non_arabic_symbols(string):
    return re.sub(r'[^\u0600-\u06FF]', ' ', string)

#supprimer les URLs
def remove_urls(string):
    regex = re.compile(r"(http|https|ftp)://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\(\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+")
    return re.sub(regex, ' ', string)

#Supprimer les nombres
def remove_numbers(string):
    regex = re.compile(r"(\d|[\u0660\u0661\u0662\u0663\u0664\u0665\u0666\u0667\u0668\u0669])+")
    return re.sub(regex, ' ', string)

#Supprimer les mots non arabes
def remove_non_arabic_words(string):
    return ' '.join([word for word in string.split() if not re.findall(
        r'[^\s\u0621\u0622\u0623\u0624\u0625\u0626\u0627\u0628\u0629\u062A\u062B\u062C\u062D\u062E\u062F\u0630\u0631\u0632\u0633
        word)])
```

Figure 3. 7 : Les fonctions de traitement

3.4.2.2 Supprimer les lettres et les mots répétés

La plupart des commentateurs utilisent la répétition de lettres et de mots pour exagérer l'expression du sentiment, il devrait donc être supprimé pour faciliter le processus de traitement.

3.4.2.3 Normalisation

Cette phase permet de rendre tous les mots à la forme normale ce qui veut dire que si il y a un mot écris avec des caractères doublons par exemple il sera réduit à une seule lettre et les doublons seront supprimés

```python
#Normalise une chaîne
def noramlize(string):
    regex = re.compile(r'[ٱإأآ]')
    string = re.sub(regex, 'ا', string)
    regex = re.compile(r'[ى]')
    string = re.sub(regex, 'ي', string)
    regex = re.compile(r'[ؤئ]')
    string = re.sub(regex, 'ء', string)
    return string
```

Figure 3. 8 : Normalise une chaine.

3.4.2.4 Élimination des mots vides

Le texte est volumineux et plein de vocabulaire et de nombreux mots, et il est naturel qu'ils n'aient pas tous la même valeur. Les mots qui se répètent fréquemment et ne portent pas de sens par leur seule présence sont considérés comme du bruit et des mots vides qu'il faut supprimer

Les lettres, outils et pronoms de la langue arabe, tels que : qui, sur, cela, est, ceux-ci, etc., sont souvent utilisés et sont bruyants dans le texte, il est donc préférable de les laisser de côté, mais concentrons-nous d'abord sur le cas général et uniquement sur les lettres et les outils, la boîte à outils de traitement du langage naturel (NLTK) nous y aidera car elle fournit un ensemble général de mots vides pour un certain nombre de langues, dont l'arabe et l'anglais, et pour l'utiliser, vous peut appliquer le code suivant :

Pour supprimer les mots de dotation du panier, nous avons créé un fichier qui contient tous les mots arabes. L'image montre une partie de ce fichier.

```
def stopwordremove (text):
    stop_word = open('stop_words.txt','r+',encoding='utf-8')
    stop_word= stop_word.read().split("\n")
    ar_stop_word= set(w.rstrip()for w in stop_word)
    ST=[]
    words= word_tokenize(text)
    for w in not in(ar_stop_word):
        ST.append(w)
    filtered_sentenc= "".join(neededword)
    return filtered_sentenc
```

Figure 3. 9 : Élimination des mots vides

Le tableau représente quelques commentaires avant et après prétraitement

Étape	Phrase
Normal	بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيمِ مرررررحبا بكم في مسابقة جديدة للتسجيل 🤓 اضغط هنا https://www.google.com/
Filtre	بسم الله الرحمن الرحيم مرررحبا بكم في مسابقة جديدة ل لتسجيل اضغط هنا
Supprimer les lettres et les mots répétés	بسم الله الرحمن الرحيم مرحبا بكم في مسابقة جديدة للتس جيل اضغط هنا
Normalisation	بسم الله الرحمن الرحيم مرحبا بكم في مسابقة جديدة للتسجيل اضغط هنا
Élimination des mots vides	بسم الله الرحمن الرحيم مرحبا مسابقة جديدة للتسجيل اضغط

Table 3. 1: Commentaires avant et après prétraitement.

3.4.2.5 Tokénisation

Dans la langue arabe, le principal séparateur entre les mots est l'espace, ce qui facilite la tâche. La boîte à outils linguistique peut être utilisée pour effectuer cette tâche :

nltk.word_tokenize (sentence**)**

3.4.2.6 Stemming et lemmatisation

Premièrement, la structure du mot doit être connue en langue arabe, comme suit :

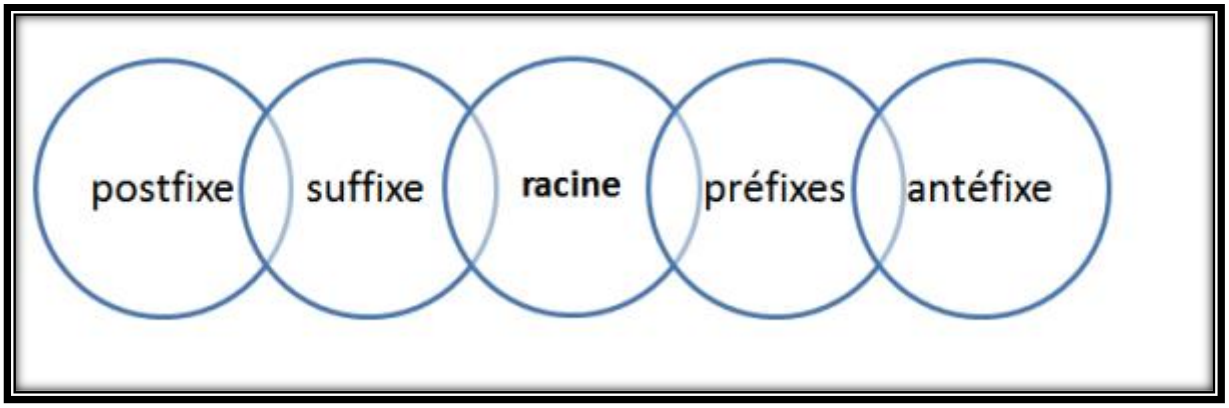


Figure 3. 10 : La structure du mot en langue arabe**.**

Le joule représente les mots avant et après Stemming et lemmatisation :

les mots	Stemming et lemmatisation
بانتظام منتظم منتظمة انتظام	نظام
ينفعنا ينتفع انتفاع	نافع
تندم يندم نادم الندم نادمة	ندم
سيء سوء إساءة مسيء يسيء	سيئ

Table 3. 2 : Représentions des mots avant et après Stemming et lemmatisation

Naïve Bayes Classifier :

Un classificateur est une fonction qui attribue une étiquette de classe à un tweet. Du point de vue de la probabilité, selon la règle de Bayes, la probabilité qu'un tweet soit positif, négatif ou neutre est donnée comme suit :

$$P(\textbf{tweet}/\text{positif}) = \frac{P(\text{tweet}/\text{positif})\,P(\text{positif})}{P(\text{tweet})}$$

$$P(\textbf{tweet} / \textbf{negatif}) = \frac{P(\text{tweet/negatif})\, P(\text{negatif})}{P(\text{tweet})}$$

$$P(\textbf{tweet} / \textbf{neutre}) = \frac{p(\text{tweet/neutra})\, P(\text{neutre})}{P(\text{tweet})}$$

Exemple : considérons le texte suivant des tweets de formation

tweet d'origine	Tweet après traitement	فخر	وطن	تعب	طلب	عمل	صبح	خير	class
نفخر بالوطن	فخر وطن	1	1	0	0	0	0	0	positif
تعبنا نطالب بعمل في الوطن	تعب طلب عمل وطن	0	1	1	1	1	0	0	Négative
نعمل بفخر في الوطن	عمل فخر وطن	1	1	0	0	1	0	0	positif
صباحكم خير	صبح خير	0	0	0	0	0	1	1	positif

Table 3. 3 : Classification des mots dans les tweets en négatif et positif

Mot	positif	négatif	neutre	P(x)	
فخر	2	0	0	2/11	0,18
وطن	2	1	0	3/11	0,27
تعب	0	1	0	1/11	0,09
طلب	0	1	0	1/11	0,09
عمل	1	1	0	2/11	0,18
صبح	0	0	1	1/11	0,09
خير	0	0	1	1/11	0,09
Total	5	4	2		
P(c)	5/11	4/11	2/11		
	0,45	0,36	0,18		

Table 3. 4 : L'application de l'algorithme Naïve Bayes sur les tweets

- Ces probabilités estimées sont utilisées pour prédire la classe du nouveau tweet suivant

tweet d'origine	Tweet après traitement	class
صباحكم خير و عمل و وطن يفخر فيكم	صبح خير وطن فخر	-

Table 3. 5 : Résultats de la classification

- P(positif / new Tweet) P(X/c1)×P(c1)= 0,0000288
- P(négative / new Tweet) P(X/c2)×P(c2)= 0,00000850
- P($neutre$ / new Tweet) P(X/c3)×P(c3)= 0,00001159
- ainsi le classificateur attribue la nouvelle classe de tweet = positif

3.5 Présentation de l'application réalisée

Notre application se compose d'un ensemble d'interfaces et de fonctions qui aident l'utilisateur aux sentiments de l'analyse du texte arabe et d'une part permet à la machine de donner compréhensible aux résultats de l'utilisateur.

3.5.1 L'interface principale de l'application

L'interface principale de l'application où l'utilisateur peut construire son processus comme le montre la figure suivante :

Figure 3. 11 : L'interface principale de l'application.

3.5.2 Seconde interface :

L'application effectue le processus de traitement et affiche le nombre de mots positifs et le nombre de mots négatifs

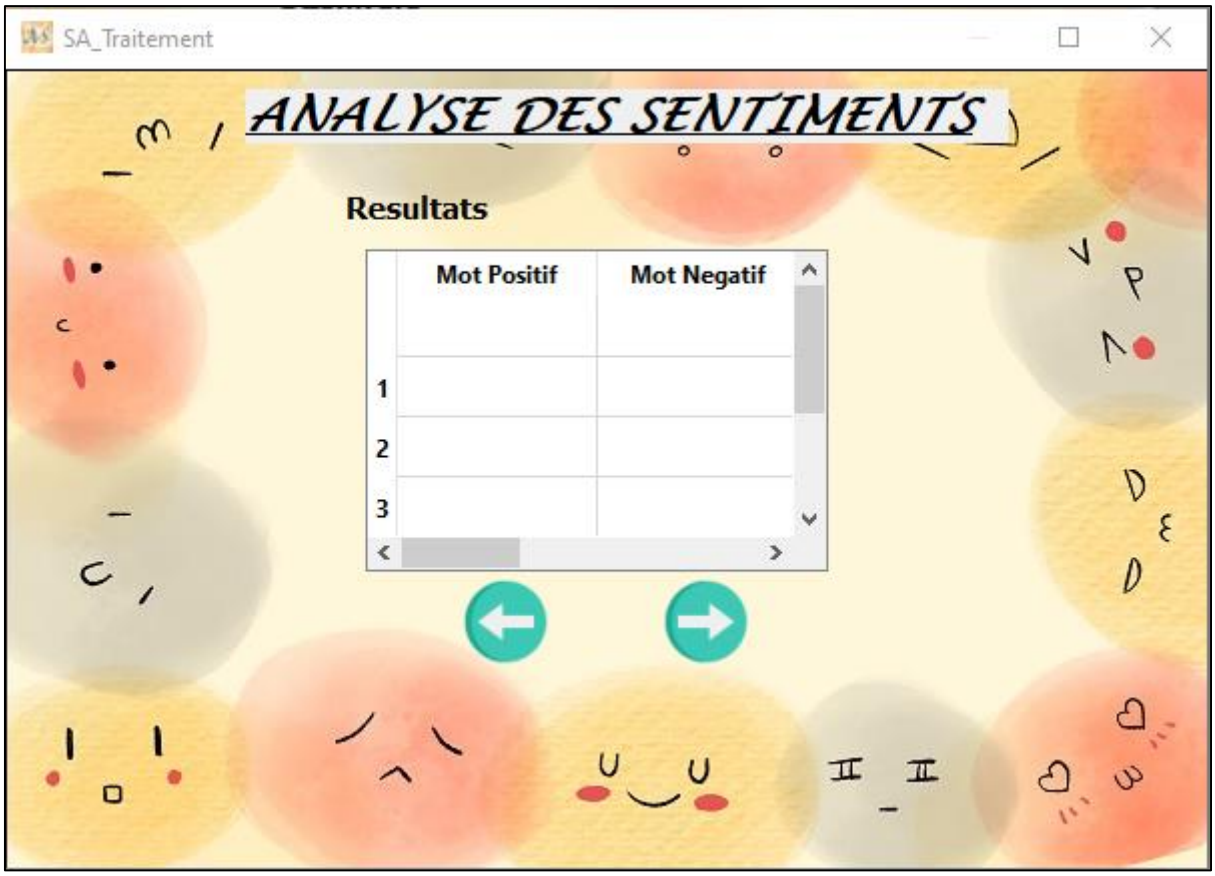


Figure 3. 12 : Seconde interface.

3.5.3 La dernière interface :

À la fin elle s'affiche si le texte que nous avons saisi est négatif ou positif

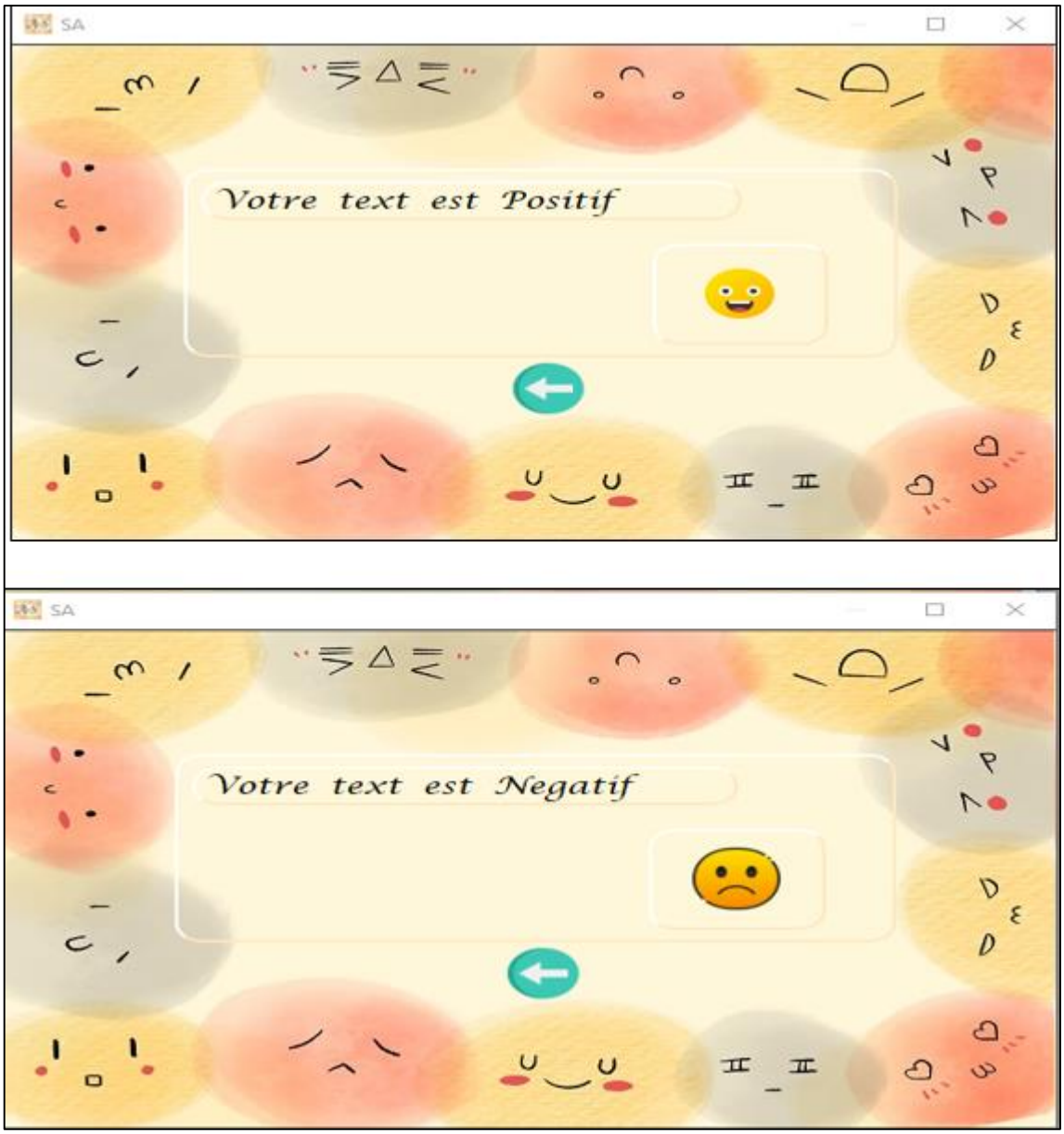


Figure 3. 13 : dernière interface.

3.6 Conclusion

Dans ce chapitre, nous avons décrit comment mettre en œuvre les différentes étapes de création d'un modèle d'analyse des sentiments, dont nous avons précédemment expliqué en détail la conception.

CONCLUSION GENERALE

CONCLUSION GENERALE

Le domaine de l'analyse des sentiments se développe très rapidement et vise à utiliser les opinions ou les textes présents dans diverses plateformes médiatiques grâce à des techniques d'apprentissage en profondeur.

Il est devenu un domaine de recherche très en vogue. De nombreuses recherches ont été effectuées dans ce domaine, mais comme l'analyse des sentiments traite de données textuelles non structurées, de nombreux problèmes subsistent.

Par conséquent, dans ce mémoire, nous introduisons d'abord le processus d'analyse des sentiments, y compris ses applications, tâches et défis de l'analyse des sentiments. De même, nous introduisons des techniques d'apprentissage en profondeur

L'objectif de notre travail est de construire un modèle pour analyser les sentiments et les opinions dans un texte arabe afin qu'il puisse être utilisé dans la planification et la prise de décision dans tous les domaines, à travers un dictionnaire de polarisation qui se compose de "positif et négatif" au cours du processus de collecter et classer manuellement les tweets.

Nous avons identifié manuellement les termes positifs et négatifs et appliqué des modèles d'apprentissage en profondeur pour classer les textes selon un lexique précédemment rapporté.

Malgré toutes les difficultés qu'on les a rencontrais telle que : le comprendre la structure interne unique de la langue arabe, sa nature, ses termes et ses règles linguistiques. Chaque forme a sa propre syntaxe et son propre vocabulaire, ce qui rend difficile la construction d'un lexique arabe. De plus, différents mots peuvent avoir le même sens ou un mot peut avoir des sens différents. Mais, nous pensons qu'on a quand même pu relever le défi.

Pour conclure, ce travail peut être amélioré à la future par la formation de systèmes d'analyse des sentiments sur les textes arabes exprimés en lettres latines, ainsi que d'autres traits discriminatoires tels que la négation et le ridicule, qui n'ont pas encore été étudiées.

Bibliographie

[1] A. Aymen, «Traitement Automatique de la langue naturelle», 22/4/2022.

[2] Chenni Rania Wissem, «Analyse des sentiments arabes en utilisant L'apprentissage en profondeur », Mémoire de Master,2020.

[3] M. Hadji, ««Analyse des sentiments : Généralités»,» August 2019.

[4] I. A. e. M. Mohamed, «A Review on Sentiment Analysis in Arabic Using Document Level» ,Faculty Informatics' & Computing, University Sultan Zainal Abidin, Besut Campus, Terengganu, Malaysia, 2018.

[5] naaima boudad, radouan faizi ,Rachid oulad haj thami , raddouane chihed,Sentiment analysis in arabic, 2017.

[6] K.R.Beesley, «Finite-state morphological analysis and generation of Arabic at Xerox Research: Status and plans in 2001,» *Presented at the ACL Workshop on Arabic Language Processing: Status and Perspective*, vol. 1, p. 1–8, 2001.

[7] *Al-Subaihin, A.A, H.S. Al-Khalifa, and A.S. Al-Salman.«A proposed sentiment analysis tool for modern arabic using human-based computing». In proceedings of the 13th international conference on information integration web-based applications and services,* 2011.

[8] H. B. a. I. B. A. Mountassir, A cross-study of Sentiment Classification on Arabic corpora. In Max Bramer and Miltos Petridis, editors, pages 259–272, London: Research and Development in Intelligent Systems XXIX, 2012.

[9] A. Al Hasan, buliding a sentiment lexicon for the Palestinian dialect., 2016.

[10] G. T. O. a. T. H.-R. Alwakil, Challenges in sentiment analysis for Arabic social networks.117: p. 89-100., 2017.

[11] N. D. B. a. M. B. Yussupova, «Applying of sentiment analysis for texts in Russian based on machine learning approach». In Proceedings of Second International Conference on Advances in Information Mining and Management., Russian, 2012.

[12] Balahur, A. Sentiment analysis in social media texts. in Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis, 2013.

[13] U. Hodeghatta, Sentiment analysis of Hollywood movies on Twitter. in 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE., ASONAM 2013, 2013.

[14] D. Saxena, «Sentiment analysis,» n° %18(3), pp. 46-51, 2019.

[15] I. Zakariyaa, «Apprentissage Supervise VS non Supervise,» [En ligne]. Available: https://analyticsinsights.io/apprentissage-supervise-vs-non-supervise. [Accès le 14 04 2022].

[16] Tokenising into Words and Sentences | what is Tokenization and it's Definition? By Great Learning Team, May 29, 2020.

[17] Sana Alowaidi, Mustafa Saleh, Osama Abulnaja Semantic Sentiment Analysis of Arabic Texts. Computer Science Department King Abdulaziz University Jeddah, Saudi Arabia.

[18] N. M. Abdelhamid, *« Techniques d'apprentissage automatique pour l'analyse et la fouille des sentiments dans les réseaux sociaux», thèse de Doctorat en informatique,* BISKRA: Université Mohamed Khider, 2021.

[19] M. &. F. H. &. A. I. Alrefai, Sentiment analysis for Arabic language: A brief survey of approaches and techniques., 16 Mars 2020.

[20] Tavish ,« Introduction to k-Nearest Neighbors: A powerful Machine Learning », 26, 2018.

[21] sunil 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R, September 11, 2017.

[22] V. N. a. S. Agrawal, « « Political sentiment analysis using hybrid approach »,» *International Research Journal of Engineering and Technology (IRJET),* vol. 3, 2016.

[23] «M. Desai and M. A. Mehta,« A hybrid classification algorithm to classify engineering students 'problems and perks »,» *Computers and Society,* p. 21–35, 2016.

[24] «« Le tutoriel Python»,» [En ligne]. Available: https://docs.python.org/fr/3/tutorial. [Accès le 12 05 2022].

[25] R. Sharma, ««Python Anaconda Tutorial: Everything You Need to Know»,» [En ligne]. [Accès le 12 05 2022].

[26] D. Mike, «Jupyter Notebook: An Introduction,» [En ligne]. [Accès le 12 05 2022].

ملخص

اليوم، مع الانتشار الواسع لوسائل التواصل الاجتماعي، يتم إنشاء كمية هائلة من البيانات في شكل وجهات نظر وآراء ومشاعر حول الأحداث الاجتماعية المختلفة والمنتجات والعلامات التجارية والسياسات .يستخدم تحليل المشاعر لتصنيف المشاعر المعبر عنها بطرق مختلفة، مثل السلبية والإيجابية.في هذا العمل، نهدف إلى تحقيق نظام تحليل المشاعر المستخرج من النصوص العربية .في الوقت نفسه، نسلط الضوء على الأبحاث الحديثة حول تنفيذ نماذج تحليل المشاعر مثل التعلم العميق .لحل مشاكل تحليل المشاعر المختلفة.

الكلمات المفتاحية: تحليل المشاعر، تحليل النصوص العربية، اللغة العربية، التعلم العميق، الشبكات الاجتماعية، تويتر.

Abstract

Today, with the widespread distribution of social media, a huge amount of data is generated in the form of views, opinions, and feelings about different social events, products, brands, policies...Sentiment analysis is used to classify the feelings expressed in different ways, such as negative, positive .In this work, we aim to achieve a sentiment analysis system extracted from Arabic texts. At the same time, we highlight recent research on the implementation of sentiment analysis models such as deep learning. To solve different sentiment analysis problems.

Keywords: Sentiment analysis, Arabic language, deep learning, social networks, twitter.

Résumé

Aujourd'hui, avec la diffusion généralisée des médias sociaux, une énorme quantité de données est générée sous forme de vues, d'opinions et de sentiments sur différents événements sociaux, produits, marques, politiques.., L'analyse des sentiments est utilisée pour classer les sentiments exprimés de différentes manières, telles que négatives, positives. Dans ce travail, nous visons à réaliser un système d'analyse des sentiments extraites à partir des textes arabes. Dans le même temps, nous soulignons les recherches récentes sur la mise en œuvre de modèles d'analyse de sentiments tels que l'apprentissage en profondeur. Pour résoudre différents problèmes d'analyse de sentiments.

Mots clés : Analyse des sentiments, langue arabe, l'apprentissage en profondeur, réseaux sociaux, twitter.