République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad DAHLAB - Blida 1

Faculté des sciences

Département d'Informatique

Mémoire présenté par :

Mlle. HABES Yasmine

Pour l'obtention du diplôme de Master

Domaine : Mathématique et Informatique

Filière : Informatique

Spécialité : Traitement Automatique de la Langue

Sujet :

Application des méthodes d'Apprentissage Automatique dans l'Analyse des Sentiments des Tweets Arabes

Soutenu le 21 janvier 2021, devant le jury composé de :

Mme. BENBLIDIA	Université de Blida 1	Président
Mr. ABBACHE	Université de Chlef	Examinateur
Mme. G. DROUA	CRSTDLA	Encadreur
Mme.M.MEZZI	Université de Blida 1	Promoteur

Résumé

Avec l'expansion spectaculaire de l'information sur Internet, les utilisateurs du monde entier expriment quotidiennement leur opinion sur le réseau social tel que Facebook et Twitter. Les grandes entreprises investissent aujourd'hui dans l'analyse de ces opinions afin d'évaluer leurs produits ou services grâce aux commentaires des gens sur leurs activités. Le processus de connaissance des opinions des utilisateurs sur les produits ou services, qu'elles soient positives ou négatives, est appelé analyse des sentiments. L'arabe est l'une des langues courantes qui ont suscité l'intérêt de cette discipline. Dans la littérature, plusieurs approches ont été proposées pour l'Analyse des Sentiments arabes et la plupart de ces approches utilisent des techniques d'apprentissage automatique. Par conséquent, dans cette étude, nous essayons d'identifier une approche simple mais réalisable pour l'Analyse des Sentiments arabes sur Twitter. Cette solution proposée se base sur les différentes techniques de l'apprentissage automatique « Machine Learning » avec deux méthodes supervisées SVM (Support Vector Machine) et K-NN (K-Nearest Neighbors).

Mots clés : Analyse des Sentiments, Analyse d'opinion, Langue Arabe, L'apprentissage automatique, Méthodes SVM et K-NN

Abstract

With the dramatic expansion of information on the Internet, users around the world are expressing their opinions daily on social networks such as Facebook and Twitter. Large companies now invest in the analysis of these opinions in order to evaluate their products or services through people's feedback on their activities. The process of knowing users' opinions about products or services, whether positive or negative, is called feelings analysis. Arabic is one of the common languages that has attracted the interest of this discipline. In the literature, several approaches have been proposed for the Analysis of Arab Feelings and most of these approaches use machine learning techniques. Therefore, in this study, we try to identify a simple but feasible approach for Arab Sentiment Analysis on Twitter. This proposed solution is based on the different techniques of machine learning «Machine Learning» with two supervised methods SVM (Support Vector Machine) and K-NN (K-Nearest Neighbors).

Keywords: Sentiment Analysis, Opinion Analysis, Arabic Language, Machine Learning, SVM and K-NN Methods

ملخص

ومع التوسع الهائل في المعلومات على الإنترنت، يعرب المستخدمون في مختلف أنحاء العالم عن آرائهم يوميا حول الشبكات الاجتماعية مثل فايسبوك وتويتر. والآن تستثمر الشركات الكبرى في تحليل هذه الآراء من أجل تقييم منتجاتها أو خدماتها من خلال ردود فعل الناس على أنشطتها. تسمى عملية معرفة آراء المستخدمين حول المنتجات أو الخدمات، سواء كانت إيجابية أو سلبية، بتحليل المشاعر. إن اللغة العربية واحدة من اللغات المشتركة التي جذبت اهتمام هذا النظام. في الأدب، تم اقتراح العديد من المناهج لتحليل المشاعر العربية ومعظم هذه الأساليب تستخدم تقنيات التعلم الآلي. لذا، نحاول في هذه الدراسة تحديد نهج بسيط ولكنه ممكن لتحليل المشاعر العربية على تويتر. يستند هذا الحل المقترح إلى الأسلوب المختلفة لتعلم الآلة »تعلم الآلة« SVM وK-NN طريقتين خاضعتين للإشراف

الكلمات المفتاحية: تحليل المشاعر، تحليل الآراء، اللغة العربية، تعلم الآلة، طرق SVM وK-NN

Dédicaces

Gloire soit rendu au Dieu tout puissant créateur de toutes choses, le très miséricordieux pour tous les bienfaits dont il m'a comblé et pour m'avoir donné le courage et la force pour réaliser ce modeste travail que je dédie à:

Mes chères parents pour leur orientations, leurs encouragements, leurs soutiens et conseils indéfectible.

Ma Petite Famille : Mon mari et Ma fille Sydra

Mon cher frère Takieddinne

Mes chères sœurs Bouthaina, Chahd et Loudjaine et Ma nièce Razane

A Mes deux chères amies Nassiba Adjani et Romaissa Benbaibeche

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué au succès de mon projet et qui m'ont aidée lors de la rédaction de ce mémoire.

Je voudrais dans un premier temps remercier, mon encadrante Mme.Droua Ghania, pour sa patience, son orientation, et la disponibilité qu'elle a témoignée pour me permettre de mener à bien ce travail et surtout ses judicieux conseils, qui ont contribué à alimenter ma réflexion.

Je tiens à témoigner toute ma reconnaissance aux personnes suivantes, pour leur aide dans la réalisation de ce projet :

Je tiens à remercier Mme .Mezzi qui m'a beaucoup aidé pour avoir relu et corrigé mon mémoire. Et avoir répondu à mes questions sur la rédaction, elle a été d'un grand soutien dans l'élaboration de ce mémoire. Ses conseils de rédaction ont été très précieux.

Et je n'oublie pas Mes Parents, Mon Mari, Ma fille, Mon Frère et Mes Sœurs, je tiens à remercier tous ses personnes pour leur soutien constant et leurs encouragements

Yasmine

Table des matières

Liste des Figures

Liste des Tableaux

Liste d'Acronymes

IA : Intelligence artificielle

AS : Analyse des Sentiments

ASA : Analyse des Sentiments Arabe

NLP : Natural Langage Processing

TLN : Traitement du LangUages Naturel

NLU: Naturel Langage Understanding

POS: Part-Of-Speech

ML: Machine Learning

K-NN: K-Nearest Neighbors

DT: Decision tree

NB: Naive Bayes

SVM: Support Vector Machine

IDE: Integrated Development Environment

API: Application Programming Interface

URL: Uniform Ressource Locator

NLTK: Natural Language Toolkit

VSM: Vector Space Model

BOW: Bag Of Words

TF : Terme de Fréquence

IDF : Inverse Terme de Fréquence

Introduction Générale

1.1 Contexte global

Avec l'expansion spectaculaire de la World Wide Web et de l'information sur internet grâce à la rapidité de réponse qu'elle offre, Internet permet de faciliter les échanges d'informations avec son correspondant et permet également de faire un partage de connaissances et de divertissements avec les proches familles et amis. Cela est encore plus vrai avec l'utilisation des sites, des communautés de relations et des différents réseaux sociaux.

Les utilisateurs du monde expriment quotidiennement leur opinion sur le réseau social tel que Facebook et Twitter. Récemment, la plupart des entreprises ont un essentiel besoin de vérification de leurs services ou produits. Ces demandes dépendent du point de vue des consommateurs sur ces services ou produits. Par conséquent, la connaissance des consommateurs sur les opinions sont devenues très difficiles.

Le traitement des enquêtes sur les opinions des utilisateurs est devenu plus accessible et plus simple. Habituellement, ces informations seraient en mode textuel. Donc, l'utilisation de technologies récentes telles que le Web mining et le web sémantique facilite l'analyse du texte qui conduit à extraire les connaissances. Tout cela est considéré comme un processus appelé analyse des sentiments.

L'analyse des sentiments, également appelée opinion mining, est une étude de l'informatique sur des opinions, des sentiments et des attitudes sur des sujets, des entités, des personnes et des événements, exprimé à travers du texte, entre autres. Il vise à attribuer un sentiment prédéfini classé aux textes comme négatifs, positifs ou neutres. Ce processus vise à une meilleure compréhension de l'opinion publique. De nombreuses études ont montré que l'analyse du sentiment est très intéressante pour les personnes qui se concentrent sur l'opinion publique, pour de nombreuses raisons personnelles, commerciales ou politiques. Ainsi, de nombreuses applications et systèmes ont déjà été développés pour cette tâche, ils sont généralement formés sur des données textuelles évaluatives traditionnelles comme sur des textes non-traditionnels les messages et les Tweets de réseaux sociaux qui deviennent une ressource très importante et précieuse pour les internautes concernant leurs opinions et leurs orientations.

Actuellement, Twitter est considéré comme l'un des microblogs les plus populaires. Il a permis aux gens de communiquer, partager des commentaires et exprimer leurs opinions sur presque tous les aspects de la vie quotidienne.

L'Analyse des Sentiments est devenue une nécessité, Mais, elle peut être difficile à implémenter comme le langage humain est complexe à interpréter pour les systèmes d'apprentissage basés sur la machine. Le traitement des sentiments s'avère plus complexe, ce qui entraine la nécessité d'emploi d'autres domaines tels que : le Traitement Automatique du Langage Naturel, l'apprentissage automatique avec ces différents classifies.

1.2 Problématique

Malgré que l'arabe fasse partie des 10 langues les plus couramment utilisées sur Internet (d'après la classification établie par Internet World State en 2018), et est parlé par des centaines de millions de personnes, les recherches effectuées sur l'analyse du sentiment en langue Arabe sont très limitées, en particulier les dialectes par rapport à d'autres langues comme l'Anglais. Cela peut s'expliquer par deux facteurs : Premièrement, le manque de ressources pour ce type de langues. Pour l'analyse du sentiment il y a un manque de corpus annotés (étiquetés). Deuxièmement, la complexité du traitement de cette langue : Le problème majeur auquel nous sommes confrontés lors du traitement des données Twitter c'est que d'une part, les dialectes sont associés à aucune forme d'écriture normalisée et contiennent du bruit, des fautes d'orthographe, des abréviations, des répétitions, et des mots qui ne suivent aucune règle grammaticale.

Un autre problème dans l'analyse des sentiments c'est le problème du classement de la polarité (Positive, Négative, Neutre) à partir de données textuelles à l'échelle Web qui est une tâche très difficile et coûteuse en raison de la grande quantité de données bruitées.

1.3 Objectifs de l'étude

La majorité des études faites dans le domaine concernent la langue anglaise. C'est pour cela que nous avons proposé une étude pour la langue arabe. Cette langue est une langue sémitique et cursive. C'est la langue de près de 168 millions de locuteurs arabes utilisant l'internet. Selon le monde Internet Classement Stat 2016, la langue arabe est classée dans les tops cinq langues les plus utilisées sur l'internet. C'est pourquoi, il est important de concevoir des systèmes et des applications automatiques capables d'analyser, de détecter et d'extraire les sentiments et les opinions des individus qui s'expriment en arabe sur les réseaux sociaux.

Dans ce projet, notre objectif est d'explorer le domaine de l'Analyse des Sentiments, et de détecter automatiquement les sentiments des internautes et leurs avis pour ou contre un produit ou un phénomène social. En développant un système d'Analyse des Sentiments pour classer les opinions en trois catégories : positif, négatif et neutre. Nous avons également l'intention de présenter l'approche d'apprentissage automatique (Machine Learning), de ses applications et de développer notre propre modèle en tant qu'une contribution à la problématique de l'analyse du sentiment

1.4 Organisation du mémoire

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

Le premier chapitre est consacré au domaine de l'Analyse des Sentiments et Opinion Mining. Nous présenterons les différents niveaux d'analyse des sentiments, leurs avantages et inconvénients spécifiques à la langue Arabe, les outils d'Analyse des Sentiments, ainsi qu'une description des caractéristiques principales des réseaux sociaux, ainsi que leur influence sur le domaine d'Analyse des Sentiments.

Par la suite, le deuxième chapitre exposera le processus de construction l'Analyse des Sentiments en langues Arabe et présentera les différentes approches de classification.

Par ailleurs, le troisième chapitre présentera les étapes nécessaires et la démarche détaillée de notre système

Ensuite, le quatrième chapitre définira les outils de programmation et l'implémentation de notre application, présentation des interfaces et les résultats d'exécution.

Finalement, nous clôturons ce mémoire par une conclusion générale.

Chapitre I:

Traitement du Langage Naturel

et l'Analyse des Sentiments

1 Introduction

Avec l'émergence de la technologie Web 2.0 et le nombre croissant de sites Web et réseaux sociaux et de forums Web, les utilisateurs actuels d'Internet ont la possibilité d'ajouter leurs avis, notes ou opinions sur des réseaux sociaux. De plus, les gens peuvent consulter souvent les forums de discussion avant de prendre une décision d'investissement en demandant à leurs amis leurs opinions et extraire manuellement de bonnes ou de mauvaises critiques de certain sujet, pour avoir tous cela l'émergence d'une nouvelle technique conduit à étudier les opinions liées à un sujet spécifique qui s'appelle l'Analyse des Sentiments.

Dans ce chapitre nous allons expliquer et exposer le Traitement du Langage Naturel et aborder le domaine de l'Analyse des Sentiments en présentant ses différents types, niveaux et aussi les spécifications liées la langue Arabe.

2 Traitement du Langage Naturel

Le Traitement du Langage Naturel est l'approche informatisée de l'analyse de texte qui repose à la fois sur un ensemble de théories et sur un ensemble de technologies [1]. C'est un domaine d'actualité en recherche et développement, il n'existe pas une seule définition convenue qui satisfasse tout le monde, mais il y a certains aspects standards dont nous allons essayer d'exposer les grandes lignes.

2.1 Définition

Le Traitement du Langage Naturel est une gamme de techniques de calcul pour analyser et représenter des textes naturels à un ou plusieurs niveaux d'analyse linguistique dans le but d'effectuer des traitements du langage pour une gamme de tâches ou d'applications.

Plusieurs éléments de cette définition peuvent être détaillés. Premièrement, la notion imprécise de «Gamme de techniques de calcul» est nécessaire car il existe plusieurs méthodes ou techniques parmi lesquelles choisir pour réaliser un type particulier d'analyse du langage.

Les «textes naturels» peuvent être de n'importe quelle langue, mode, genre, etc. Les textes peuvent être oraux ou écrit. La seule exigence est qu'ils soient dans une langue utilisée par les humains pour communiquer entre eux. De plus, le texte analysé ne doit pas être spécifiquement construit aux fins de l'analyse, mais plutôt que le texte soit recueilli à partir de l'usage.

La notion de «niveaux d'analyse linguistique» renvoie au fait qu'il existe plusieurs types de traitement du langage connus lorsque les humains produisent ou comprennent le langage. On pense

que les humains utilisent normalement tous de ces niveaux puisque chaque niveau véhicule différents types de sens [1].

2.2 Objectif

L'objectif de la NLP, comme indiqué ci-dessus, est «d'accomplir un traitement du langage semblable à l'homme». Le choix du mot «traitement» est très délibéré et ne doit pas être remplacé par 'compréhension'. Car bien que le domaine de la NLP ait été à l'origine appelé Naturel Language Understanding (NLU) dans les premiers jours de l'Intelligence Artificielle (IA), il est bien admis aujourd'hui que si l'objectif de la NLP est le vrai NLU, cet objectif n'a pas encore été atteint. Dans un NLU complet Le système serait capable de [1]:

1. Paraphraser un texte d'entrée.
2. Traduire le texte dans une autre langue.
3. Répondre aux questions sur le contenu du texte.
4. Tirer des inférences à partir du texte.

2.3 Niveaux de Traitement du langage naturel

La méthode la plus explicative pour présenter ce qui se passe réellement dans le système de traitement du langage utilise l'approche des «niveaux de langue». C'est également appelé modèle synchronique du langage et se distingue du modèle séquentiel, qui émet l'hypothèse que les niveaux de traitement du langage humain se succèdent de manière strictement séquentielle.
La description de certains niveaux sera présentée séquentiellement [1] :

- **Phonologie :** Ce niveau traite de l'interprétation des sons de la parole dans les mots. Là en fait, trois types de règles utilisées dans l'analyse phonologique:
  - les règles phonétiques : pour les sons dans les mots
  - les règles phonémiques : pour les variations de prononciation lorsque les mots sont parlés ensemble
  - les règles prosodiques : pour la fluctuation de la tension et de l'intonation à travers une phrase
- **Morphologie :** Ce niveau traite de la nature composante des mots, qui sont composés de Morphèmes, les plus petites unités de sens. Par exemple, le mot préinscription peut être analysé morphologiquement en trois morphèmes distincts: le préfixe, la racine et le suffixe.
- **Lexique:** À ce niveau, les humains, ainsi que les systèmes NLP, interprètent le sens des mots individuels. Plusieurs types de traitement contribuent à la compréhension au niveau du mot -

le premier d'entre eux étant l'attribution d'une seule étiquette de partie de discours à chaque mot.

- **Syntaxe :** Ce niveau se concentre sur l'analyse des mots d'une phrase afin de découvrir la grammaire ou la structure de cette dernière.
- **Sémantique :** Ce niveau se concentre sur l'analyse des mots d'une phrase afin de découvrir le sens.

2.4 Applications du Traitement du langage naturel

Le Traitement du Langage Naturel fournit à la fois de la théorie et des implémentations pour une gamme d'applications. En fait, toute application utilisant du texte est candidate à la NLP. De plus, les applications fréquentes utilisant la NLP sont les suivantes [1]:

Recherche d'informations : est le domaine qui étudie la manière de retrouver des informations dans un corpus. Celui-ci est composé de documents d'une ou plusieurs bases de données, qui sont décrits par un contenu ou les métadonnées associées. Les bases de données peuvent être relationnelles ou non structurées, telles celles mises en réseau par des liens hypertexte comme dans le World Wide Web, l'internet . Le contenu des documents peut être du texte, des sons, des images ou des données.

Extraction d'informations (EI) : est la tâche d'extraire automatiquement des informations structurées à partir de documents lisibles par machine non structurés et / ou semi-structurés et d'autres sources représentées électroniquement. Dans la plupart des cas, cette activité concerne le traitement de textes en langage humain au moyen du TLN.

Traduction automatique : désigne la traduction d'un texte (ou d'une conversation audio, en direct ou en différé) entièrement réalisée par un ou plusieurs programmes informatiques, sans qu'un traducteur humain n'ait à intervenir. On la distingue de la traduction assistée par ordinateur où la traduction est en partie manuelle, éventuellement de façon interactive avec la machine.

Systèmes de dialogue : est un système d'informatique destiné à l'humain pour faire la communication à travers les conversations ou discours. Les systèmes de dialogue utilisaient un ou plusieurs des modes comme texte, parole, graphiques, gestes.

Les domaines du Traitement du Langage Naturel que nous avons vu ci-dessus il y a d'autre domaine comme l'Analyse des Sentiments auquel nous intéressons tout particulièrement. Il va faire l'objet des sections suivantes de ce chapitre.

3 Analyse des Sentiments

Dans cette section nous allons parler de L'Analyse des Sentiments aussi appelée Opinion Mining.

3.1 Définition

C'est un domaine du NLP et une interprétation et la classification des sentiments (positives, négatives et neutres) dans les données de texte à l'aide de techniques d'analyse de texte. L'Analyse des Sentiments permet de déterminer le sentiment qui se cache derrière une série de mots. Par exemple il permet aux entreprises d'identifier le sentiment des clients envers les produits, les marques ou les services dans les conversations et les commentaires .La figure suivante (Figure 1) illustre une Analyse des Sentiments par des émojis :

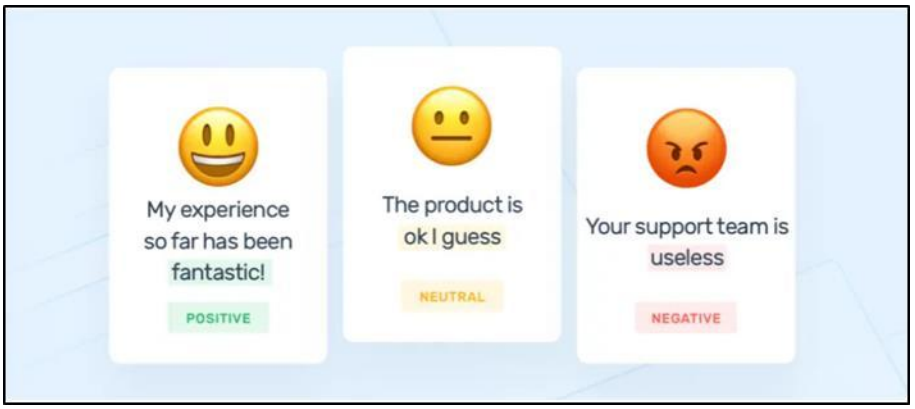

Figure 1: L'Analyse des Sentiments (Positive, Négative, Neutre) pour une entreprise

3.2 Types d'Analyse des Sentiments

L'Analyse des Sentiments prend diverses formes, des modèles qui se concentrent sur la polarité (positif, négatif, Neutre) détectent les sentiments et les émotions (en colère, heureux, triste, etc.), ou même des modèles qui les intentions (par exemple, intéressé, contre non intéressé).

Voici quelques-uns des types d'analyses de sentiments les plus populaires [2] :

5.1.1 Analyse fine des sentiments

Si le degré de précision d'une polarité est important pour une entreprise, elle peut par exemple envisager d'élargir les catégories de polarité pour une meilleure analyse: très positif, positif, neutre, négatif, très négatif

Ceci est généralement appelé Analyse des Sentiments à grain fin et peut être utilisé pour interpréter les notes en 5 étoiles dans un avis, par exemple : très positif (5) vs. Très négatif (1)

5.1.2 Détection d'émotion

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.

5.1.3 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.

3.3 Niveaux d'Analyse des Sentiments

L'Analyse des Sentiments est faite selon des niveaux est selon l'objectif [3] :

3.3.1 Au niveau du document

Détermine la polarité d'un texte entier. L'hypothèse est que le texte n'exprime qu'une seul opinion sur une seule entité (par exemple, un seul produit)

3.3.2 Au niveau de la phrase

Détermine la polarité de chaque phrase contenue dans un texte. L'hypothèse est que chaque phrase dans le texte exprime une opinion unique sur une entité unique.

3.3.3 Au niveau des aspects

Effectue une analyse plus fine que les autres niveaux. Il est basé sur l'idée qu'une opinion consiste d'un sentiment et une cible (d'opinion). Par exemple, la phrase «L'iPhone est très bon, mais il faut encore travailler sur la durée de vie de la batterie et les problèmes de sécurité» évalue trois aspects : iPhone (positif), la durée de vie de la batterie (négative) et la sécurité (négative).

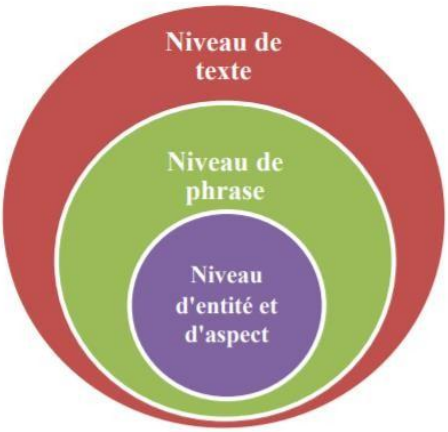

Figure 2: Les niveaux d'Analyse des Sentiments

3.4 Le fonctionnement de l'Analyse des Sentiments

L'Analyse des Sentiments est une technologie basée sur le traitement automatique du langage humain (appelé plus communément Traitement du Language Naturel).

L'Analyse des Sentiments va de la détection des émotions (colère, bonheur, peur) au sarcasme et à l'intention (p. ex. plaintes, rétroaction, opinions). Dans sa forme la plus simple, l'Analyse des Sentiments attribue ensuite une polarité (positive, négative, neutre) à un texte [4].

3.5 Processus général de l'Analyse des Sentiments

Dans cette section nous illustrons l'architecture générale de l'Analyse des Sentiments présenté dans la figure au-dessous (Figure 3)

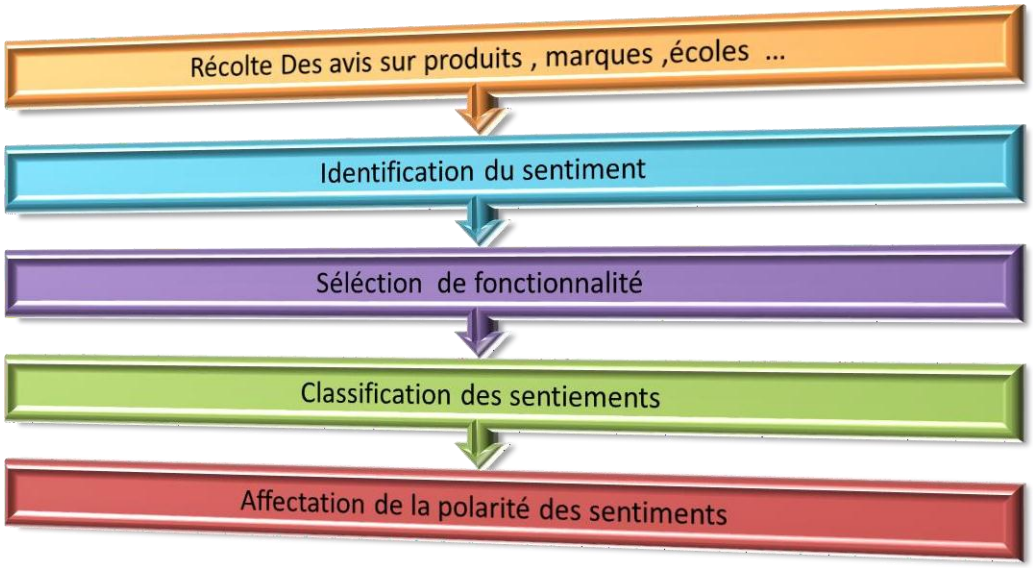

Figure 3: Processus d'Analyse des Sentiments

3.6 Quelques outils de l'Analyse des Sentiments

Il existe des outils permettant d'identifier le sentiment dégagé par un texte. Voici une liste non exhaustive des outils les plus connus :

3.6.1 Sentiment140

Sentiment140 (anciennement connu sous le nom "Twitter sentiment") est un outil en ligne gratuit qui a été créé par trois étudiants en computer science de Stanford, Il s'agit d'un projet académique. Cet outil, contrairement à la plupart des autres sites d'Analyse des Sentiments, n'utilise pas de listes de mots positifs ou négatifs mais est fondé sur les algorithmes d'apprentissage automatique [6].

Sentiment140 permet de découvrir des sentiments des Tweets d'une marque, un produit ou un sujet sur Twitter.

3.6.2 Tweetfeel

Tweetfeel est un service qui s'appuie sur les capacités temps réels de Twitter pour donner le sentiment des utilisateurs de Twitter sur un mot clé, une marque ou encore une star. L'évaluation

de Tweetfeel se fait sur la base de présence de mots clés précis dans les tweets tels que Good, Bad, etc... (Uniquement anglais pour l'instant). Ensuite un pourcentage est calculé selon le nombre de tweets positifs ou négatifs pour avoir un sentiment global de Twitter sur la marque [7].

3.6.3 Twitrratr

Twitrratr est un outil en ligne gratuit, qui a émergé à partir d'un projet Startup Weekend. Twitrratr fonctionne à partir d'une liste de mots positifs et d'une liste de mots négatifs. Cet outil classe une opinion sur le mot clé de la requête s'il est capable de le croiser avec un mot d'une des deux listes. Les mots positifs et négatifs qui servent à classer les tweets sont surlignés dans l'interface [8].

3.6.4 Tweet Sentiments Analyses

Tweet Sentiments Analyses est un outil en ligne gratuit et open source d'analyse du sentiment sur Twitter. Il peut donner des sentiments positifs, négatifs et neutres des tweets sur le mot clé lancé dans la requête. Il peut travailler sur 12 langues. Il donne les résultats sous forme graphique [9].

3.7 Les Avantages de l'Analyse des Sentiments

On estime que 80% des données mondiales ne sont pas structurées, c'est-à-dire qu'elles ne sont pas organisées. D'énormes quantités de données textuelles (e-mails, tickets d'assistance, chats, conversations sur les réseaux sociaux, sondages, articles, documents, etc.) sont créées chaque jour, mais il est difficile d'analyser, de comprendre et de trier, sans parler de temps et d'argent.

Cependant, l'Analyse des Sentiments aide les entreprises à comprendre tout ce texte non structuré en le marquant automatiquement [5].

Les avantages de l'Analyse des Sentiments comprennent :

- Traitement des données à grande échelle peut-on imaginer de trier manuellement des milliers de Tweets, conversations de support client ou avis clients ? Il y a tout simplement trop de données à traiter manuellement. L'Analyse des Sentiments aide les entreprises à traiter d'énormes quantités de données de manière efficace et rentable.
- Analyse en temps réel L'Analyse des Sentiments peut identifier les problèmes critiques en temps réel, par exemple une crise de relations publiques sur les réseaux sociaux s'intensifie-t-elle ? Un client en colère est-il sur le point de se retourner ? Les modèles d'Analyse des Sentiments peuvent aider à identifier immédiatement ce type de situations, afin que d'agir agir immédiatement.

Critères cohérents On estime que les gens ne s'entendent que 60 à 65% du temps pour déterminer le sentiment d'un texte particulier. Le marquage de texte par sentiment est très subjectif, influencé par des expériences personnelles, des pensées et des croyances. En utilisant un système centralisé d'analyse des sentiments, les entreprises peuvent appliquer les mêmes critères à toutes leurs données, ce qui les aide à améliorer la précision et à obtenir de meilleures informations.

4 L'Analyse des Sentiments en langue Arabe

La majorité des recherches dans le domaine ont été effectuée en langue Anglaise, car c'est la langue qui domine le monde de la science actuellement. Récemment, quelques chercheurs se sont intéressés à l'application de l'Analyse des Sentiments dans d'autres langues, une de ces langues étant l'Arabe. La figure ci-dessous montre la différence entre la recherche qui a été menée en Arabe et en Anglais [10].

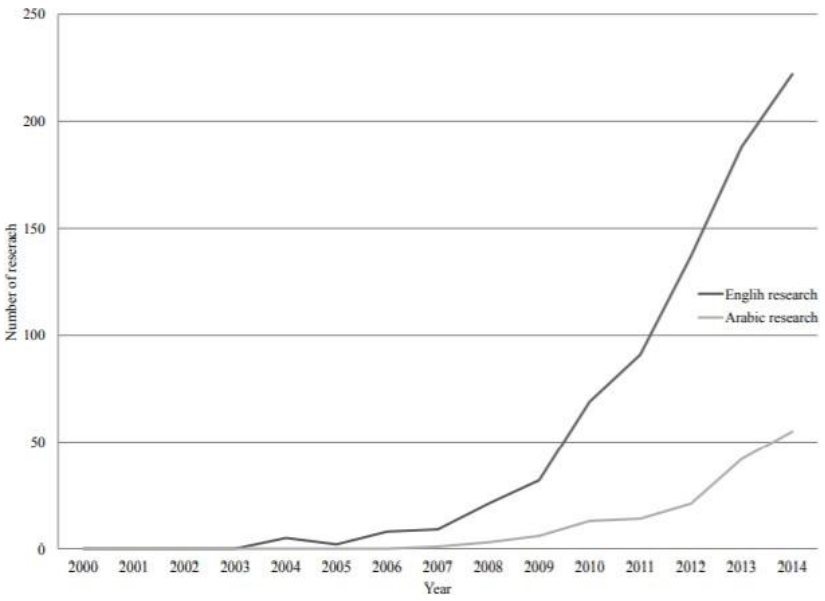

Figure 4: La différence de la recherche entre l'arabe et l'anglais

Ces données collectées à l'aide de mots clés pertinents dans le champ d'Analyse des Sentiments dans les deux langues.

Il est clair qu'il existe un grand écart entre le travail réalisé en arabe et l'anglais, cela peut être dû aux limitations des outils ou des ressources du NLP de l'arabe.

En outre, il peut révéler que l'arabe nécessite un traitement spécial en raison de sa complexité nature et structure.

4.1 Définition

L'Analyse des Sentiments est l'un des domaines du traitement du langage naturel, dédié à l'exploration d'opinions ou de sentiments subjectifs recueillis auprès de diverses sources sur un sujet particulier.

Pour chaque linguiste , on trouve une définition de l'Analyse des Sentiments. Nous avons regroupé quelques-unes ci-après [10] :

Selon, Sharda et. al : « c'est une technique utilisée pour détecter des opinions favorables et défavorables à l'égard de produits et services spécifiques en utilisant un grand nombre de sources de données textuelles »

L'arabe est actuellement classé comme la quatrième langue utilisée sur le Web, et il y a environ 168 millions d'internautes arabes.

La plupart des travaux d'Analyse des Sentiments se concentrent sur la langue anglaise et il y a la complexité de la langue arabe et le manque de ressources disponibles pour l'Analyse des Sentiments en arabe comme les lexiques et les ensembles de données sont les principaux obstacles à l'analyse de sentiment Arabe.

L'Analyse des Sentiments en langue arabe (ASA) a été nécessaire depuis que le public arabe, qui utilise Internet et beaucoup d'applications, a augmenté [7].

La figure ci-dessous, présente clairement que le nombre de recherches sur l'ASA a augmenté progressivement au cours des cinq dernières années ; c'est-à-dire qu'il a atteint 53 articles en 2018.

Il est évident que l'ASA est récemment un sujet de recherche absolument d'actualité. Par conséquent, il a connu une croissance rapide et un intérêt accru de la part des chercheurs.

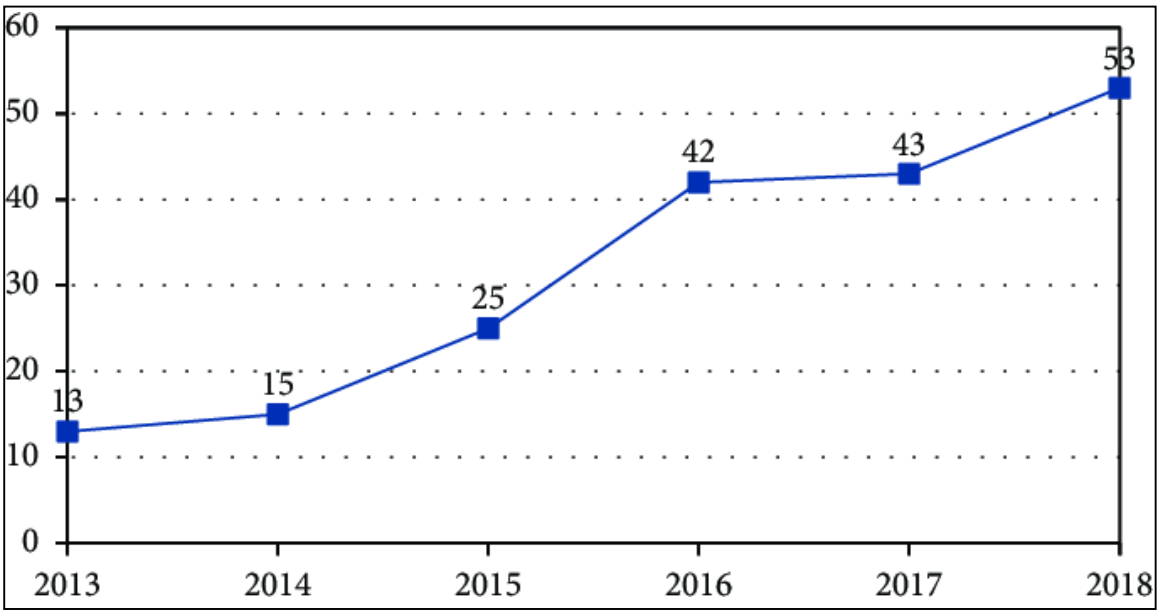

Figure 5: Distribution des articles examinés par l'ASA sur des années

Selon les études examinées, la plupart des tâches utilisées dans l'ASA sont réparties en [11] : Classification des sentiments et le renforcement des ressources

De plus, il ressort que la tâche de construction de la classification des ressources et des sentiments domine par rapport aux autres tâches. En effet, la langue arabe manque encore de ressources et d'outils qui peuvent être utilisés pour soutenir la classification de Sentiment, tous cela présentés dans la figure au-dessous

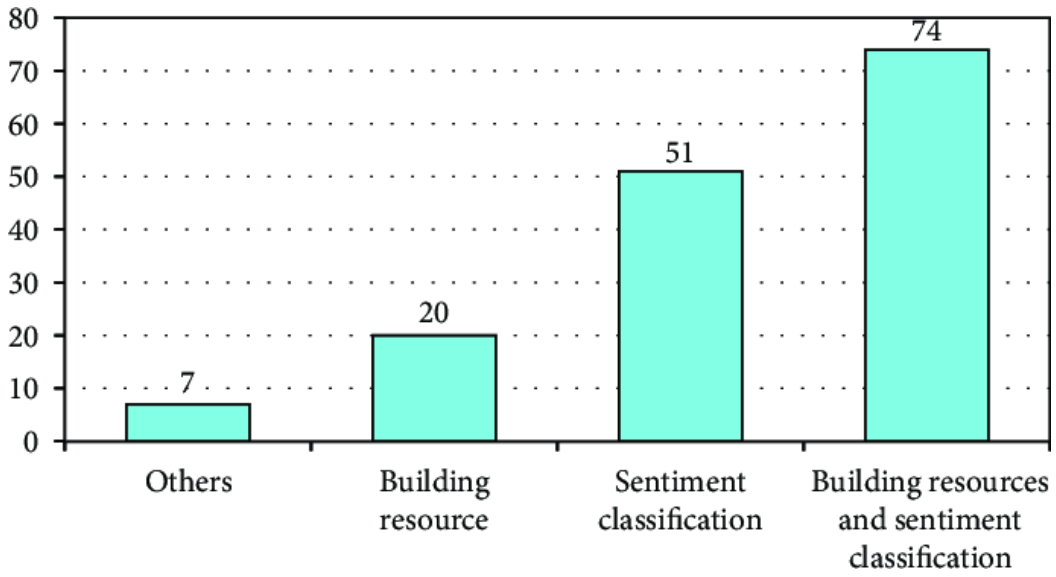

Figure 6: Nombre d'articles ciblant les tâches d'ASA

4.2 La langue Arabe

L'Arabe est la langue officielle de 22 pays, soit plus de 300 millions de locuteurs natifs. Le taux de croissance (soit 8,616.0 %) des internautes arabes a été classé le plus rapide en 31 décembre 2017 par les statistiques mondiales de l'Internet (Internet world stats) 1 par rapport à 3,434.0 % pour le japonais, 647.9 % pour l'anglais et 152.0 % pour le français. Les utilisateurs arabes représentent 50.3 % du monde, plus de219 041 264 d'utilisateurs d'Internet (Figure 5) [34].

Les dix langues les plus utilisées sur le Web - Décembre 31, 2017 (Nombre d'internautes par langue)					
LES DIX LANGUES LES PLUS UTILISEES SUR INTERNET	Population mondiale pour cette langue (estimation 2018)	Utilisateurs d'Internet par langue	Pénétration d'Internet (% de la population)	Croissance du nombre d'internautes (2000 - 2018)	Utilisateurs d'Internet % du total mondial (participation)
Anglais	1,462,008,909	**1,052,764,386**	72.0 %	647.9 %	25.3 %
Chinois	1,452,593,223	**804,634,814**	55.4 %	2,390.9 %	19.4 %
Espagnol	515,759,912	**337,892,295**	65.5 %	1,758.5 %	8.1 %
Arabe	435,636,462	**219,041,264**	50.3 %	8,616.0 %	5.3 %
Portugais	286,455,543	**169,157,589**	59.1 %	2,132.8 %	4.1 %
Indonésien / Malaisien	299,271,514	**168,755,091**	56.4 %	2,845.1 %	4.1 %
Français	127,185,332	**118,626,672**	93.3 %	152.0 %	2.9 %
Japonais	143,964,709	**109,552,842**	76.1 %	3,434.0 %	2.7 %
Russe	405,644,599	**108,014,564**	26.6 %	800.2 %	2.8 %
Allemand	94,943,848	**84,700,419**	89.2 %	207.8 %	2.2 %
TOP 10 LANGUES	5,135,270,101	**3,206,613,856**	62.4 %	1,091 %	77.1 %
Reste des langues	2,499,488,327	**950,318,284**	38.0 %	935 %	22.9 %
TOTAL MONDIAL	7,634,758,428	**4,156,932,140**	54.4 %	1,051 %	100.0 %

Figure 7: Les dix langues les plus utilisés dans le web

La langue Arabe est incluse dans la figure ci-dessus parmi les dix premières langues d'internet et elle est citée comme quatrième langue d'internet utilisée

La MSA est la langue officielle du monde arabe et elle est syntaxiquement, morphologiquement et phonologiquement basée sur l'arabe classique (CA Classical Arabic). L'arabe classique est la langue du Coran. Alors que les dialectes arabes sont de véritables formes de langue maternelle, ils sont utilisés dans la communication informelle quotidienne et ne sont pas enseignés dans les écoles ou standardisés. Dans le contexte des dialectes, la MSA est généralement une langue écrite et non parlée. Les dialectes arabes sont peu proches à l'arabe classique [34]. Il y a beaucoup de dialectes arabes, et ils sont différents dans de nombreux aspects, principalement la géographie et les classes sociales. Une façon de diviser les dialectes Arabes est basée sur l'aspect géographique comme suit (Figure 8):

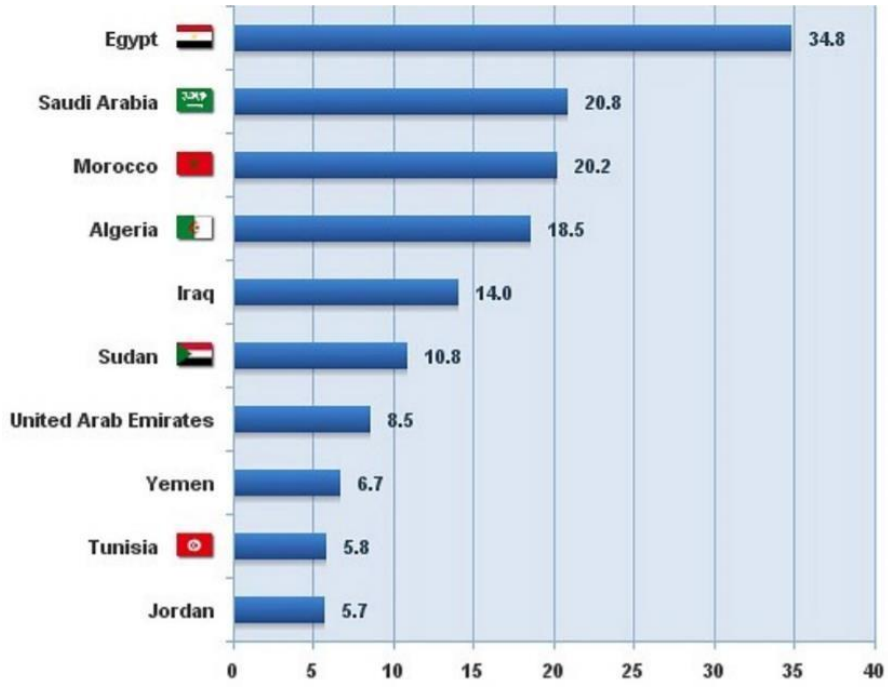

Figure 8: Les différents dialectes Arabes

Le dialecte le plus courant est l'arabe égyptien, qui couvre la vallée du Nil (Égypte et Soudan).

♣ L'arabe du Golfe comprend les dialectes des pays du Golfe (Emirats Arabes Unis, Arabie Saoudite, etc.).

♣ L'arabe levantin couvre les dialectes de la Syrie, du Liban, de la Jordanie, de la Palestine et d'Israël.

♣ L'arabe maghrébin couvre les dialectes d'Algérie, de Tunisie et du Maroc.

♣ L'arabe irakien couvre l'Irak et combine des éléments des dialectes du Levantin et du Golfe.

Chaque groupe dialectal est complètement homogène du point de vue linguistique. L'arabe est une langue sémitique qui possède un système d'inflexion très riche et est considérée comme l'une des langues les plus riches en termes de morphologie [Habash 2009]. Les formes de phrases arabes sont divisées en deux types ; les constructions nominales 37 et 37 verbales [Farra et al. 2010]. Dans le domaine verbal, l'arabe a deux modèles d'ordre des mots (sujet-verbe-objet et verbe-sujet-objet). Dans le domaine nominal, un modèle normal consisterait en deux mots consécutifs, un nom (c'est-à-dire un sujet) puis un adjectif (descripteur du sujet) [34].

4.3 Défis de la langue Arabe

L'Arabe est l'une des dix langues officielles des Nations Unies et c'est une langue morphologiquement riche. Cette langue a deux formes principales, standard et dialectales. Modern Standard Arabic (MSA) est utilisé dans les discours et l'écriture formels comme les livres et les journaux tandis que l'arabe dialectal est utilisé dans l'écriture informelle spécialement dans les médias sociaux et il varie d'un pays à l'autre. L'analyse du texte arabe est très compliquée, les points suivants expliquent la complexité de l'application de l'Analyse des Sentiments à la langue arabe [10] :

- Chaque pays arabe a son propre dialecte qui varie d'un autre pays, et les gens ont tendance à utiliser leur dialecte au lieu d'utiliser le MSA
- Le même mot peut être écrit par différents utilisateurs de différentes façons comme les mots qui se terminent par Ta 'marbootah (ة), par exemple (مؤنثة) peut s'écrire (مؤنثه)
- Les mots de négation peuvent inverser le sens de la phrase, donc si la phrase avait un sentiment positif, après avoir ajouté le mot de négation dans la phrase, elle aura eu un sentiment négatif. Dans le Traitement du Langage Naturel (NLP)

- Le même verbe peut être écrit de différentes manières en fonction du sujet singulier ou pluriel, féminin ou masculin par exemple "هو يحب السيارات"(Il aime les voitures) "هي تحب سيارات" (elle aime les voitures)
- Certains noms arabes sont dérivés d'adjectifs, et le nom lui-même n'a pas de sentiment tandis que l'adjectif peut avoir un sentiment. Par exemple, le nom Jameelah et l'adjectif Beautiful ont la même forme en arabe (جميلة)
- Les arabophones pourraient utiliser des idiomes pour exprimer leurs opinions, et ces idiomes ont une opinion implicite. Par exemple (حسبي الله و نعم الوكيل) est une opinion négative alors qu'il n'a pas de mot négatif

4.4 Les récents travaux réalisés sur l'Analyse des Sentiments en Arabe

La littérature arabe SA a de nombreuses tentatives pour s'attaquer au problème, et la plupart du travail est basé sur des algorithmes d'apprentissage automatique conventionnels avec peu de tentatives d'utilisation du deep learning. Parmi les travaux connus [23] :

- Dans un autre travail (Shoeb et Ahmed, 2017), les auteurs ont appliqué SA sur des Tweets en utilisant Naive Bayes (NB) et K-NN, ils ont obtenu des résultats relativement bons.
- (Al-Ayyoub et coll. 2015) ont également créé un grand lexique de termes arabes extraits d'articles de presse. Sur la base de leur lexique, ils ont construit un système SA et l'ont testé sur des données collectées sur Twitter.
- Dans (Elmasry et al., 2014), les auteurs visaient à s'attaquer au problème des dialectes. Ils ont construit un lexique des mots et des idiomes sentimentaux d'argot (SSWIL) et ont mené des expériences en utilisant SVM et le nouveau lexique.
- les travaux de (Abdulla et al., 2013) ont introduit un ensemble de données de 2000 Tweets, que les auteurs ont utilisé pour mener une expérience avec des systèmes basés sur le lexique et ML. Ils ont constaté que la combinaison des deux approches permettrait d'obtenir de meilleurs résultats.
- (Abdul-Mageed et coll. 2014) ont proposé un système de SA pour les médias sociaux. Dans leur travail, ils ont expérimenté et étudié une grande variété de fonctionnalités. Ils ont également étudié l'effet des dialectes et la richesse morphologique de l'arabe.
- De plus, dans (Abdul-Mageed, 2017a, b), les auteurs ont étudié les différentes manières de gérer la richesse morphologique arabe pour SA. Ils ont étudié l'effet de la segmentation

dans la représentation de l'entrée lexicale, ils ont également essayé d'étudier le poids et l'importance de ces segments pour l'AS.

- (El-Beltagy et coll. 2017) ont été classés premiers dans la tâche SemEval 2017 pour Arabic SA. Ils ont utilisé un ensemble de fonctionnalités conçues à la main et basées sur le lexique .
- Le deuxième rang dans la même tâche était pour le travail de (Jabreel et Moreno 2017), qui ont introduit un riche ensemble de fonctionnalités principalement basées sur le modèle sac de mots en plus de certaines fonctionnalités extraites de l'intégration de mots.
- Dans (Alayba et al., 2017), les auteurs ont présenté leur propre ensemble de données SA des opinions sur les services de santé. Ils ont construit un système SA et il a été testé sur le nouvel ensemble de données. Leurs expériences comprenaient l'utilisation de nombreux algorithmes ML, y compris les CNN.
- (Al-Sallab et coll. 2015) ont expérimenté différents modèles d'apprentissage profond tels que l'auto encodeur récursif (RAE), les réseaux de croyances profondes (DBN) et encodeur automatique profond (DAE). Ils se sont appuyés sur le lexique ArSenL (Badaro et al., 2014) pour construire les vecteurs de caractéristiques. Dans (Al-Smadi et al., 2017b), les auteurs ont abordé l'Analyse des Sentiments basée sur les aspects (ABSA). Dans leurs expériences, ils ont utilisé des RNN et SVM comme Méthodes, les résultats ont montré que SVM était supérieur.
- (Alayba et coll. 2018) construit un système SA basé sur une combinaison de CNN et LSTM. Ils ont testé leur modèle sur deux ensembles de données, les ensembles de données Ar-Twitter et Arabic Health Services, où ils ont obtenu des précisions de 88,1% et 94,3% respectivement.
- Dans (Al-Smadi et al., 2018), les auteurs ont proposé une système d'analyse des sentiments, leur modèle est basé sur un Bi-LSTM et un champ aléatoire conditionnel (CRF).Ils ont testé leur modèle sur l'ensemble de données des avis des hôtels arabes, ils ont obtenu un score F de 70%.
- (Elshakankery et Ahmed2019) ont proposé un système hybride pour Arabic SA, qui utilise des approches basées sur le lexique et l'apprentissage automatique. Dans leur travail, ils ont expérimenté plusieurs ensembles de données tels qu'ASTD et ArTwitter , ils ont utilisé différents Méthodes pour la tâche, qui varie de l'utilisation de l'apprentissage automatique conventionnel à modèles d'apprentissage

Nous ne sommes pas au courant de tous outil open-source publié pour Arabic SA, considérée comme l'une des plus grandes limitations en NLP arabe. Bien qu'il y en ait beaucoup d'outils NLP en arabe pour diverses tâches, y compris la segmentation, le marquage POS et la diacritisation (Pasha et al., 2014; Abdelali et al., 2016), la communauté de recherche pour la NLP arabe manque toujours d'un outil pour l'analyse des sentiments.

5 L'Analyse des Sentiments sur les Réseaux Sociaux

L'Analyse des Sentiments a reçu beaucoup d'attention non seulement de la part de la recherche scientifique mais aussi par les domaines de la publicité et du marketing. Le développement du web 2.0 a entraîné un intérêt de ces domaines pour les équipes marketing, souvent soumises à un déluge de données. La solution se connecte d'elle-même aux différents réseaux sociaux (Facebook, Twitter, LinkedIn, etc.), blogs, forums, commentaires d'articles, etc. Au fur et à mesure de l'indexation des données trouvées, la solution d'Analyse des Sentiments détermine par un système de notation si le contenu global recueilli est positif, négatif ou neutre. Elle peut même être capable d'identifier des propos sarcastiques ou ironiques. [18]

Aussi la rapidité du relais de l'information, les grandes masses de données réelles issues des réseaux sociaux sont largement utilisées pour l'analyse des sentiments. Analyser les messages récents issus des réseaux sociaux pourrait donner l'opinion générale des utilisateurs envers un sujet spécifique [19].

Voici comment Nick Martin, expert mondial de l'engagement sur les médias sociaux chez Hootsuite, définit le sentiment sur les médias sociaux : « Le sentiment sur les médias sociaux correspond à l'impression positive ou négative qui se dégage d'une publication ou d'une interaction» [20].

5.1 Réseaux Sociaux

Dans le domaine des technologies, un réseau social consiste en un service permettant de regrouper diverses personnes afin de créer un échange sur un sujet particulier ou non. En quelque sorte, le réseau social trouve ses origines dans les forums, groupes de discussion et salons de chat introduits dès les premières heures d'Internet .

Les réseaux sociaux en ligne désignent les sites Internet et applications mobiles qui permettent aux utilisateurs de se constituer un réseau d'amis ou de relations, et qui favorisent les interactions sociales entre individus, groupes d'individus ou organisations. Le réseautage social, ou social networking, désigne l'utilisation des réseaux sociaux [15].

Les réseaux sociaux les plus connus sont Facebook, Twitter, Linkedin, Viadeo,Pinterest, etc. Youtube peut également être considéré partiellement comme un réseau social dans la mesure où le service a développé des outils d'interactions entre ses membres [16].

5.1.1 Le Réseau Social Twitter

Pour notre part nous avons opté poursur le Réseau Social Twitter .

Twitter est un réseau social de microblogage géré par l'entreprise Twitter Inc. Il permet à un utilisateur d'envoyer gratuitement de brefs messages appelés Tweets sur internet, par messagerie instantanée ou par SMS, limités à 280 caractères. Twitter a été créé le 21 mars 2006 par Jack Dorsey, Evan Williams, Biz Stone et Noah Glass à San Francisco. Le service en ligne est rapidement devenu populaire.

Il compte 313 millions d'utilisateurs actifs par mois, 500 millions de Tweets envoyés par jour et est disponible dans plus de quarante langues [21].

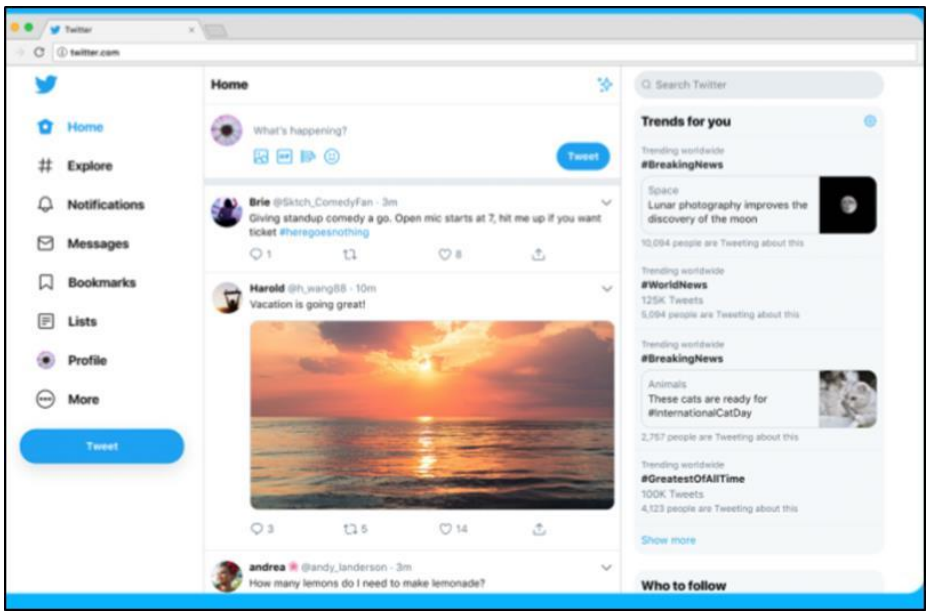

Figure 9: La page d'actualité de Twitter

5.1.2 Les Caractéristiques du Tweet

Le Tweet est un court message, il est limité à l'origine à 140 caractères (jusqu'en septembre 2016). Il contraint les utilisateurs à être concis dans leur rédaction. Initialement, Twitter pouvait être utilisé par l'intermédiaire des SMS. Ceux-ci étant limités à 160 caractères, Twitter prend cette limite et conserve 20 caractères pour ajouter son nom d'utilisateur [21].

Figure 10 : Exemple d'un Tweet

5.2 Twitter pour l'Analyse des Sentiments

Nous avons choisi le Twitter comme un terrain et objet d'étude, car le Twitter génère une grande quantité de données riches de sentiments sous la forme de Tweets exprimé par l'utilisateur, et leur collecte et traitement automatique via l'API du service est facile.

De plus, Twitter présente des avantages intéressants comme la brièveté des Tweets (140 caractères) ainsi que sa réactivité. Aussi, Twitter est ouvert et les textes qui y sont soumis et sont accessibles à tous grâce à un service web ce qui facilite l'exploitation des données.

D'après les chercheurs et linguistes du monde, Twitter présente 50% des sources de données utilisées dans les articles et Twitter est l'applicationla plus fréquemment utilisée des médias sociaux et les articles examinés.

Il a un grand potentiel pour explorer la vie des gens ainsi que leurs opinions et intérêts [2].

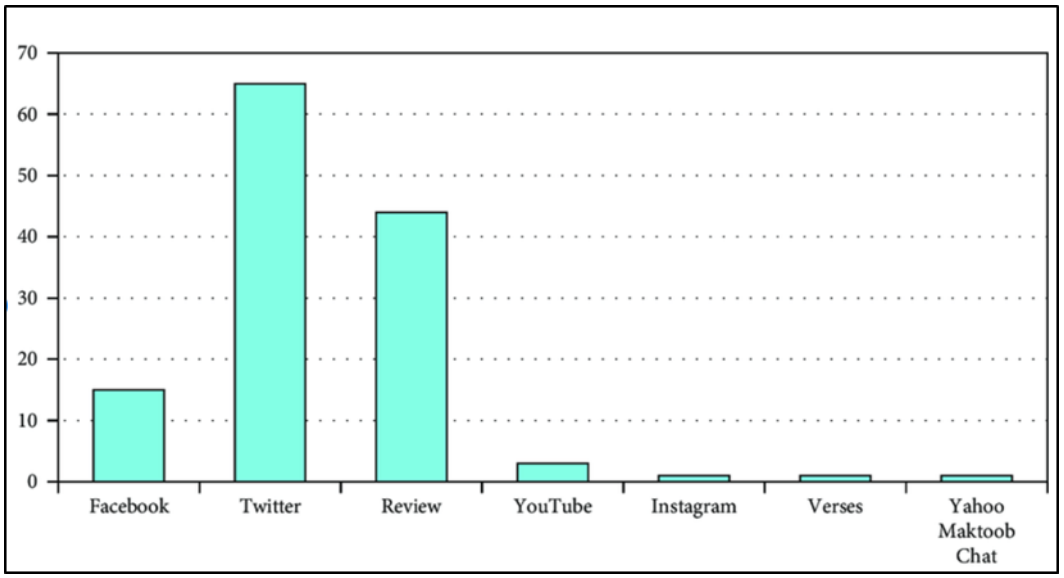

Figure 11: Les sources de l'ensemble de données utilisées dans l'Analyse des Sentiments

Grâce aux avancées de l'apprentissage automatique et de la NLP, il est désormais possible de créer des modèles qui apprennent à partir d'exemples et peuvent être utilisés pour traiter et organiser des données de texte [2].

5.3 Les avantages de l'Analyse des Sentiments sur Twitter

Nous résumons ci-dessous, les différents avantages de l'Analyse des Sentiments sur le Twitter dans ces points [22] :

Évolutivité : supposons que l'on voudrait analyser des centaines de Tweets mentionnant une marque. Bien que ça puisse être fait manuellement, cela prendrait des heures et des heures de traitement manuel et finirait par être incohérent et impossible à mettre à l'échelle. En effectuant une Analyse des Sentiments sur Twitter, on peut automatiser cette tâche et obtenir des résultats rentables en très peu de temps.

Analyse en temps réel : l'Analyse des Sentiments sur Twitter est essentielle pour remarquer les changements soudains de l'humeur des clients, détecter si les critiques et les plaintes augmentent et prendre des mesures avant que le problème ne s'aggrave. On peut surveiller la marque en temps réel et obtenir des informations précieuses qui vont permettre d'apporter des modifications ou des améliorations en cas de besoin.

Les Critères cohérents : dans un texte l'Analyse des Sentiments est une tâche subjective. Une fois fait manuellement, le même Tweet peut être perçu différemment par deux membres de la même équipe, et les résultats seront probablement biaisés. En formant un modèle d'apprentissage automatique pour effectuer une Analyse des Sentiments sur Twitter, on peut définir les paramètres pour analyser toutes les données et obtenir des résultats plus cohérents et précis.

6 Conclusion

D'après ce que nous avons parcouru dans ce chapitre, nous concluons que l'Analyse des Sentiments et plus particulièrement l'Analyse des Sentiments pour la langue Arabe est un sujet d'actualité dans le domaine NLP. Elle est considérée comme l'un des domaines grandissants qui concernent la découverte des sentiments des individus envers un événement, une marque ou autre chose.

Dans le prochain chapitre nous allons discuter sur le processus de l'Analyse des Sentiments et ses différentes approches.

Chapitre II :

Les approches du l'Analyse des Sentiments

1 Introduction

Dans la littérature 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.

Pour cela il existe de nombreuses méthodes et algorithmes pour mettre en œuvre des systèmes d'analyse des sentiments, que l'on peut classer en trois catégories : l'approche de Lexicon, l'approche de Machine Learning, l'approche Hybride.

Dans ce chapitre nous commençons par la présentation du processus de construction de l'Analyse des Sentiments pour la langue arabe et par la suite nous allons explorer les différentes approches trouvées pour ce processus.

2 Processus de l'Analyse des Sentiments pour la langue Arabe

L'Analyse des Sentiments est considérée comme des techniques d'exploration de texte qui peut être utilisées pour détecter des opinions favorables et défavorables à l'égard de services ou de produits spécifiques, et elle peut être appliquée à différents domaines tels que la gestion de marque, la politique, les soins de santé, ou dans le filtrage des e-mails en priorisant les E-mails reçus [6].

L'Analyse des Sentiments pour l'arabe en est encore à ses débuts et la recherche a augmenté rapidement au cours des dernières années, depuis 2014, 2015. Nous avons parcouru presque toutes les recherches d'analyse du sentiment arabe et les références dans les enquêtes d'Analyse des Sentiments des Tweets arabe montrent des rapports statistiques sur les littératures collectées, et comme nous l'avons vu dans le chapitre précédent, les recherches ont commencé à partir de 2006 et elles ont rapidement augmenté ces dernières années. Et selon l'enquête l'Analyse des Sentiments arabes est toujours un domaine de recherche ouvert.

Le processus de construction du module d'Analyse des Sentiments comprend en général quatre étapes principales : L'ensemble de données, prétraitement, classification des sentiments et l'évaluation du système.

2.1 L'ensemble de données (Dataset)

Un ensemble de données (ou Dataset) est une collection de données. Il correspond à une ou plusieurs tables de base de données, Dans le cas des données tabulaires, où chaque colonne d'une table représente une variable particulière, et chaque ligne correspond à un enregistrement donné de l'ensemble de données.

Pour ce domaine le Dataset est une collection de données collectées d'après le Twitter. Pour chaque donnée de l'ensemble de données doit être affectée à une classe spécifique comme positive, négative ou neutre durant le processus d'annotation [10].

Il existe deux approches utilisées pour le processus d'annotation : la première est l'approche d'annotation manuelle où deux ou plusieurs annotateurs classent chaque instance de l'ensemble de données manuellement et la seconde approche est l'approche de sourcing où une foule d'utilisateurs d'Internet classent les instances dans l'ensemble de données en utilisant une application Web créée par les auteurs [10]

2.2 Le Prétraitement du Dataset

Dans la deuxième étape de la construction du module d'Analyse des Sentiments permis de supprimer et nettoyer les Tweets ou les données du Dataset par l'utilisation des différents techniques comme la suppression des mots vides , la tokenisation , la normalisation, stemming etc .

Ces opérations de prétraitement seront appliquées à chaque instance de l'ensemble de données pour éviter le bruit du certains données et faciliter le traitement d'ASA.

De nombreuses techniques de prétraitement ont été utilisées dans ces étapes, les points suivants décrivent ces techniques :

- Normalisation du texte et suppression des lettres répétées : transformation de certaines lettres arabes en lettre générale (remplacé إ ,أpar ا) et suppression de la lettre répétée utilisée dans les revues pour intensifier les opinions comme (رااائع)

- Filtrage de texte : suppression de mots et de symboles qui n'ont aucun effet sur la sortie, comme certains mots vides, ponctuations et signes diacritiques, etc. Presque toutes les recherches ont appliqué cette technique [6].

- Stemming du Texte : remplacer le mot par sa racine. Cette technique a un grand impact en minimisant le stockage requis et en éliminant les termes redondants car il y a beaucoup de mots

arabes possédant la même racine avec laquelle les mots seront remplacés. Par exemple يكرهون ،يكره ،أكره ،نكره qui est كره donc tous ces mots seront remplacés par le mot كره.

Khoja Stemmer est une source ouverte pour dériver du texte arabe, et il a déjà été utilisé dans de nombreux ouvrages arabes

- Convertisseur de dialecte : mise en correspondance du mot d'argot à MSA.

-Négation : détection des phrases de négation et d'intensification dans l'ensemble de données. Cette technique est très importante pour l'extraction de sentiments afin de donner la polarité opposée en cas de négation et de donner plus de poids à la polarité actuelle en cas d'intensification, par exemple négation : هذا المطعم غير جيد le motغير reflètera la polarité de sortie sachant que le mot جيد obtiendra une polarité positive mais la sortie finale devra être négative

- Part-Of-Speech (POS): obtenir le type du mot tel que nom, verbe, adjectif, etc., et donner le plus poids du mot selon de ces types.

- Extraction d'entités : pour la formation de méthodes du Machine Learning, les données annotées doivent être converties en vecteur de fonction, et une combinaison d'entités spécifiques donnera une classe spécifique. La fonction unique est produite en convertissant un morceau de texte en fonction.

Dans les études d'analyse des sentiments, de nombreuses fonctionnalités ont été utilisées, les caractéristiques les plus courantes : fréquence de terme, uni-grammes, parties de discours (POS) et négation.

2.3 La classification des Sentiments (les Approches d'Analyse des Sentiments)

La troisième étape du processus d'Analyse des Sentiments dépend de l'approche utilisée pour la classification [10].

Dans le domaine de l'analyse des sentiments, les textes appartiennent à des classes positives ou négatives. Il peut également y avoir des classes à valeurs multiples ou plus comme positive, négative et neutre.

Les approches de classification des sentiments peuvent être catégorisées en l'approche: approche d'apprentissage automatique, approche de lexicon-based et l'approche de hybride chacune pouvant être classée en sous-catégories [10] comme il est illustré dans la Figure au-dessous (Figure 12) :

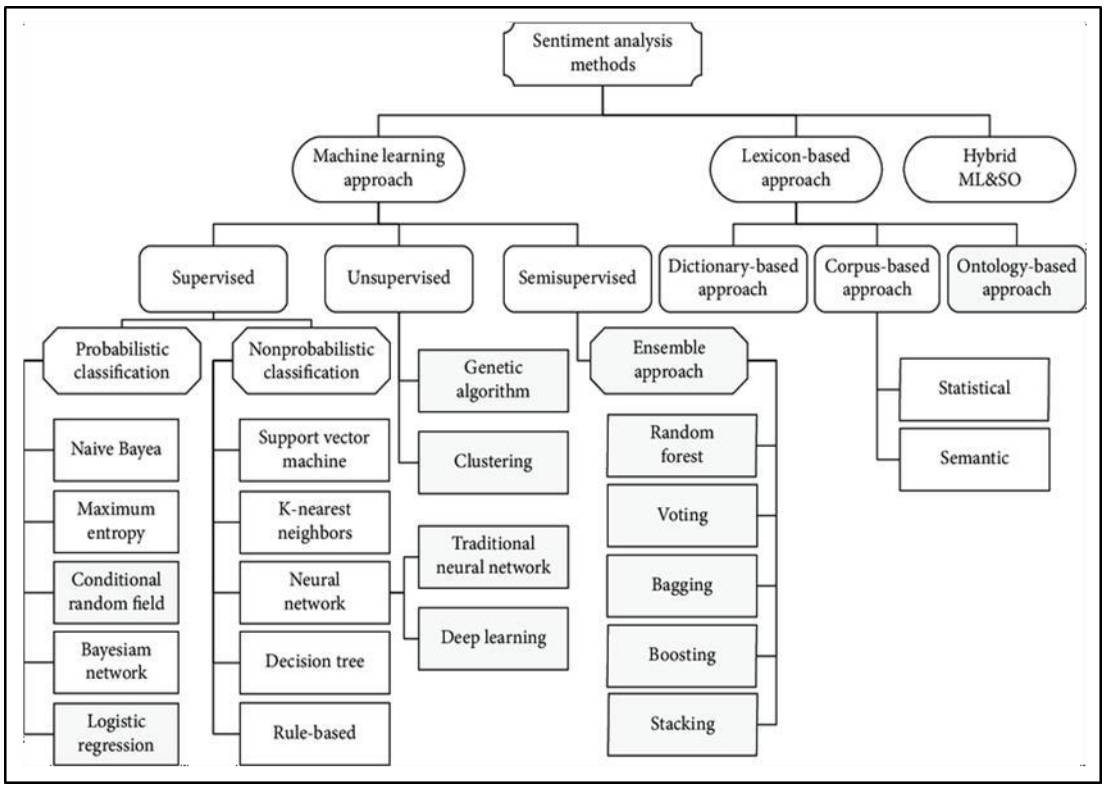

Figure 12: les différentes approches d'Analyse des Sentiments

2.3.1 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 [10].

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.

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

2.3.1.1 L'approche basée sur le dictionnaire

Cette approche fonctionne comme suit, à partir de l'ensemble initial de mots de sentiment avec une orientation positive et, négative connue. Exploitez ensuite les thésaurus et corpus disponibles comme WordNet pour trouver des synonymes et des antonymes pour chaque mot de la liste. Le mot nouvellement trouvé trouver est ajouté à la liste de départ et l'itération suivante commence. Le processus sera terminé lorsqu'aucun nouveau mot ne pourra être trouvé.

L'inconvénient majeur de cette approche est de ne pas trouver de mots d'opinion avec orientation de domaine, par exemple, si l'on dit: le haut-parleur du téléphone est silencieux, cela indique un avis négatif, mais si l'on dit: la voiture est silencieuse, cela indique une opinion positive.

2.3.1.2 L'approche basée sur le corpus

Cette approche repose sur des modèles statistiques ou syntaxiques par une liste de départ de mots d'opinion avec une polarité connue pour trouver de nouveaux mots de sentiment avec leur polarité dans un grand corpus.

Pour un modèle statistique, le nouveau mot de sentiment peut être trouvé par sa fréquence d'occurrence dans un grand corpus annoté, donc si le mot apparaît plus fréquemment dans les documents positifs que les documents négatifs, il sera ajouté à la liste de mots en tant que mot positif, et s'il apparaît plus fréquemment dans un document négatif puis il sera ajouté en tant que mot négatif. En d'autres termes, le mot sera ajouté en tant que mot positif s'il apparaît plus fréquemment dans un document positif, et il sera ajouté en tant que mot négatif s'il apparaît plus fréquemment dans un document négatif. Dans le modèle syntaxique, les mots ayant une opinion similaire apparaissent ensemble dans le corpus, et ce modèle suppose que si les mots apparaissent fréquemment ensemble dans les documents, ils ont probablement la même polarité. Ainsi, le mot qui n'a pas de valeur de polarité et qui apparaît fréquemment avec un autre mot avec une polarité connue doit avoir la même valeur de polarité ou une valeur de polarité opposée du mot connu. en fonction du mot connecté entre eux comme le mot (و) par exemple le mot (واسعة) dans cette phrase (السيارة مريحة وواسعة) aura la même valeur de polarité que le mot (مريحة). Pour construire et développer le lexique arabe, il y a deux façons de le faire dans la littérature: Manuelle et Automatique. Dans les techniques manuelles, le lexique est construit en traduisant le contenu de Senti-WordNet ou Senti-Strength, et pour chaque mot traduit, on doit trouver ses synonymes et ajouter le mot avec ses synonymes au lexique.

En technique automatique, le lexique est construit en partant du lexique collecté et annoté manuellement (lexique de base), puis en augmentant la taille du lexique en ajoutant des synonymes

et des antonymes. La méthode manuelle de création de lexique est plus précise que la méthode automatique, mais la méthode automatique demande moins de temps et de travail.

Par conséquent, de nombreux chercheurs ont proposé différentes méthodes et techniques pour surmonter ces inconvénients. Par exemple les chercheurs ont proposé de combiner les ressources disponibles: English Senti WordNet (ESWN), Arabe WordNet (AWN), et l'Analyseur Morphologique Arabe Standard (SAMA), la combinaison a été faite en fusionnant deux lexiques afin de créer une grande échelle Arabic Sentiment Lexicon (ArSenL). Le premier lexique a été créé en faisant correspondre AWN à ESWN et le deuxième lexique a été développé en faisant correspondre les lemmes du lexique SAMA à ESWN, ArSenL a 28780 lemmes et il est accessible au public. Un autre exemple où les auteurs ont présenté un lexique de sentiment des idiomes / proverbes arabes à grande échelle de MSA et familier pour les tâches d'analyse des sentiments, le lexique des idiomes / proverbes a été collecté et annoté manuellement, il contenait 3632 idiomes et proverbes, les auteurs ont prouvé que le lexique des idiomes / proverbes peut améliorer le processus de classification des sentiments.

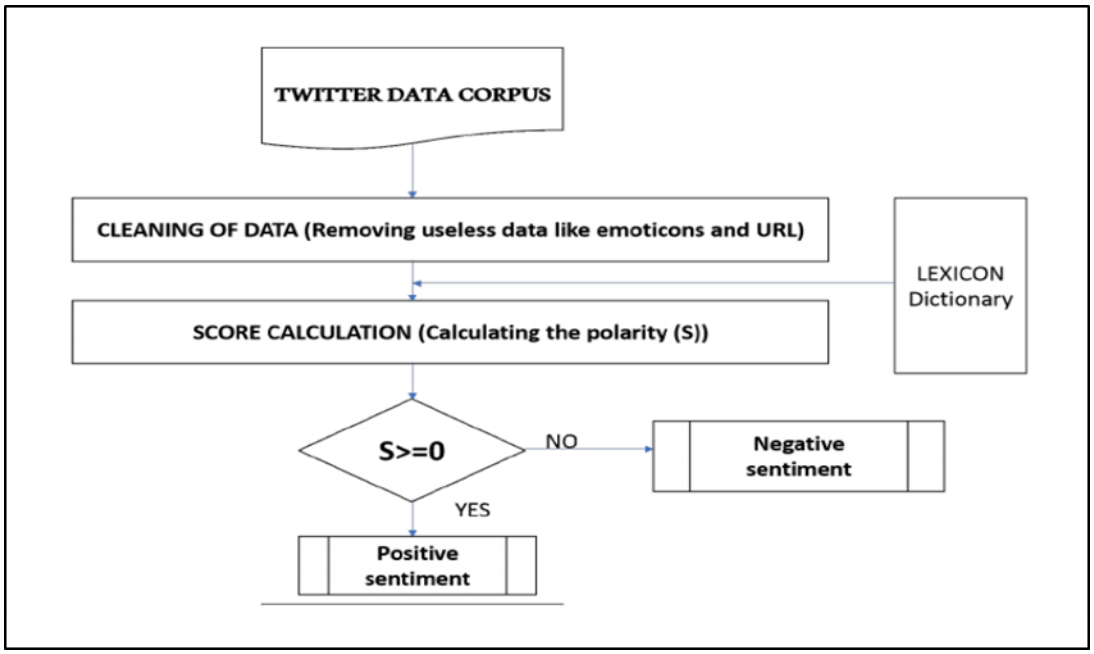

Figure 13: l'organigramme de lexicon approche

2.3.2 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 basée sur les lexiques [10].

2.3.2.1 Les méthodes d'apprentissage automatique

Peuvent être classé en deux catégories : l'apprentissage non supervisé et l'apprentissage supervisé. L'algorithme d'apprentissage non supervisé ne nécessite pas de documents annotés il découvrira automatiquement une structure dans les documents donnés. Contrairement à l'apprentissage supervisé, dans l'apprentissage supervisé consiste à utiliser des documents déjà annotés pour l'apprentissage (training) afin de former le modèle.

Après avoir préparé et annoté l'ensemble de données, il sera divisé en ensembles de données d'entraînement et de test.

L'ensemble de données d'entraînement sera utilisé pour entraîner les modèles et l'ensemble de données de test sera utilisé pour les tests et évaluation de la Méthode. Dans ce processus de formation, la méthode apprendra des documents annotés et cela lui permettra de faire une prédiction sur les documents qui pourraient venir dans le future.

- **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.

- **Méthodes Non-Probabilistes**
  - **SVM (Support Vector Machines)**

Parfois, pour classer les données, nous devons changer leur représentation. Il est toujours possible de transformer n'importe quel ensemble de données afin que les classes qu'il contient puissent être séparées linéairement [27]. Le problème consiste à utiliser un bon nombre de dimensions et un noyau adapté.

SVM (qui est l'acronyme de Support Vector Machines, soit machines à vecteurs support en français, parfois traduit par séparateur à vaste marge pour garder l'acronyme) est un algorithme d'apprentissage automatique, et qui est très efficace dans les problèmes de classification .SVM (Cortes, et al., 1995) est l'un des algorithmes les plus populaires. IL appartient à la catégorie des méthodes linéaires (qui utilisent une séparation linéaire des données), et qui dispose de sa méthode à lui pour trouver la frontière entre les catégories.

Les calculs de séparation des points de données dépendent d'une fonction du noyau. Il existe différentes fonctions de noyau: linéaire, polynomiale, gaussienne, fonction de base radiale (RBF) et sigmoïde. En termes simples, ces fonctions déterminent la fluidité et l'efficacité de la séparation des classes, et jouer avec leurs hyper-paramètres.

- **SVM pour classes binaires**

La méthode des SVM géométriques peut être considérée comme la tentative de trouver, parmi toutes les surfaces σ1, σ2, ... d'un espace de dimensions |T| ce qui sépare les exemples d'apprentissage positifs des négatifs. L'ensemble d'apprentissage est donné par un ensemble de vecteurs associés à leur classe d'appartenance [33] : $(\vec{x}_1, c_1), \ldots . , (\vec{x}_n, c_n), \vec{x}_n \in \mathbb{R}^m, c_n \in \{+1, -1\}$ avec

– c_n représente la classe d'appartenance. Dans un problème à deux classes la première classe correspond à une réponse Positive ($c_n = +1$) et la deuxième classe correspond à une réponse Négative ($c_n = -1$)

– $\vec{x}_n$représente le vecteur du texte numéro j de l'ensemble d'apprentissage.

La méthode SVM sépare les vecteurs à classe Positive des vecteurs à classe Négative par un hyperplan défini par l'équation suivante : $\vec{w}' \vec{x} + b = 0$, $\vec{w} \in \mathbb{R}^m$, $b \in \mathbb{R}$ (Figure 14).

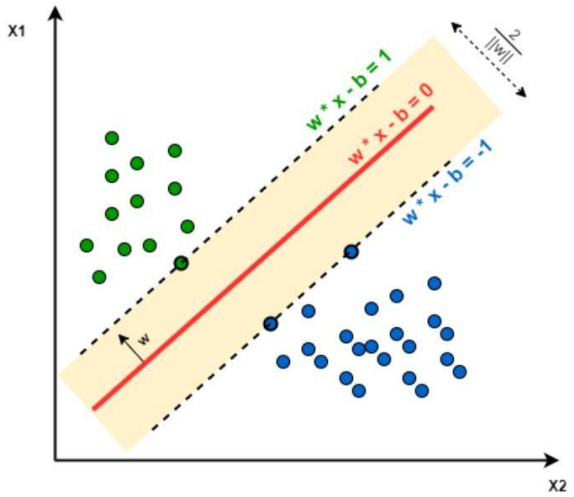

Figure 14: Le modèle du SVM

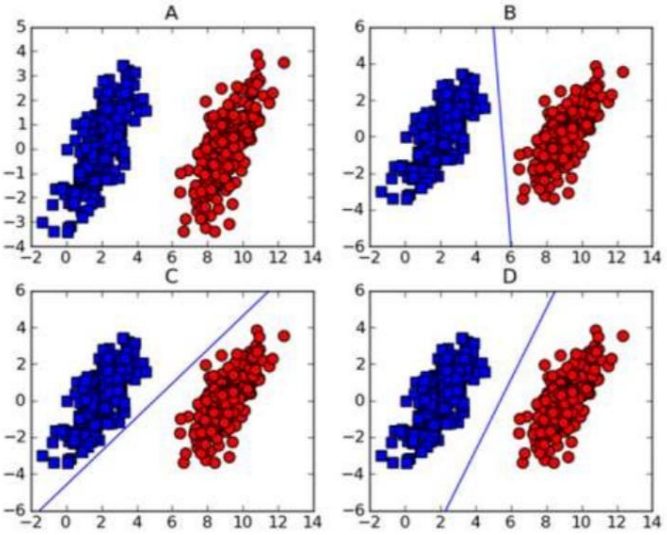

Figure 15: L'ensemble de données et séparation des hyperplans

En général, un tel hyperplan n'est pas unique. La méthode SVM détermine l'hyperplan optimal en maximisant la marge : la marge est la distance entre les vecteurs étiquetés positifs et les vecteurs étiquetés négatifs. L'ensemble d'apprentissage n'est pas nécessairement séparable linéairement, des variables d'écart □□ sont introduites pour tous les □□.

Ces □□ prennent en compte l'erreur de classification, et doivent satisfaire les inégalités suivantes :

– : □ ' □ + □ ≥ 1 − □□

– □ ' □ + □ ≤ 1 + □□

En prenant en compte ces contraintes, nous devons minimiser la fonction d'objectif suivante :

$$\frac{1}{2}\|\square\|^2 + \square \sum_{\square=1}^{\square} \square$$

Le premier terme de cette fonction correspond à la taille de la marge et le second terme représente l'erreur de classification, avec $\square$ représentant le nombre de vecteurs de l'ensemble d'apprentissage. Trouver la fonction objective précédente revient à résoudre le problème quadratique suivant : trouver la fonction de décision $\square$ telle que : $\square(\square) = \square\square\square\square\square(\square(\square))$ dans laquelle la fonction g(X) est :

$$\square(\square) = \sum_{\square=1}^{\square} \square_{\square}\square_{\square}\square_{\square} * \square + \square$$

Avec :

– $\square\square\square\square\square(\square)$ représente la fonction suivante :

– S*i* $\square > 0$ alors $\square\square\square\square\square(\square) = 1$

– Si $\square < 0$ alors $\square\square\square\square\square(\square) = -1$

– Si $\square = 0$ alors $\square\square\square\square\square(\square) = 0$

–$\square_{\square}$représente la classe d'appartenance,

– $\square_{\square}$ représente les paramètres à trouver,

– $\square_{\square} * X$ représente le produit scalaire du vecteur $\square_{\square}$avec le vecteur X.

- **SVM pour multi-classes**

Dans son type le plus simple, SVM ne prend pas en charge la classification multi-classes de manière native. Il prend en charge la classification binaire et la séparation des points de données en deux classes. Pour la classification multi-classe, le même principe est utilisé après avoir décomposé le problème de la multi-classification en plusieurs problèmes de classification binaire [32].

L'idée est de mapper les points de données sur un espace de grande dimension pour obtenir une séparation linéaire mutuelle entre deux classes. C'est ce qu'on appelle une approche **One-to-One**, qui décompose le problème multi-classe en plusieurs problèmes de classification binaire. Un classificateur binaire pour chaque paire de classes.

Une autre approche que l'on peut utiliser est **One-to-Rest**. Dans cette approche, la répartition est définie sur un modèle binaire pour chaque classe.

Le seul SVM effectue une classification binaire et différencier deux classes. De sorte que, selon les deux approches de ventilation, on classe les points de données à partir d'un ensemble de données de m classes:

- Dans l'approche **One-to-Rest**, le modèle peut utiliser m SVMs.
  Chaque SVM prédirait l'appartenance à l'une des classes m.
- Dans l'approche **One-to-One**, le modèle peut utiliser des $\frac{(-1)}{2}$ SVMs.

On Prend un exemple de problème de classification de 3 classes; vert, rouge et bleu, comme l'image suivante [32]:

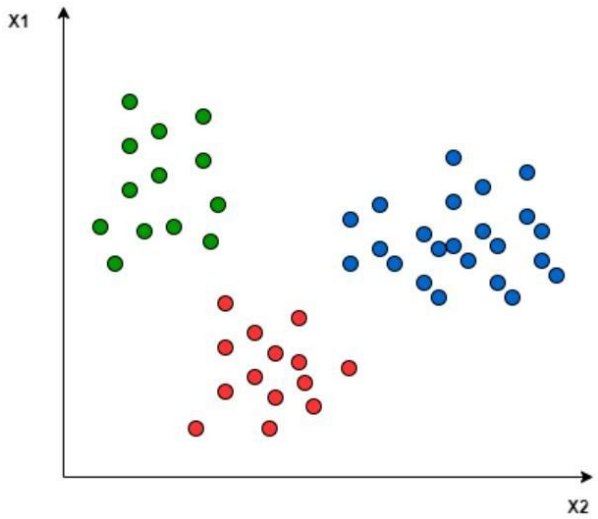

Figure 16: L'ensemble de données sur le SVM

L'application des deux approches à cet ensemble de données donne les résultats suivants:

Dans l'approche **One-to-One**, nous avons besoin d'un hyperplan pour séparer toutes les deux classes, en négligeant les points de la troisième classe. Cela signifie que la séparation ne prend en compte que les points des deux classes dans la répartition actuelle. Par exemple, la ligne rouge-bleue essaie de maximiser la séparation uniquement entre les points bleu et rouge. Cela n'a rien à voir avec les points verts:

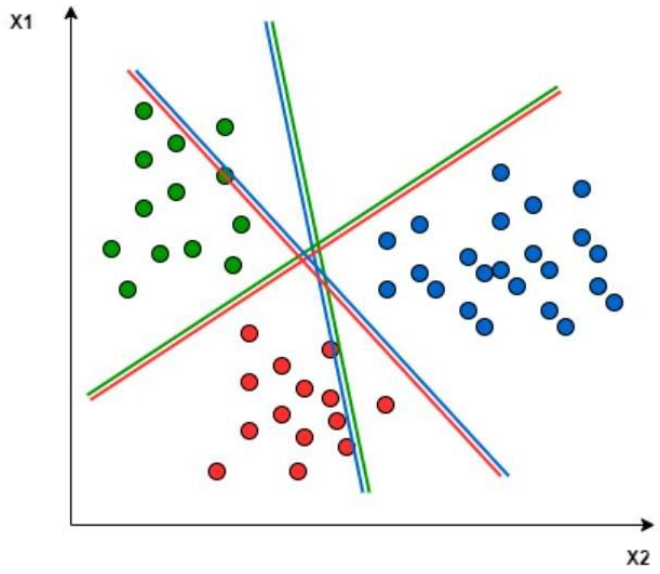

Figure 17: Modèle One-to-One du SVM

Dans l'approche **One-to-Rest**, nous avons besoin d'un hyperplan pour séparer une classe de toutes les autres à la fois. Cela signifie que la séparation prend tous les points en compte, les divisant en deux groupes; un groupe pour les points de classe et un groupe pour tous les autres points. Par exemple, la ligne verte essaie de maximiser la séparation entre les points verts et tous les autres points à la fois:

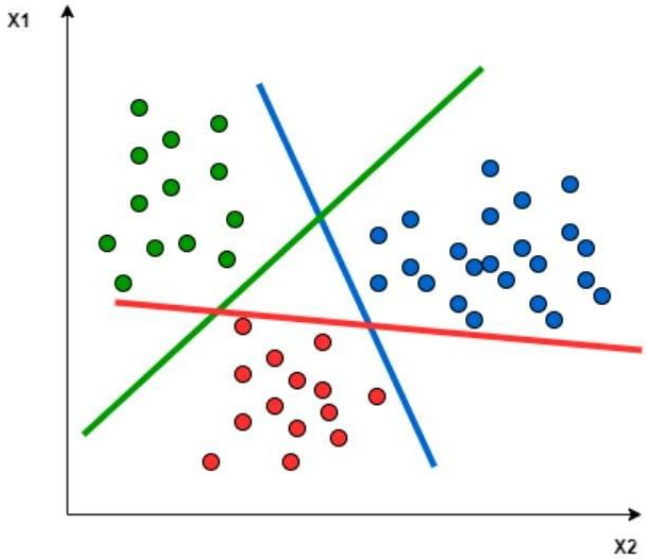

Figure 18: Modèle One-to-Rest du SVM

L'un des problèmes les plus courants dans le monde réel pour la classification multi-classes en utilisant SVM est la classification de texte. Par exemple, classer des articles de presse, des Tweets ou des articles scientifiques.

- **K-NN (K- Neirest Neighbors)**

K-NN a été désigné comme l'un des dix algorithmes non probabilistes les plus populaires et intéressants. Il est connu pour être très simple et facile. K-NN est un exemple basé sur un groupe d'apprentissage [28]. K-NN est effectué en recherchant le groupe de k objets dans les données d'apprentissage les plus proches (similaires) dans de nouvelles données ou des tests de données. Généralement la formule de distance euclidienne est utilisée pour définir la distance entre des objets d'apprentissage et de test.

$$□□□ = \sqrt{\sum_{□=1}^{□} (□_□ - □_□)^2}$$

➢ **Méthodes Probabilistes**

- **Naive Bayes**

C'est une méthode probabiliste qui peut apprendre le modèle en examinant des ensembles un ensemble de documents classés [2]. Elle compare le contenu avec la liste de mots pour classer les documents à leur bonne catégorie ou classe. Soit d le Tweet et c * une classe qui est assignée à d, où

$$□^* = \arg\max_{□} □_{□□(□|□)}$$

$$□_{□□(□|□)} = \frac{(□(□)) \sum_{□=1}^{□} □((□|□))^{□□(□)}}{□(□)}$$

À partir de l'équation ci-dessus, «f» est une «caractéristique», nombre de caractéristiques (fi) est noté $□_{□(□)}$ et il est présenté dans d qui représente un Tweet. Les paramètres P (c) et P (f | c) sont calculés par maximum des estimations de vraisemblance et le lissage est utilisé pour les fonctionnalités.

- **Maximum Entropy**

Dans la méthode Maximum Entropy, aucune hypothèse n'est prise concernant la relation entre les caractéristiques extraites à partir de l'ensemble de données [2]. Cette méthode essaie toujours de maximiser l'entropie du système en estimant la distribution conditionnelle de l'étiquette de classe.

L'entropie maximale gère même la fonction de chevauchement et est identique à la méthode de régression logistique qui trouve la distribution sur des classes. La distribution conditionnelle est définie comme MaxEnt ne fait aucune hypothèse d'indépendance pour ses fonctionnalités, contrairement à Naïve Bayes.

Le modèle est représenté par la formule qui suit:

$$\square_{\square\square}(\square|\square,\lambda) = \frac{\square\square\square\left[\sum_{\square}\lambda_{\square}\square_{(\square)}\right]}{\sum_{\square}\square\square\square\left[\sum_{\square}\lambda_{\square}\square_{(\square)}\right]}$$

Où « c » est la classe, « d »est le Tweet et « λi » est le poids. Les vecteurs de poids décident de l'importance d'une caractéristique classification.

- **L'apprentissage Non-supervisé**

La classification ou la régression, est un thème de recherche majeur en apprentissage automatique, en analyse et en fouille de données où l'objectif est, la répartition en classes d'un ensemble de données non étiquetées.

Dans des nombreuses études Il a été rapporté que SVM dépasse les autres méthodes dans la classification pour l'ASA.

La figure suivante montre les Méthodes les plus couramment utilisées pour l'ASA et le nombre d'articles qui les ont appliqués. Comme nous pouvons le voir, il existe quatre Méthodes principaux qui sont SVM, NB, K-NN et DT [10].

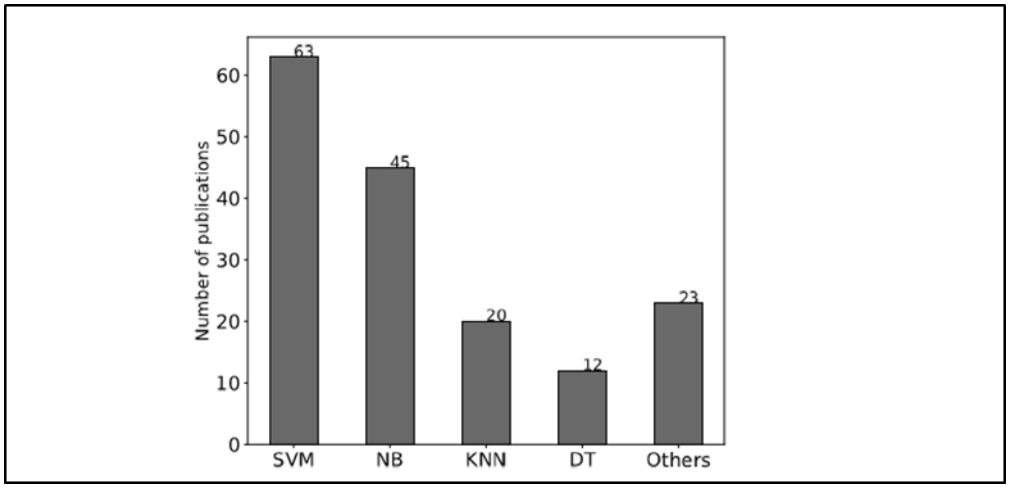

Figure 19: Le nombre de publications pour Les Méthodes de Machine Learning

SVM est le plus appliqué là où 63 papiers ont utilisé SVM qui représente environ 39% du nombre total de papiers qui appliquait les Méthodes du ML. NB vient ensuite avec 45 papiers.

K-NN et DT sont moins courants avec 20 et 12 articles, respectivement.

D'autres Méthodes ont également été appliqués pour la ASA dans quelques études comme Conditional Random Field, Hybrid SVM & K-NN, Functional Tree , Genetic-Reduce et Random Sub Space (RSS) & SVM ,Toutes les autres Méthodes forment environ 14% de tous les papiers.

2.3.2.2 Avantages et les inconvénients des Méthodes vu

Pour toutes les Méthodes, celles mentionnées ci-dessous ont leurs avantages et leurs inconvénients [13]. Nous les résumons dans le tableau suivant :

Méthode	Catégorie	Avantages	Inconvénients
SVM	Supervisé	- 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	-Problème lorsque les classes sont bruitées -Le traitement des problèmes multi-classes reste une question ouverte -Problème lorsque les classes sont bruitées
Naïve Bayes	Supervisé	-Rapide dans la tâche d'entrainement et de classification. - La facilité et la simplicité de leur implémentation.	-Moins précis que SVM -ces performances sont limitées quand il s'agit d'une grande quantité de lexiques à traiter
K-NN	Supervisé	-Entraînement très rapide. -Simple et facile à comprendre. -La méthode des k plus proches voisins n'utilise pas de modèle pour	-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,

		classifier les documents.	avant de décider quelle classe à choisir. -Haute complexité de calcul. -la grande capacité de stockage qu'elle nécessite pour le traitement des corpus.

Tableau 1: Les avantages et les inconvénients des Méthodes ML

2.3.3 L'approche Hybride

L'approche Hybride combine à la fois l'apprentissage automatique et l'approche basée sur le lexique. Cette approche a été signalée comme une meilleure approche que l'apprentissage automatique et les approches basées sur le lexique [10].

La combinaison peut être faite en ajoutant des fonctionnalités extraites du lexique dans une Méthode d'apprentissage automatique ainsi que d'autres fonctionnalités telles que les fonctionnalités au niveau de la phrase.

Une autre façon de combiner les deux approches consiste à supprimer chaque mot qui n'existe pas dans le lexique de toutes les instances afin que les mots de sentiment restent seuls. Les mots supprimés n'affectent pas le sentiment, mais les garder déroutera la méthode et diminuera la précision. Par conséquent, en supprimant ces mots, la précision devrait s'améliorer.

Par rapport aux approches précédentes, un peu de travaux ont utilisé l'approche hybride pour l'ASA.

Parmi ces derniers, nous présentons le travail d'une approche hybride d'Analyse des Sentiments pour classer des Tweets. Dans ce travail, les chercheurs ont combiné les deux approches d'Analyse des Sentiments l'apprentissage automatique et les approches basées sur le lexique.

Dans l'approche d'apprentissage automatique, trois fonctionnalités ont été utilisées : uni-gramme, bi-gramme et tri-gramme.

Et pour l'une approche basée sur le lexique, ils ont collecté les mots de sentiment de leur corpus et ont donné à chaque mot un poids en fonction de sa fréquence.

La combinaison entre les deux approches a été réalisée en ajoutant les mots de sentiment en tant que fonctionnalités dans l'approche d'apprentissage automatique. Et dans les résultats, la précision de l'approche hybride était meilleure.

Un autre exemple de travail intéressant, Ce travail a présenté une approche d'Analyse des Sentiments hybride en supprimant tous les mots non-sentimentaux de l'ensemble de données d'entraînement afin que l'ensemble de données d'entraînement contienne des mots qui décrivent le sentiment dans les instances et leur étiquette.

Les auteurs de ce travail ont utilisé deux Méthodes, SVM et K-NN, et ils ont comparé l'approche proposée avec l'approche d'apprentissage automatique en utilisant les mêmes Méthodes. Les résultats ont confirmé que l'approche proposée à une meilleure précision que l'approche d'apprentissage automatique.

2.3.4 Les avantages et les inconvénients d'approches

Le tableau suivant montre les avantages et les inconvénients de l'approche basée sur ML et de l'approche basée sur le lexique [11] que nous avons vue précédemment.

Approches	Avantages	Inconvénients
Approche basé sur le lexique	-Il ne demande aucune donnée d'entrainement ou des données étiquetées et ceci permet d'introduire moins d'opérations de calcul	-Moins de capacité de classification en fonction du contexte ou du domaine. -Exige l'existence de ressources linguistiques puissantes qui ne sont pas toujours disponibles
Approche d'apprentissage automatique	-Il peut être transformé en ce que le domaine demande pour mieux travailler. -Un dictionnaire n'est pas nécessaire. -Donne de meilleurs résultats en termes de haute précision de classification.	- peut être affecté par les variations de classes et aussi par l'effet des changements linguistiques. -Les Méthodes qui se sont entrainées sur un domaine spécifique, dans la plupart des cas ne fonctionnent pas avec un autre

Tableau 2: Les avantages et les inconvénients d'approches lexique et Machine Learning

2.4 L'Évaluation du système

Dans le domaine de l'analyse des sentiments, la plupart des travaux de la littérature ont utilisé les métriques d'extraction d'informations communes Précision, Rappel et F-mesure afin d'évaluer leurs méthodes [12].

2.4.1 Précision

La précision est le nombre de documents pertinents retrouvés rapporté au nombre de documents total proposé pour une requête donnée.

Le principe est le suivant : quand un utilisateur interroge une base de données, il souhaite que les documents proposés en réponse à son interrogation correspondent à son attente. Tous les documents retournés superflus ou non pertinents constituent du bruit. La précision s'oppose à ce

bruit documentaire. Si elle est élevée, cela signifie que peu de documents inutiles sont proposés par le système et que ce dernier peut être considéré comme « précis ». On calcule la précision avec la formule suivante :

$$\text{précision} = \frac{\text{nombre de documents pertinents retrouvés}}{\text{nombre de documents retrouvés}}$$

En statistique, la précision est appelée valeur prédictive positive.

2.4.2 Rappel

Le rappel est défini par le nombre de documents pertinents retrouvés au regard du nombre de documents pertinents que possède la base de données.

Cela signifie que lorsque l'utilisateur interroge la base, il souhaite voir apparaître tous les documents qui pourraient répondre à son besoin d'information. Si cette adéquation entre le questionnement de l'utilisateur et le nombre de documents présentés est importante alors le taux de rappel est élevé. À l'inverse, si le système possède de nombreux documents intéressants mais que ceux-ci n'apparaissent pas dans la liste des réponses, on parle de silence. Le silence s'oppose au rappel. Le rappel est donc calculé comme suit :

$$\text{rappel} = \frac{\text{nombre de documents pertinents retrouvés}}{\text{nombre de documents pertinents}}$$

En statistique, le rappel est appelé sensibilité.

2.4.3 F-mesure

Une mesure qui combine la précision et le rappel est leur moyenne harmonique, nommée F-mesure ou F-score :

$$F - \text{mesure} = 2 * \frac{(\text{précision} * \text{rappel})}{(\text{précision} + \text{rappel})}$$

Cela est dû à l'évaluation du processus de la classification est considérée comme une tâche de recherche d'informations. Cette évaluation est parfaite si elle fournira des réponses dont la précision et le rappel sont égaux à 1 (l'algorithme trouve la totalité des documents pertinents - rappel - et ne fait aucune erreur - précision). Dans la réalité, les algorithmes de recherche sont plus ou moins précis et plus ou moins pertinents. Il est possible d'obtenir un système très précis (par

exemple un score de précision de 0,99), mais peu sensible (par exemple avec un rappel de 0,10, qui signifiera qu'il n'a trouvé que 10 % des réponses possibles). De même, un algorithme dont le rappel est fort (par exemple 0,99), mais la précision faible (par exemple 0,10) fournira en guise de réponse de nombreux documents erronés en plus de ceux pertinents : il sera donc difficilement exploitable.

3 Conclusion

Après ce parcours bibliographique, nous allons présenter dans le chapitre suivant commencer la première phase de développement c'est la conception et modélisation du système pour assurer un bon processus de travail et avoir un bon système. Nous allons montrer nos expérentations avec l'Analyse des Sentiments appliqué à notre corpus.

Chapitre III :
Conception et modélisation de la solution

1 Introduction

De nos jours, l'Analyse des Sentiments est devenue un sujet de recherche émergent en raison du grand nombre de données disponibles sur les blogs et les réseaux sociaux. Suivre différents types d'opinions et les résumer peut fournir des informations précieuses aux différents types d'opinions des utilisateurs qui utilisent les réseaux sociaux pour obtenir des commentaires sur tout produit, service ou autre. L'analyse des opinions et sa classification sur la base de la polarité (Positive, Négative, Neutre) est une tâche difficile. Beaucoup de travaux ont été fait sur l'Analyse des Sentiments des données de Twitter et mais il reste beaucoup à faire.

2 Notre système d'Analyse des Sentiments

Le cadre d'Analyse des Sentiments en langue arabe proposé sur un réseau social a pu analyser les commentaires et les différentes opinions, points de vue et émotions à partir du texte, discours, Tweets et sources de bases de données via le Traitement du Language Natural (TLN) et les classer en trois catégories positives, négatives et neutres.

Pour notre système d'analyse les sentiments nous avons choisi le réseau social Twitter. Twitter actuellement est une plate-forme de micro-blogging la plus populaire .Elle couvre des différents sujets du jour, tout en limitant le nombre de caractères utilisés dans un Tweet à 140 caractères pour la langue Arabe.

Dans cette étude, nous visons à établir une étude par les techniques d'apprentissage automatique « Machine Learning » supervisés : SVM« Support Vector Machine » K-NN « k – Nearest Neighbors », et utiliser des fonctionnalités simples pour obtenir de bons résultats.

2.1 Architecture du système

Notre système est un modèle qui s'appelle l'Analyse des Sentiments des Tweets Arabes que nous avons développé pour analyser les données de Twitter « Tweets » et donner pour chaque Tweet sa prédiction (Positive, Négative, Neutre). Pour notre solution, nous sommes basés sur l'approche de Machine Learning « l'approche d'apprentissage automatique »et ces techniques. Ce modèle vise à étudier les performances de l'approche d'apprentissage automatique pour classer les critiques de sentiments collectées sur Twitter.

Le cadre de la méthode proposée du modèle ASA se compose de six phases principales : le Dataset, Prétraitement, Extraction des caractéristiques, Préparer les ensembles d'entrainement et test, la Classification par les techniques de ML (SVM et K-NN), et finir par l'évaluation du système comme indiqué sur le schéma au-dessous (Figure 20).

Tout d'abord, nous avons commencé par la création du Dataset , Cette phase vise à clarifier les données en déterminant les détails de son contenu par une base de données qui présente plus de 12483 Tweets annotés , chaque Tweet est étiqueté en trois classes (Négative , Positive, Neutre) .

Après pour la deuxième phase nous avons procédé par un prétraitement des Tweets du Dataset par la Suppression des caractères inutiles, puis nous avons effectué une Normalisation et Tokenisation. Tout cela pour éliminer le bruit pour chaque Tweet. Puis nous avons effectué une Extraction des fonctionnalités, Cette étape est critique car le type d'entités extraites et la manière dont elles sont construites influence la performance des Méthodes ML. Les différents types de caractéristiques étaient extraits grâce aux méthodes BoW, TF-IDF, n-Grams. Après nous avons passé à une étape essentielle du système la Classification du Tweets qui se base sur les Méthodes d'apprentissage automatique supervisés (SVM et K-NN).

Finalement nous avons évalué notre système et fait une comparaison entre ces deux Méthodes (SVM et K-NN) pour connaitre la meilleur Méthode pour notre système.

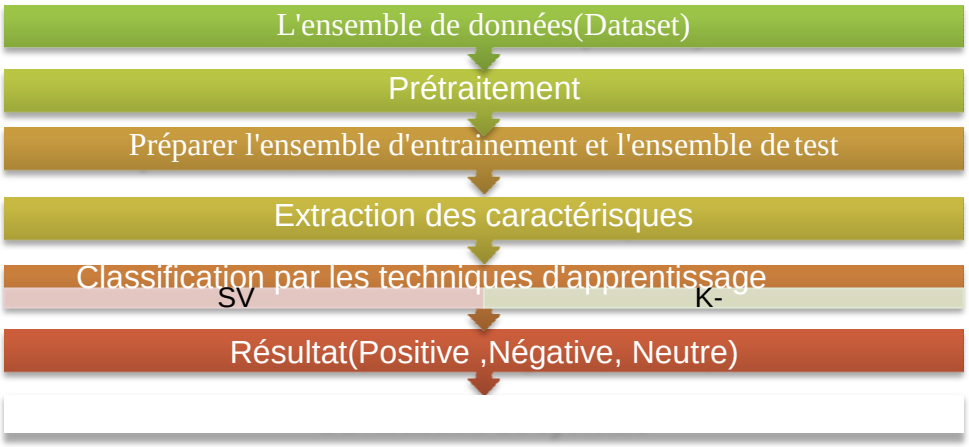

Figure 20: Processus d'Analyse des Sentiments du notre système

Ci-dessous, nous allons décrire les différentes étapes de notre processus (figure 20) d'Analyse de Sentiments pour la langue Arabe dans le détail.

2.1.1 L'ensemble de données (Dataset)

Cette phase de Dataset vise à clarifier l'ensemble de données en déterminant les détails de leur contenu.

Pour notre système, le Dataset qui a été utilisé présente une collection de Tweets en langue arabe « #darija » et « #fosha » qui ont été collectés à des fins d'Analyse des Sentiments. Cet ensemble de données contient des Tweets et chaque Tweet a été étiqueté par un type de classe (Positive, Négative, Neutre).

Le contenue du Dataset contient 12483Tweets étiquetés (6000Tweets positifs et 5000Tweets négatifs, 1483Tweets Neutres) liés à différents sujets par exemple «Politique», «arts» etc.

2.1.2 Prétraitement des Tweets

Le Prétraitement du texte est une étape essentielle pour faciliter le traitement et l'analyse de données, Nous nettoyons les Tweets de la collection dans le but de les normaliser en éliminant les données non pertinentes de Twitter et les données inutiles.

Le Prétraitement comprend les étapes suivantes : Nettoyage des données, Normalisation, Tokenisation et Suppression des mots vides, ainsi que le Stemming

2.1.2.1 Nettoyage des données

Le nettoyage des données ou data cleanning est une tâche critique pour gérer le bruit des données Twitter.

Cette étape consiste à supprimer des éléments des Tweets qui n'incluent aucun sentiment. En tant que tel, certains éléments que nous avons supprimé incluent : les URLs, les chiffres, les caractères non alphabétiques (par exemple, <÷ >-+ =% $) et les signes de ponctuation du la langue arabe (par exemple : ","_" ;" ":"-"."/" ! " ,"؟") comme le montre dans le tableau 3 au-dessous (ces caractères sont remplacé par " ") .

2.1.2.2 Normalisation

La tâche de Normalisation est importante pour produire des formes de mots cohérentes.

Pour le texte Arabe, nous allons faire la normalisation selon les étapes suivantes :

- Élimination des signes diacritiques

  tashkeel (# Tashdid# Fatha # Tanwin Fath # Dama # Tanwin #Damma # Kasra# Tanwin# Sukun) par exemple" ٱلْعَرَبِيَّةُ " à "العربية"

- Élimination Tatweel :"العربيــة" à"العربية"
- Remplacement de la lettre "ة" par "ه"
- Remplacement de la lettre "ى" par "ي"
- Remplacement des lettres "آ-إ-أ" par "ا"
- Normalisation des lettres répétées: par ex. "سعاااادة " à" سعادة"

Tableau ci-dessous présente tous les caractères que nous avons remplacés ou éliminés (remplacer par « «ou éliminé)

Valeur pour remplacer	Valeur Remplacer par
إ	ا
أ	ا
آ	ا
ة	ه
"_"	" "
"-"	" "
"/"	" "
"+"	" "
"="	" "
"×"	" "
"."	" "
"،"	" "
و	و
يا	يا
' " '	" "
"-"	" "
" ' "	" "
" ى "	ي
"\\"	" "
'\n'	" "
'\t'	" "
'؟'	'؟'
'؟'	'؟'
'!'	'!'
"،"	" "
"."	" "
'وو'	'و'
'يي'	'ي'
'اا'	' ا '

Tableau 3: La Normalisation de Tweets

2.1.2.3 Stemming

Stemming est une technique qui aide à réduire la dimensionnalité élevée de l'espace des fonctionnalités dans la classification de textes. Plusieurs approches de Stemming existent pour la langue arabe, chacune produisant un ensemble différent de racines.

En arabe, les approches de Stemming les plus connues utilisées sont root-based Stemmers, Arabic Light Stemming .

Pour cette étape nous avons utilisé Arabic Light Stemming .Cette approche n'est pas utilisée pour produire la racine linguistique d'une forme de surface arabe donnée, mais pour supprimer les suffixes et préfixes les plus fréquents. Les suffixes le plus courant comprend les duals et les pluriels pour les formes masculines et féminines, possessives, les articles définis et les pronoms.

L'exemple suivant illustre ce principe : Prenons cette liste des mots ou Tokens ['دعاء', 'إلى','قول','أن','يود','ام','لك','تحول','الذين','نحن'] après l'utilisation du Light Stemming elle devient ['دعاء' ,'إلى' ,'قول' ,'أن' ,'ود' ,'ام' ,'لك' ,'حول' ,'الذين' ,'نحن']

2.1.2.4 Tokenisation

Au cours de cette étape, le texte du Tweet a été divisé en une séquence de jetons où chaque jeton représente un seul mot basé sur des espaces blancs. Un exemple d'après notre projet:

Nous prenons cette Tweet"السلام عليكم ورحمة الله" après la Tokenization du cette Tweet elle devient ['السلام', 'عليكم', 'ورحمة', 'الله']

2.1.3 Préparer les ensembles de données d'entrainement et de test

Pour cette étape nous commençons par diviser les données en deux ensembles de données, Train et Test. L'ensemble de données d'entraînement pour s'adapter au système, il facilite les prédictions des données inconnues. Après des prédictions seront effectuées sur l'ensemble de données de test. Dans notre travail nous avons utilisé les techniques supervisées SVM et K-NN ces techniques ont été appliqués dans des travaux de recherche antérieurs sur ASA, elles sont basées sur les deux ensembles de données un ensemble de données étiquetées et un ensemble de données de test « Test », car ces techniques du L'apprentissage automatique sont supervisés. Alors elles se basent sur des données annotés ou étiquetés pour donner une bonne prédiction de Tweets du l'ensemble « Test ».

Alors le principe de l'ensemble d'entrainement, il fournit des étiquettes au modèle supervisé pendant le processus comme il est présenté dans le schéma ci-dessous (Figure 21). En plus ces ensembles de données étiquetés sont formés pour obtenir des résultats significatifs lorsqu'ils sont rencontrés lors de la prise de décision [24].

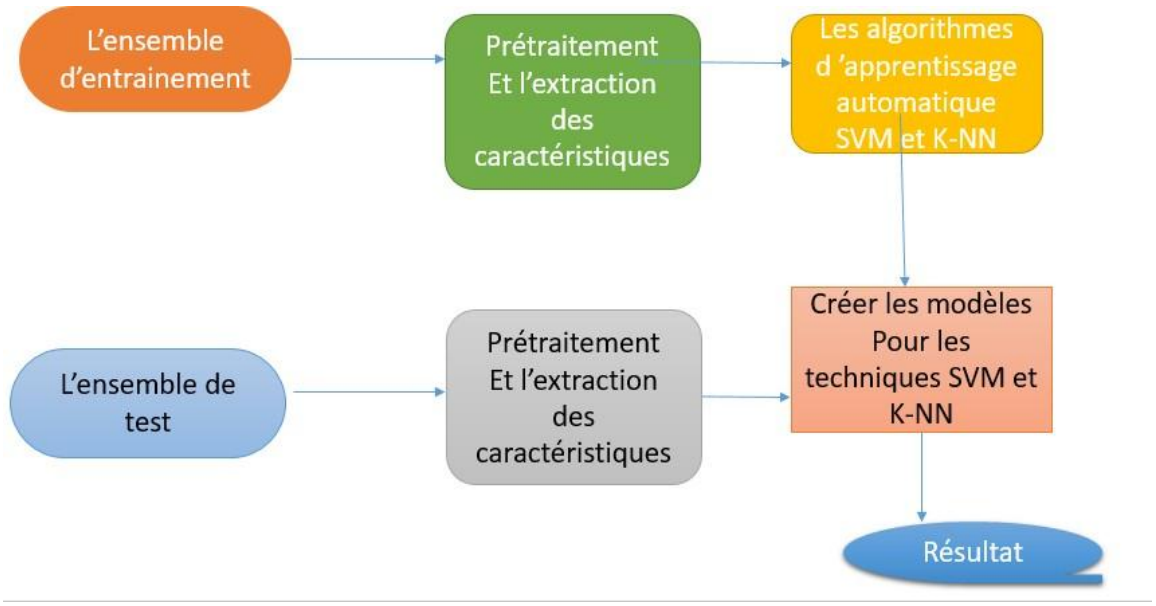

Figure 21 : L'ensemble d'entrainement et l'ensemble de test dans la classification

Ce tableau ci-dessous montre comment nous avons divisé l'ensemble de données en ensemble d'entraînement « Train »et l'ensemble de Test et le nombre de Tweets pour chaque ensemble. Les données de formation « Train » couvrent 70% du corpus et les données de Test représentent les 30% restants.

Dataset	Positive	Négative	Neutre
Training	4200	3500	1083
Test	1800	1500	400
Total	6000	5000	1483

Tableau 4: Les statistiques du Dataset utilisé

2.1.4 Extraction des caractéristiques et Vectorisation des mots

La classification des sentiments peut être définie comme la tâche générale consistant à classer une séquence de texte d'entrée selon des critères. L'extraction des caractéristiques est une étape essentielle à l'efficacité de l'analyse de séquence, car les séquences de texte ne peuvent pas être facilement décrites comme des vecteurs de caractéristiques.

Dans cette étape nous avons un processus général de transformation d'une collection de Tweets en vecteurs de caractéristiques numériques, cette représentation est nécessaire pour les algorithmes de ML supervisés et la majorité des approches ML utilisent le VSM , où chaque document est représenté sous la forme d'un vecteur de caractéristiques pondérées. Il existe de nombreuses méthodes pour convertir des données texte en vecteurs que le modèle peut comprendre, mais la méthode la plus populaire, de loin et que nous avons utilisés c'est la méthode TF-IDF qui signifie « Fréquence du terme - inversé Document Fréquences » en plus Sac de Mots (Bag Of Words), N-gram modèles.

2.4.1 Sac de Mots (Bag Of Words « BOW »)

Le modèle BOW utilisé dans la plupart des applications de classification de texte, il est efficace dans le domaine de l'Analyse des sentiments. Pour construire le vecteur de caractéristiques, il considère les mots de base comme informatifs des aspects des textes. Il se compose de mots distincts qui apparaissent dans l'ensemble de données après le prétraitement des Tweets.

Ainsi, pour le modèle Bag Of Words nous devons présenter les Tweets par des vecteurs numériques car les techniques de la classification du texte ne fonctionnent qu'avec des nombres. Pour créer le modèle du sac de mots, nous devons créer une matrice où les colonnes correspondent aux mots ou tokens les plus fréquents de notre corpus (les Tweets après le prétraitement) où les lignes correspondent aux Tweets. Dans cette matrice toutes les cellules de la matrice sont remplies avec 0 ou 1, selon l'occurrence du mot.

Nous prenons un exemple ci-dessous .la première étape nous devons calculer pour chaque token du Tweet une fréquence correspondante. Après nous remplissons la matrice du BOW par les mots les plus fréquents dans les colonnes et les Tweets dans les lignes, la figure (Figure 22).

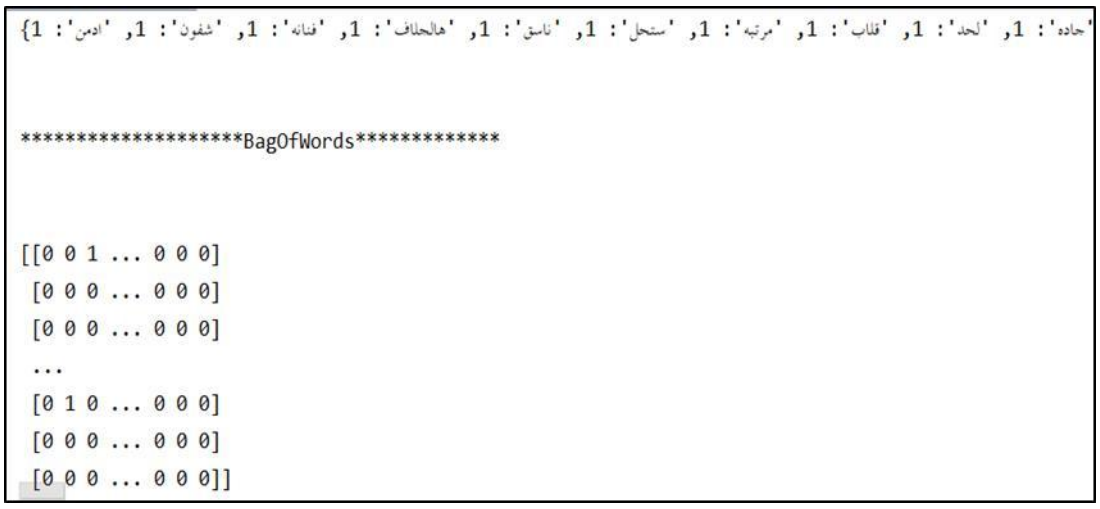

Figure 22: Un modèle du Sac de Mots

2.4.2 Générer des modelés N-grams

Nous avons montré comment une pondération du système peut influencer sur la classification du système en utilisant des modèles N-grams (l'uni-gram, le bi-gram et tri-gram) et vérifier leur impact sur la classification.

Cette fonction présente une méthode de vérification de «n» mots à partir d'une séquence donnée Tweet.

Uni-gram fait référence à n-grams de taille 1, Bi-gram se réfère à n-grams de taille 2, Tri-gram se réfère à n-gramme de taille 3. Un n-grams plus élevé fait référence à quatre grammes, cinq grammes, etc.

La méthode n-grams peut être expliquée à l'aide de l'exemple suivant: Un exemple typique de phrase peut être considéré comme " الاعنو و الاعانيۃ و األ من".

• Les uni-gram : «الاعنو »,«و »,«الاعانيۃ»,«و »,«األ من »où un seul mot est considéré .

. • Les bi-grams: «الاعنو و»,«و الاعانيۃ »,«الاعانيۃ و »,«و األ من »où une paire de mots est considérée.

• Les tri-grams: «الاعنو و الاعانيۃ», «الاعانيۃ و األ من», «و الاعانيۃ و»»où un ensemble de mots comptant trois est considéré.

Pour cela nous avons opté l'uni-gram pour les Tweets de de notre Dataset que nous présentons pour chaque colonne dans par le Bag Of Words.

2.4.3 La fonction TF-IDF

Avant le processus de classification, nous devons calculer la pondération des mots en utilisant le terme schéma de pondération fréquence et fréquence inverse des documents (TF-IDF). TF-IDF est le taux de fréquence d'un terme dans les Tweets, il permet de réduire le poids des fonctionnalités qui apparaissent dans les Tweets.

- **TF :**

TF est l'abréviation de l'anglais Term Frenquency (fréquence du terme). Il détermine la fréquence relative d'un mot ou d'une combinaison de mots dans un document. Cette fréquence du terme sera comparée à la survenance de tous les autres mots restants du texte, du document ou du site web analysé. Cette formule se présente comme suit :

[illegible]

Cette formule atteste qu'une augmentation visible du mot-clé dans le texte ne mène pas à une amélioration de sa valeur dans le calcul. Alors que la densité du mot-clé calcule principalement la distribution en pourcentage d'un seul mot dans le texte (en relation avec le nombre total de mots restant), le Term Frenquency factorise également en proportion de tous les mots utilisés dans le texte.

- **IDF :**

L'IDF calcule l'Inverse Document Frequency (la fréquence inverse du document) et complète l'analyse de l'évaluation du mot. Il agit en tant que correctif du TF. L'IDF inclut dans le calcul la fréquence des documents pour un mot précis, autrement dit l'IDF compare le chiffre correspondant à tous les documents connus avec le nombre de textes contenant le mot en question. L'IDF est calculé avec la formule suivante :

[illegible]

En conséquence, l'IDF détermine la pertinence d'un texte en considérant un mot clé précis. Par exemple, lorsqu'un document de 15 mots contient le terme «جميل» 1 fois, le TF pour le mot «جميل» est TF = 1/15 c'est-à-dire 0,066

L'IDF (fréquence de document inverse) d'un mot est la mesure de la signification de ce terme dans l'ensemble du corpus. Par exemple, disons que le terme «جميل»apparaît x fois dans un corpus de

12 000 de Tweets. Supposons qu'il y ait 500 de documents qui contiennent le terme «لجمڹ», alors l'IDF (i.e. log {DF}) est donnée par le nombre total de documents (12 000) divisé par le nombre de documents contenant le terme «لجمڹ» (500).

□□□ = 1,38 Alors □ «لجمڹ» = (□□ * □□□) = 0,08

TF-IDF sont des scores de fréquence de mots qui tentent de mettre en évidence les mots les plus intéressants. Fréquents dans un document mais pas entre les documents. Par la suite nous pouvons présenter par la matrice du Bag Of Word pour chaque token du Tweets son TF-IDF, comme il est présenté dans la figure (Figure 23) ci-dessous :

```
********************BagOfWordTF-IDF*************


[[0.         0.         0.         ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.2807117  0.         0.        ]
 [0.28545621 0.         0.         ... 0.         0.         0.        ]
 ...
 [0.         0.         0.         ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]]
```

Figure 23 : Bag Of Words TF-IDF

Le modèle présenté dans la figure en-dessus est différent du modèle simple de Bag Of Words car il ne représente pas le Tweet en tant que vecteurs de «0» et «1», mais attribue plutôt des valeurs TF-IDF plus précises entre 0 et 1. Le modèle TF-IDF simple fonctionne bien et donne l'importance des mots rares plutôt que de traiter tous les mots considérés comme égaux dans le cas du modèle simple de Bag Of Words.

Maintenant, les trois sous-tâches que nous avons vu précédemment Calcul TF-IDF et BOW, N-gram sont appliqués à chacun des mots du Tweet du notre Dataset. Ces valeurs sont utilisées comme poids pour chaque mot où le mot avec le poids le plus élevé est considéré comme une bonne fonctionnalité pour un document Tweet.

2.1.5 La classification par SVM et K-NN

Dans cette étape, la représentation résultante est fournie à l'algorithme de classification ML « Machine Learning »pour créer et apprendre un modèle à partir des Tweets « train » étiquetés qui peuvent prédire les étiquettes de sentiment des nouveaux Tweets sans étiquettes « test ».

Des techniques d'apprentissage automatique comme (K-NN), et Support Vector Machine (SVM) ont réalisé de grands succès dans l'Analyse des Sentiments [24].

Sur la base de notre enquête, nous avons constaté que les algorithmes de classification supervisée Non-probabilistes comme Support Vector Machine (SVM), K-NN sont les plus utilisés dansla ASA.

La raison derrière l'utilisation d'une telle technique repose sur leur succès et capacité efficace à traiter le texte catégorisation où le nombre de fonctionnalités utilisé est énorme pour détecter le sentiment.

2.1.5.1 La méthode SVM « Support Vector Machine »

SVM est un algorithme qui transforme les données d'apprentissage en dimension supérieure en utilisant une fonction de mappage non linéaire, et dans la nouvelle dimension, il détermine le meilleur séparateur linéaire entre différentes classes.

Nous avons appliqué la méthode SVM avec les fonctionnalités proposées (TF-IDF) comme définie précédemment.

2.1.5.2 La méthode K-NN « k-Nearest Neighbors »

La deuxième méthode ML appliqué en ASA est (K-NN) qui est aussi non probabiliste.

K-Nearest Neighbours (K-NN) est un algorithme de classification supervisée non paramétrique, qui est simple mais efficace dans de nombreux cas

Nous avons appliqué la méthode K-NN avec les fonctionnalités proposées (TF-IDF) comme définie précédemment.

2.1.6 L'Evaluation du Système

Dans ce travail, différentes approches de représentation de Tweet ont été expérimentées pour déterminer la meilleure approche qui amélioré les performances du modèle ASA.

Les résultats obtenus feront l'objet du prochain chapitre.

3 Conclusion

Après qu'nous avons présenté l'architecture du notre système et comment on doit construire ce système, nous pouvons commencer la partie essentiel le développement du l'application. D'après la conception du système on conclut qu'Il existe différentes techniques d'apprentissage automatique pour identifier les sentiments à partir des Tweets en langue arabes comme les techniques que nous avons utilisé SVM et K-NN. Nous trouvons que ces techniques du Machine Learning sont plus simples et efficaces pour le domaine d'analyse les sentiments. Dans le prochain chapitre nous présentons les différents outils utilisés et exposer l'application.

Chapitre IV :
Implémentation de la solution

1 Introduction

Après la phase de la conception et modélisation nous allons dans ce chapitre, présenter l'étape d'implémentation de notre système en passant en revue le matériel et logiciel utilisé ainsi que le framework de développement et les différentes librairies pour l'implémentation de notre solution d'Analyse des Sentiments pour la langue arabe en terminant par les différentes pages de l'application et certains codes source.

Mais avant cela, nous allons présenter les résultats des tests que nous avons effectués

2 Résultats d'évaluation

La performance de l'approche développée a été évaluée à l'aide de la F-mesure. C'est la moyenne harmonique entre la précision et le rappel. La précision et le rappel sont deux paramètres d'évaluation standard largement utilisés pour évaluer l'efficacité des algorithmes de classification.

La Précision (P), est le nombre de Tweets Positifs divisé par le nombre de Tweets étiquetés comme Positifs par le système. Le même calcule pour les Tweets Négatifs et Neutre. Il se défini comme ça :

□□é□□□□□□[□□□□□□□] = □□□□ □□□□□□□ / (□□□□ □□□□□□□ + □□□□ □□□□□□□)

□□é□□□□□□[□é□□□□□] = □□□□ □é□□□□□ / (□□□□ □é□□□□□ + □□□□ □é□□□□□)

□□é□□□□□□[□□□□□□□] = □□□□ □□□□□□□ / (□□□□ □□□□□□□ + □□□□ □□□□□□□)

Où Vrai Positif (ou Négatif, Neutre) est le nombre de phrases correctement classées et Faux Positif (ou Négatif, Neutre) est le nombre de phrases incorrectement classées.

Le Rappel (R), est le nombre de Tweets positifs divisé par le nombre de Tweets positifs présenté dans le corpus. Le même calcule pour les Tweets Négatifs et Neutres. Il se définit comme ça :

□□□□□□[□□□□□□□] = □□□□ □□□□□□□ / (□□□□ □□□□□□□ + □□□□ □é□□□□□ + □□□□ □□□□□□□)

□□□□□□[□é□□□□□] = □□□□ □é□□□□□ /(□□□□ □é□□□□□ + □□□□ □□□□□□□ + □□□□ □□□□□□□)

□□□□□□[□□□□□□□] = □□□□ □□□□□□□ / (□□□□ □□□□□□□ + □□□□ □□□□□□□ + □□□□ □é□□□□□)

Où Vrai Positif (ou Négatif, Neutre) est le nombre de phrases correctement classées et Faux Positif (ou Négatif, Neutre) est le nombre de phrases qui n'ont pas été classées à tout. Quant à la f-mesure, elle se calcule comme suit :

$$\square - \square\square\square\square\square\square = \frac{\square\square é\square\square\square\square\square * \square\square\square\square\square\square * 2}{(\square\square é\square\square\square\square\square + \square\square\square\square\square\square)}$$

Accuracy(Acc), est une mesure liée au nombre total d'éléments correctement identifiés par rapport à une quantité totale de données dans l'analyse que nous avons effectué. En d'autres termes, c'est un pourcentage de Tweets qui se sont déroulées correctement.

$$\square\square\square\square\square\square\square\square = \frac{\square\square\square\square\square\square \square\square \square\square\square\square\square\square é\square\square\square\square\square}{\square\square\square\square\square\square\square\square\square\square\square\square\square\square\square}$$

Cela nous faisons une analyse comparative entre les deux Méthodes SVM et K-NN. . En calculant, pour chaque méthode sa Précision et son Rappel, ainsi que F-mesure.

Pour mieux mesurer la qualité de la classification, nous avons présenté les résultats des deux Toutes ces mesures sont effectuées pour examiner les résultats des deux méthodes: K-NN et SVM.

Une Matrice de Confusion comme il est présenté dans les deux tableaux au-dessous (Tableau 7, Tableau 9). Cette matrice est un résumé des résultats de prédiction pour le problème de classification. Le nombre de prédictions correctes et incorrectes est résumé avec des valeurs de comptage selon les différentes classes « Positif », « Négatif », ou « Neutre ».

Nous utilisons la matrice de confusion pour la multi-classe .Les métriques standard utilisées en mode multi-classe sont les mêmes que celles utilisées dans le cas d'une classification binaire. La métrique est calculée pour chaque classe en le traitant comme un problème de classification binaire après avoir regroupé toutes les autres classes dans la seconde classe. Nous avons utilisé ce modèle de la matrice de confusion présenté dans le tableau ci- dessous :

		Tweets prédites par le système		
		Positive [1]	**Neutre [0]**	**Négative [-1]**
Tweets Actuelles Du Dataset	**Positive [1]**	Vrai Positif	Faux Neutre	Faux Négatif
	Neutre [0]	Faux Positif	Vrai Neutre	Faux Négatif
	Négative [-1]	Faux Positif	Faux Neutre	Vrai Négatif

Tableau 5 : Un modèle de la Matrice de Confusion

2.1 Pour la méthode SVM

La méthode SVM est appliquée avec les fonctionnalités proposées (TF-IDF). En fait, l'évaluation a été effectuée en utilisant les métriques communes de la recherche d'informations qui sont la Précision, le Rappel et la F-mesure. Comme nous l'avons montré dans le tableau 6.

La moyenne précision, rappel et f-mesure obtenus par la méthode SVM sont respectivement de 80% ,77% ,79%pour chacune des mesures utilisées.

	Précision	**Rappel**	**F-mesure**
Négative [-1]	70%	66%	68%
Positive [1]	71%	76%	73%
Neutre [0]	100%	89%	94%
Moyenne	80%	77%	79%

Tableau 6: Précision, Rappel et F-mesure pour la méthode SVM

	Négative [-1]	**Neutre [0]**	**Positive [1]**
Négative [-1]	1012	0	514
Neutre [0]	17	399	31
Positive [1]	426	1	1346

Tableau 7: Matrice de Confusion pour la méthode SVM

D'après les deux tableaux ci-dessus (Tableau 6, Tableau 7), nous pouvons déduire que Accuracy de la méthode SVM est 73,59 %. Cette mesure présente la quantité de Tweets qui ont été analysés déroulés correctement dans notre processus. Ce résultat est calculé comme suit :

$$\square\square\square\square\square\square\square\square\square(\square\square\square) = \frac{\square\square\square\square\square\ \square\square\square\ \square\square\square\square\square\square\ \square\square é\square\square\square\ \square\square\square\square\square\square\square\square\square\square\square\square\square\square\square}{\square\square\square\square\square\ \square\square\square\ \square\square\square\square é\square\square\ \square\square\square\square\square\square\square é\square\square}$$

Tels que :

- Total des données utilisés = 3746
- Totale des Tweets prédits correctement = 1012+399+1346=2757
- Accuracy(SVM)= 2757 / 3331= 0,7359

D'après le tableau 6 nous avons trouvé que la Précision des phrases Positives et Neutre est supérieure à leur Rappel. Tandis que la Précision des phrases négatives est inférieure à leur Rappel. En d'autres termes, ces résultats sont justifiés par le fait que nous avons plus de deux classes. Lorsque la méthode est incapable de détecter la bonne classe d'une phrase. Cette dernière est automatiquement classée dans la classe erronée. Par conséquent, toute correspondance incorrecte peut augmenter le FP (Faux Positif) de la classe ce qui conduit à augmenter le FN (Faux Négatif et Neutre) des autres classes de la même manière.

2.2 Pour la méthode K-NN

La méthode K-NN est appliquée avec les fonctionnalités proposées (TF-IDF) sur les données. Le tableau 8 montre les résultats obtenus:

	Précision	Rappel	F-mesure
Positive [1]	98%	22%	45%
Négative [-1]	85%	30%	45%
Neutre [0]	16%	100%	28%
Moyenne	66%	51%	36%

Tableau 8: Précision, Rappel et F-mesure pour la méthode K-NN

	Négative [-1]	Neutre [0]	Positive [1]
Négative [-1]	446	1008	10
Neutre [0]	0	455	0
Positive [1]	79	1337	411

Tableau 9: Matrice de Confusion pour la méthode K-NN

D'après les deux tableaux ci-dessus (Tableau 8, Tableau 9), nous pouvons déduire que l'Accuracy de la méthode K-NN (=35.02%). Cette mesure présente la quantité de Tweets qui ont été analysé correctement dans notre processus. Ce résultat est calculé comme suit :

$$\text{Accuracy}(K\text{-}NN) = \frac{\text{Total des Tweets prédits correctement}}{\text{Total des données utilisés}}$$

Tels que :

- Total des données utilisés = 3746
- Total des Tweets prédits correctement = 446+455+411=1312
- Accuracy (K-NN)= 1312 / 3331= 0,3502

D'après le tableau 9, nous avons trouvé que la Précision du Négatif et Neutre Tweets sont plus que ses Rappel, tandis que la précision des phrases Négatives est inférieure à son rappel. Dans un autre sens, dans ce système nous avons plus de deux classes, quand la méthode est classé incorrectement une phrase automatiquement il classera dans les autres classes. Par conséquent, toute correspondance incorrecte peut augmenter le Faux Tweets trouvés de la classe, conduit à

augmenter le Faux des autres classes de la même manière. Vice versa, toute correspondance incorrecte peut réduire le Faux de la classe, conduit à réduire le Faux classe des autres classes de la même manière.

	Précision	**Rappel**	**F-mesure**
SVM	80%	77%	79%
K-NN	66%	51%	36%

Tableau 10: Précision, Rappel et F-mesure pour SVM et K-NN

Certains remarques pour les méthodes SVM et K-NN que nous trouvons :

D'après la Précision moyenne, le Rappel et la f-mesure que nous avons obtenus par les deux méthodes K-NN et SVM (présenter dans le Tableau 10). Nous trouvons que la Précision est supérieure au Rappel dans les deux méthodes ça veut dire que ces deux méthodes sont presque plus précises et pertinents.

Comme il est montré dans le tableau 10, SVM a surpassé l'autre méthode en atteignant 80% de précision, 77% rappel et79% de la F-mesure. Ceci peut être justifié par le fait que SVM.

Nous trouvons que SVM surpasse lorsque nous rencontrons de nouvelles données non prédites, contrairement au K-NN.

La méthode K-NN est appelée k-voisin le plus proche car la classification dépend des k voisins les plus proches. Alors nous avons obtenus que si nous descendons pour la méthode K-NN, l'augmentation de la valeur k augmente le nombre de voisins, ce qui peut entraîner une diminution des performances. D'après d'autres études ils trouvent qu'il est préférable de prendre k le nombre de voisin comme un nombre impair.

EN plus nous apercevons que K-NN est moins intensif d'informatique que SVM, puisque K-NN est facile à implémenter.

Une autre remarque mauvaise pour cette méthode K-NN, nous trouvons que que la métrique de distance est calculée à chaque fois si nous rencontrons un ensemble de nouvelles données non prédites alors le K-NN ne fonctionne pas rapidement et elle perdus beaucoup de temps et ceci représente le point faible.

En outre, d'après les résultats et les observations finales, nous pouvons juger que le corpus utilisé la méthode SVM est plus fiable. L'algorithme qui garantit une détection fiable dans des situations

imprévisibles dépend des données. Si les points de données sont hétérogènes distribués, les deux devraient bien fonctionner. Si les données sont homogènes à regarder, on pourrait être en mesure de mieux classer en mettant un noyau dans le SVM. Pour la plupart des problèmes pratiques, K-NN est un mauvais choix car il évolue mal s'il y a un million d'exemples étiquetés, il prendrait beaucoup de temps pour trouver les K voisins les plus proches. La figure 24 montres la comparaison entre les deux méthodes.

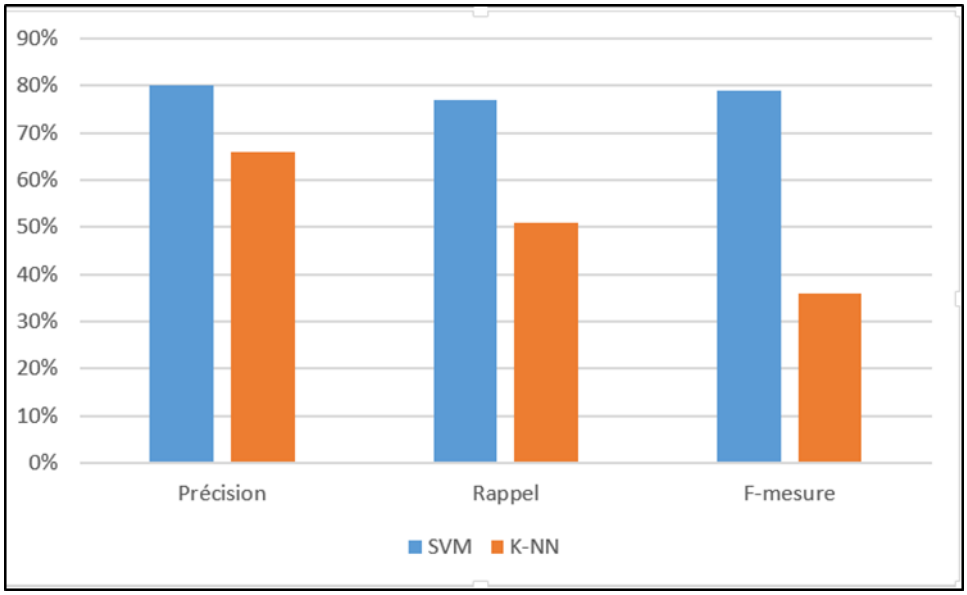

Figure 24: La comparaison entre les deux Méthodes SVM et K-NN

3 Matériels et Bibliothèques utilisées

3.1 Matériels utilisés

3.1.1 Langage de programmation (Python)

Dans ce projet on a utilisé le langage de programmation Python. Il est l'un des langages de programmation les plus intéressants du moment. Facile à apprendre, python est souvent utilisé en exemple lors de l'apprentissage de la programmation [31].

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.

L'interpréteur Python et sa vaste bibliothèque standard sont disponibles librement, sous forme de sources ou de binaires, pour toutes les plateformes majeures depuis le site Internet https://www.python.org/ et peuvent être librement redistribués.

L'interpréteur Python peut être facilement étendu par de nouvelles fonctions et types de données implémentés en C ou C++. Python est également adapté comme langage d'extension pour personnaliser des applications.

Lors de la création de la Python Software Foundation, en 2001, et durant les années qui ont suivi, le langage Python est passé par une suite de versions que l'on a englobées dans l'appellation Python 2.x (2.3, 2.5, 2.6…). Depuis le 13 février 2009, la version 3.0.1 est disponible et la version la plus récente étant la version 3 .9 .

Python est à la fois simple et puissant, il permet d'écrire des scripts très simples mais grâce à ses nombreuses bibliothèques, on peut travailler sur des projets plus ambitieux.

* Web: Aujourd'hui python combiné avec le framework Django est un très bon choix technologique pour des gros projets de sites internet.

* Système : Python est également souvent utilisé par les admins du système pour créer des tâches dites répétitives ou simplement de maintenance. D'ailleurs si on veut créer des applications java en codant en python, c'est possible grâce au projet Jython.

De plus l'environnement de python est riche en différentes bibliothèques comme streamlit, nltk et tweepy etc. Ces bibliothèques facilitent notre développement et donnent des bons résultats.

3.1.2 L'IDE Pycharm

Nous avons utilisé pour notre programmation, le Pycharm IDE pour implémenter notre projet en langage Python

PyCharm est un environnement de développement intégré utilisé pour programmer en Python [30].

Il permet l'analyse de code et contient un débogueur graphique. Il permet également la gestion des tests unitaires, l'intégration de logiciel de gestion de versions, et supporte le développement

web avec Django.

Développé par l'entreprise tchèque JetBrains, c'est un logiciel multi-plateforme qui fonctionne

sous Windows, Mac OS X et Linux. Il est décliné en édition professionnelle, diffusé sous licence propriétaire, et en édition communautaire diffusé sous licence Apache.

3.2 Différentes bibliothèques utilisées

Python a une communauté active dans laquelle la plupart des développeurs créent des bibliothèques pour leurs propres besoins et les publient plus tard au public à leur avantage. Voici quelques-unes des bibliothèques de Machine Learning courantes que nous avons utilisé dans ce projet pour implémenter une application d'Analyse des Sentiments pour des Tweets en arabe [29]:

- **Scikit-learn :** elle dispose d'un large éventail d'algorithmes d'apprentissage supervisés et non supervisés qui fonctionnent sur une interface cohérente en Python. La bibliothèque peut également être utilisée pour l'exploration de données et l'analyse de données. Les principales fonctions d'apprentissage automatique que la bibliothèque Scikit-learn peut gérer sont la classification, la régression, la mise en cluster, la réduction de dimensionnalité, la sélection de modèle et le prétraitement.
- **Pandas :** elle est en train de devenir la bibliothèque Python la plus populaire utilisée pour l'analyse de données avec prise en charge de structures de données rapides, flexibles et expressives conçues pour fonctionner à la fois sur des données « relationnelles » ou «étiquetées»
- **Numpy :** Numpy est une extension du langage de programmation Python, destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux.

  Plus précisément, cette bibliothèque logicielle libre et open source fournit de multiples fonctions permettant notamment de créer directement un tableau depuis un fichier ou au contraire de sauvegarder un tableau dans un fichier, et manipuler des vecteurs, matrices et polynômes.

- **Joblib :** Joblib est un ensemble d'outils pour fournir un lightweight pipelining en Python. En particulier, il permet une mise en cache transparente sur disque des fonctions et (modèle de mémorisation). Joblib est optimisé pour être rapide et robuste en particulier sur des données volumineuses et possède des optimisations spécifiques pour les tableaux Numpy. Il est sous licence BSD [29].
- **Flask:** Flask est un framework application Web lightweight. Il est conçu pour faciliter et accélérer la mise en route, avec la possibilité de s'adapter à des applications complexes. Il a commencé comme un simple wrapper autour de Werkzeug et Jinja et est devenu l'un des frameworks d'application Web Python les plus populaires [29].

Par ailleurs, nous avons utilisé d'autres bibliothèques, notamment : **sklearn-features , sklearn_Dataset , sklearn metrics.**

4 Présentation de l'application et le code source

4.1 Le code source

- Le code source du prétraitement (certaines fonctions) :

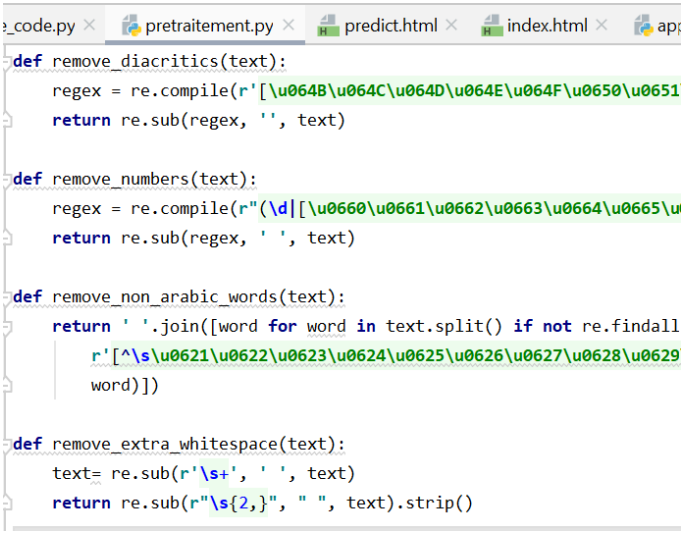

Figure 25: Le code source du prétraitement

- Le code source pour présenter les techniques d'extraction des caractéristiques : Bag Of Words et TF-IDF, N-Grams :

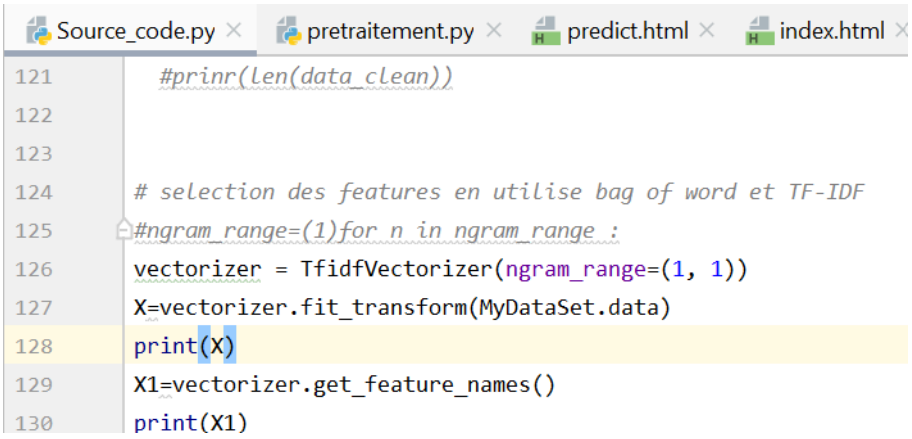

Figure 26: Bag Of Words et TF-IDF, N-Grams

- Le code source présente la division du Dataset (« Train Data » et « Test Data ») et la création des modèles SVM et K-NN :

```
# # #faire l'apprentissage
# # #LinearSVC
LSVC=LinearSVC().fit(X_train, target_train)

# # k-nearest neighbor
KNN = KNeighborsClassifier(3)
KNN.fit(X_train, target_train)

# # # faire le test pour chaque model(predict l'ensemble de  test (test_set))
print('\n\n********************LSVC CLASSIFIER*************\n\n')

# # #lsvc
#target_names=['1', '0','-1']
y_pred_lsvc=LSVC.predict(X_test)
print("\nResults pour  LSVC...")
```

Figure 27: création des modèles SVM et K-NN

4.2 Présentation les interfaces d'application

- La page d'accueil

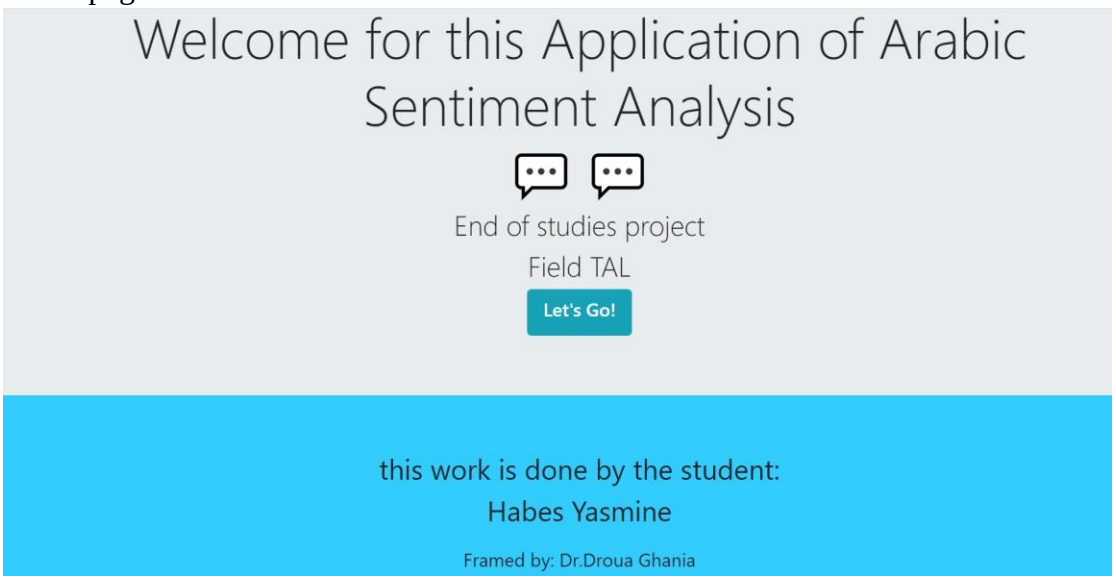

Figure 28 : La page d'accueil du notre application

- La page d'Analyse des Sentiments pour les Tweets

Figure 29: La page d'Analyse des Sentiments pour les Tweets

5 Conclusion

Dans ce chapitre, nous avons présenté l'essentiel de notre travail qui consiste à créer un système d'analyse d'opinion pour détecter les sentiments dans le réseau social Twitter. Pour l'implémentation, nous avons utilisé l'une des méthodes de classification les plus connue SVM et K-NN du Machine Learning. Pour l'implémentation, nous avons choisi des outils et matériels et différentes bibliothèques de python pour développer cette application en examinant les Tweets pour les Méthode. Notre système s'intègre dans le domaine de l'intelligence artificielle précisément "Machine Learning".

Conclusion Générale

Synthèse

Dans ce travail, nous avons exploré le domaine de l'Analyse des Sentiments qui, comme tous les autres domaines du traitement du langage naturel, a connu une évolution majeure depuis les années 2000 et a réalisé une évolution majeure et un grand intérêt depuis la naissance de l'apprentissage automatique.

L'Analyse des Sentiments arabes est considérée comme l'un des domaines en croissance qui concernent la découverte des sentiments des individus vers un événement spécifique, une marque ou autre chose. Il est clair que le domaine de l'Analyse des Sentiments pour l'Arabe en est à ses débuts mais les recherches dans ce domaine ont rapidement augmenté au cours des dernières années.

Afin d'atteindre ces résultats, nous avons passé beaucoup de temps à lire et réviser des publications, des articles et des livres pour voir et comprendre les concepts et comment appliquer un modèle de Machine Learning à notre problème.

Dans cette enquête, nous avons étudié l'Analyse des Sentiments arabes et élaboré une littérature complète, examen de ce sujet important. Nous proposons un système de classification subjective des opinions des utilisateurs du réseau social Twitter sur un produit ou un événement en trois catégories : positif, négatif et neutre. Nous sommes basés sur l'apprentissage automatique par l'application des différents modèles de classification comme SVM, et K-NN qui sont les algorithmes les plus utilisés.

Perspectives

Notre système d'Analyse des Sentiments est encore en phase de développement et loin d'être complet. Pour l'amélioration, nous proposons :

- Tester notre modèle sur d'autres ensembles de données.
- L'utilisation des approches hybrides sur la base des méthodes de Machine Learning pour avoir de meilleurs résultats.
- Elargir l'utilisation de ce travail vers d'autres objectifs comme l'analyse des tendances et l'extraction de connaissances à partir des réseaux sociaux.
- Il serait intéressant de pouvoir utiliser notre approche pour classer les sentiments dans d'autres dialectes comme darija algérienne ou encore d'autres langues comme « amazighia ».

Bibliographie

[1] (03 Mars 2020) Liddy, E. (2015). Scholarship 2 1 Natural Language Processing.

[2] (03 Mars 2020) Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets", 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-175.

[3] (16 Mars 2020) hang L., Liu B. (2016) Sentiment Analysis and Opinion Mining. In: Sammut C., Webb G. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA.https://doi.org/10.1007/978-1-4899-7502-7_907-1

[4] Paul-Louis Valat (12 octobre 2020) « Comment fonctionne l'analyse de sentiment? » dans le titre Le guide de l'analyse de sentiment: fonctionnement et bonnes pratiques, sur le site www.meltwater.com .consulté le 15 octobre 2020 https://www.meltwater.com/fr/blog/analyse-sentiment

[5] mehdihadji " l'analyse des sentiments" dans le titre "les avantages de l'analyse des sentiments" consulté le 05 Mars 2020 ,sur le site https://medium.com/@mehdihadji/analyse-des-sentiments

[6] Sentiment140 2013, dans le titre "About" consulté 20 octobre 2020, sur le site http://www.sentiment140.com

[7] http://www.Tweetfeel.com

[8] Reed "TwitrRatr",dans le titre "Twitratr"consulté le 20 Octobre 2020 , sur le site http://twitrratr.com/

[9] http://smm.streamcrab.com

[10] (16 Mars 2020)Alrefai, Mo'ath & Faris, Hossam & Aljarah, Ibrahim. (2018). Sentiment analysis for Arabic language: A brief survey of approaches and techniques.

[11] Mahdjoubi "l'analyse du sentiment utilisant le DEEP Learning" consulté le 04 Janvier 2021 sur le site https://pmb.univ-saida.dz/butecopac/doc_num.php?explnum_id=870

[12] Wikimedia "Précision et Rappel" dans les titres " Précision ,Rappel" consulté le 02 Janvier 2021 sur le site https://fr.wikipedia.org/wiki/Pr%C3%A9cision_et_rappel

[13] (16 Mars 2020) H. Hilali-;-hilali, Application de la classification textuelle pour l'extraction des règles d'association maximales, 2009.

[15] Pierre Bidet "les réseaux sociaux" ,consulté le 20 Septembre 2020 sur le site https://www.mercator-publicitor.fr

[16] https://www.definitions-marketing.com

[17] overblog 2012 "les réseaux sociaux" sur le site http://nmstpe.over-blog.com

[19] Alwine Lambert, Gabriel Bellard, Guillaume Lorre , Karim Kouki , Analyse des SentimentsTwitter

[21] wikiMedia « Twitter » dans le titre « Tweet » , consulté le 15 Mars 2020 sur le site www.wikipedia.org

[22]M. Severo and R. Lamarche-perrin, L'analyse des opinions politiques sur Twitter, Revue française de sociologie, vol.59, issue.3, p.507, 2018

[23](05 Aout 2020)Abu Farha, Ibrahim & Magdy, Walid. (2019). Mazajak: An Online Arabic Sentiment Analyser. 10.18653/v1/W19-4621.

[24] (13 Septembre 2020) Vishal.A.Kharde, Prof. Sheetal.Sonawane. (April 2016).Sentiment Analysis of Twitter Data. A Survey of Techniques International Journal of Computer Applications139(11): 5-15.10.5120/ijca2016908625.ArXiv:1601.06971.

[27] (20 Octobre 2020) Stefania Pecore. Analyse des sentiments et des émotions de commentaires complexes en langue française. Linguistique. Université de Bretagne Sud, 2019. Français. ⟨ NNT : 2019LORIS522⟩ . ⟨ tel-02903247⟩

[28] (20 Octobre 2020) Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," in Proc. - 2017 2nd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2017, vol. 2018-January, pp. 294-298, 2018

[29] Python software foundation « Find, install and publish Python packages with the Python Package Index», dans le titre « the python package » sur le site www. pypi.org .consulté le 03 septembre 2020 https://pypi.org/

[30] Wikimedia « Pycharm » dans le titre « pycharm » consulté le 04 Aout 2020 sur le site https://fr.wikipedia.org/wiki/PyCharm

[31] Python software foundation « le tutoriel python », sur le site www. python.org .consulté le 03 septembre 2020 https://docs.python.org/fr/3/tutorial

[32] Baeldung «Multiclass classification using support vector machines », dans le titre « Classification », sur le site www.baeldung.com . Consulté le 03 septembre 2020 https://www.baeldung.com/cs/svm-multiclass-classification

[33](05 Janvier 2021)Grzegorz Dziczkowski. Analyse des sentiments : système autonome d'exploration des opinions exprimées dans les critiques cinématographiques. Automatique / Robotique. École Nationale Supérieure des Mines de Paris, 2008. Français. ffNNT : 2008ENMP1637ff. fftel-00408754f

[34](05 Janvier 2021) Habash, N. (2010). Introduction to Arabic natural language processing. Synth Lect Hum Lang Technol, 3(1), pp. 1–187.