REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE IBN KHALDOUN - TIARET

MEMOIRE

Présenté à :

FACULTÉ MATHEMATIQUES ET INFORMATIQUE
DÉPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

MASTER

Spécialité : Génie Logiciel

Par :

BADRANE Sarra

Sur le thème

Exploitation des contenus sociaux pour l'analyse des sentiments "en dialecte Algérien"

Soutenu publiquement le 23 / 09 / 2021 à Tiaret devant le jury composé de :

Mr SI ABDELHADI Ahmed	M.A.A	Université Ibn-Khaldoun Tiaret	Président
Mme LAKHDHARI Aicha	M.A.A	Université Ibn-Khaldoun Tiaret	Encadreur
Mr KHARROUBI Sahraoui	M.C	Université Ibn-Khaldoun Tiaret	Examinateur

2020-2021

Dédicace

“

Je dédie ce modeste travail :

À mes anges gardiens, ***mes parents*** *qui sont la source de mon bonheur au quotidien, je suis immensément fier d'être votre fille et je ferai de mon mieux pour que vous ressentiez la même chose. Vos encouragements, vos motivations, votre aide et vos efforts sont au-delà de toute description, vous êtes les meilleurs parents du monde.*

À mes sœurs ***Nour elhouda ,Belkiss et Hadjer*** *et mes frères* ***Mohammed elfateh*** *et* ***Khalil Rahmane*** *qui ont toujours été si aimants et encourageants .*

À mes chères amies qui m'a encouragé et soutenu de tout leurs cœur et tous ceux que je connais de près ou de loin .

”

- Sarra

Tout d'abord, **الحمد لِلّه** (merci Allah) qui me donne la santé et la capacité de réaliser et de présenter cette thèse et pour m'avoir donné la puissance de lutter dans cette vie.

Je tiens à exprimer ma profonde gratitude à **Mme. LAKHDARI Aicha** , pour avoir encadré et dirigé mes recherches. Je la remercie pour m'avoir soutenu et appuyé tout au long de ma thèse. Ses précieux conseils, son exigence et ses commentaires ont permis d'améliorer grandement la qualité de mes travaux.

Nous présentons également tous nos respects et nos sincères remerciements aux membres de jury **Mr.KHAROUBI** et **Mr.SIABDELHADI** qui ont accepté d'évaluer notre travail.

Nous adressons nos remerciements à tous les professeurs, pour leurs conseils et leurs critiques qui ont guidé nos réflexions durant nos recherches, et nous remercions également tous nos collègues et amis du département d'informatique de l'université Ibn Khaldoun Tiaret.

Nous souhaitons exprimer nos profondes gratitudes nos parents qui nous ont soutenus tout au long de notre projet, ainsi que toute la famille, les amis pour leur soutien indéfectible. Enfin, on remercie tous ceux qui nous ont aidés de près ou de loin dans l'élaboration de ce travail.

Résumé

L'Internet est devenu un besoin de plus en plus important dans notre vie quotidienne. Grâce au web 2.0, les utilisateurs de l'Internet deviennent des producteurs de contenu plus que de simples consommateurs, la recherche et l'industrie se sont mises en quête des moyens pour analyser automatiquement les opinions et émotions exprimées à travers les contenus sociaux sur les réseaux sociaux.

L'analyse des sentiments sur internet est exploitée dans de nombreux domaines. En commerce électronique, elle aide à évaluer les produits et services et à analyser l'importance et la réputation numérique de ses produits sur la base des avis et commentaires laissés par les clients. Elle aide encore au positionnement en permettant d'identifier la clientèle cible d'une entreprise de production, ou de déterminer le succès d'une campagne de communication et sert d'appui aux systèmes de recommandation (ne pas publier des produits qui ont des mauvaises notes par exemple). En politique, elle permet d'évaluer une politique et lors d'élections d'analyser les opinions des citoyens sur les candidats. Dans le domaine cinématographique encore, elle permet l'analyse des critiques (reviews) des films.

Tant de nouvelles questions ont permis aux travaux sur la fouille d'opinions, que ce soit au niveau académique ou industriel, de prendre de plus en plus d'ampleur, mais les recherches menés sur SA en Arabe reste insuffisantes par rapport aux autres langues et surtout en Algérien .

L'analyse des sentiments (SA), ou fouille d'opinions, est une branche du traitement automatique des langues naturelles NLP. SA exploite les approches et les techniques de traitement standard et celles de l'apprentissage automatique pour analyser les opinions, les sentiments, les attitudes et les émotions des interlocuteurs envers des entités telles que les produits, les services, les organisations, les individus, les problèmes, les événements.

Dans ce projet de fin d'études , nous nous intéressons à ces techniques, méthodes et approches pour analyser le sentiment des commentaires-Youtube et des formulaires-Google écrits en arabe et en dialecte Algérien avec les deux scripts arabe et arabizi sur le domaine politique (Hirak) .

Mots clés : Analyse de sentiment , Dialect Algérien, Apprentissage automatique ,Hirak Algérien,arabe ,arabizi.

Table des matières

Table des figures

Liste des tableaux

Liste des sigles et acronymes

SA	*Sentiment Analysis*
BOW	*Bag Of Words*
API	*Application Programming Interface*
IA	*Intelligence artificielle*
ML	*Machine Learning*
KNN	*Support Vector Machine*
SVM	*K-Nearest Neighbor*
DT	*Decision Tree*
NB	*Naive Bayes*
LR	*Logistic Regression*
NLP	*Natural Language Processing*
TALN	*Traitement Automatique De La Langue Naturelle*
DALG	*Dialect Algérien*
SC	*Sentiment Classification*
TF	*Term Frequency*
TF-IDF	*Term Frequency-Inverted Document Frequency*
WE	*Word Embedding*

Introduction générale

Contexte

Le dialecte Algérien est une langue vivante et est utilisée quotidiennement par les interlocuteurs dans tous les comportements de la société et les dialogues familiaux ou autres.

Le dialecte Algérien a évolué à travers le temps tout comme les autres langues, l'apparition de l'écriture en Algérien est récente, seul l'arabe et le français sont les langues de l'écrit en Algérie. Actuellement, les textes publicitaires, les annonces dans les journaux, les communications dans les réseaux sociaux et les messages téléphoniques (SMS) se font en partie en Algérien. Cette évolution et l'usage quotidien de cette langue nous pousse à observer et étudier le fonctionnement du dialecte.

La langue arabe est utilisée à l'écrit au même titre que le français (journaux, documents officiels, et littérature) étant donné que l'Algérie est un pays bilingue. Le dialecte Algérien a toujours existé, c'est la langue de communication à la maison avec la famille, dans la rue avec les amis, mais ce dialecte a acquis une importance primordiale ces dernières années, car il s'utilise de plus en plus à l'écrit dans plusieurs domaines (publicité, annonce, sous-titrage, revue, radio, etc.). Instinctivement l'Algérien parle son dialecte spontanément, contrairement à l'arabe standard, surtout via son espace virtuelle (les medias sociaux). Cependant, avec l'émergence du Web social ou le Web 2.0, les plateformes sociales, qui font partie de notre quotidien, comme Facebook, Twitter, Instagram, LinkedIn,Youtube... permettent aux internautes d'exprimer leurs sentiments et leurs opinions sur un sujet particulier (politique, commercial ou individuel).

L'aspect des données de ces plateformes, telles que les commentaires Youtube , génèrent une riche fouille de données sur les acteurs et les sujets impliquées dans la communication. Un gouvernement quelconque souhaite toujours connaitre les points de vue et les préoccupations de ces citoyens et de refléter leurs humeurs et leurs opinions par rapport à l'actualité et à la politique instauré. De même, des entreprises qui souhaitent connaître les réactions des consommateurs vis-à-vis de leurs produits. Cela pourrait les aider à changer de stratégie pour améliorer la qualité de leurs produits et celle de leurs services. En effet, Ces données vont jouer un rôle important dans la prise de décision pour de nombreuses entreprises et organisations. Pour cette raison, il est important d'analyser le contenu des réseaux sociaux qui ont acquis une grande popularité dans le monde entier.

L'analyse des sentiments SA(Sentiment Analysis ou Opinion Mining) forme une méthode de traitement automatique du langage naturel (Naturel Langage Processing) qui

tente de localiser la présence des sentiments ou d'émotions exprimés dans un texte (statut, post, tweet, commentaire, formulaire, ...etc).

Motivation

-Actuellement, les textes publicitaires, les annonces dans les journaux, les communications dans les réseaux sociaux et les messages téléphoniques se font en partie en Algérien. Cette évolution et l'usage quotidien de cette langue nous pousse à observer et étudier son fonctionnement.

-Le développement technologique en Algérie nous permet d'observer encore plus l'usage du dialecte Algérien en comparaison aux autres langues qui s'utilisaient beaucoup plus dans le passé. Le français était la langue utilisée par les Algériens pour la communication de masse, sauf que ces dernières années l'usage du dialecte algérien s'est de plus en plus généralisé surtout via les medias sociaux.

-Le manque de ressources et de collection de données (dataset) standard est la cause principale qui a motivé la création d'un dataset d'opinions pour l'arabe standard et le dialecte Algérien dans notre travail.

Problématique

Les recherches effectuées sur l'analyse du sentiment en langue arabe sont presque limitées, en particulier le dialecte Algérien par rapport à d'autres langues comme l'anglais et le français. Cela peut s'expliquer :

- Pour l'analyse du sentiment, il n'existe aucune ressource standard (Dataset) annotée (étiquetée) pour **le dialecte Algérien**
- La langue vernaculaire, qui est connu comme 'āmmiya ou dardja', est une langue rassemblant plusieurs variétés d'arabe dialectal parlées en Algérie . L'arabe Algérien a de nombreux dialectes et accents régionaux, ce qui rend la tâche d'analyse des textes extraits de la plateforme Youtube et des formulaires Google très difficile.
- La complexité du traitement des **données Youtube** : le dialecte Algérien n'est associé à aucune forme d'écriture normalisée et contient du bruit, des fautes d'orthographe, des abréviations, des répétitions, et des mots qui ne suivent aucune règle grammaticale.
- Le manque de codification et de consentement entre chercheurs. Le dialecte Algérien est écrit soit en lettre latine (arabisé) ou en lettre arabe, et cela dépend de l'interlocuteur et sa préférence linguistique ou son niveau éducatif, et de la variété d'arabe dialectal parlées en Algérie selon les régions.
- Enfin, le problème du classement de la polarité (positive, négative, neutre) à partir de données textuelles qui est une tâche très difficile et coûteuse en raison de la grande quantité de données bruitées.

Objectif du projet et solutions proposées

Nous nous intéressons plus particulièrement à développer un outil d'analyse de sentiment de contenus sociaux (texte) extraits du réseau social Youtube et des formulaires de Google, où le dialecte algérien est utilisé comme support de communication. Ce système va permettre de collecter, traiter, classer la polarité des textes selon la catégorie (positifs, négatifs) et de lister les statistiques des sentiments qui dominent la communication des internautes Algériens du Hirak en utilisant des techniques de l'apprentissage automatique.

En outre, Notre dataset sur le Hirak est crée et conçu manuellement par nous même à partir de Youtube et des formulaires Google, Il est sous forme d'un ensemble de contenus textuels collectés, chaque contenu social représente un commentaire. Le dataset contient 9108 commentaires,35562 mots et 14362 vocabulaire.

L'étiquetage ou l'annotation des opinions est une tâche humaine (manuelle) qui nécessite d'énormes efforts. Et afin de réaliser ce processus, pas mal d'amis-abonnés ont travaillé sur l'annotation de ces commentaires collectés. Pour cela nous étions amenés à construire notre propre ressource des commentaires Algériens afin de déterminer par la suite la polarité des sentiments exprimée là-dedans.

Organisation du mémoire

Afin de parvenir à notre but, le présent document est divisé en deux volets. Un volet théorique ainsi qu'un volet pratique.
Le volet théorique (partie 1) dénommée « Etat de l'art » comprend les chapitres 1, 2 et 3.

Chapitre 1 :Généralités sur l'analyse de sentiments

Une étude sur les généralités de l'analyse des sentiments est présentée.

Chapitre 2: Généralités sur l'apprentissage automatique
Des généralités sur l'apprentissage automatique, son positionnement, ses domaines d'application et ses types d'apprentissages accompagnés d'algorithmes sont clairement décrits.

Chapitre 3 : Approches et langages d'analyse de sentiment
Un survol sur les principales approches et techniques orientées classification des sentiments en ML est présenté. Et Enfin, le chapitre se termine par une synthèse de travaux de recherches sur les dialectes arabes.

Le volet pratique dénommée « État des lieux ,Conception et Mise en œuvre » comprend les chapitres 4, 5 et 6:

Chapitre 4: Etat des lieux
Le contexte de notre étude de cas est sur le " Hirak d'Algériens". Sa chronologie, son linguistic et aussi ses catégories d'opinions sont bien présentés dans cet état des lieux.
Chapitre 5: Conception de notre outil d'AS
La conception de notre outil d'AS passe automatiquement par traitement, analyse et

classification des données textuelles extraites de Youtube et de formulaires Google.
Chapitre 6: Réalisation, tests et résultats
Les différentes technologies et bibliothèques utilisées marque l'accomplissement de notre travail et expose les résultats de l'évaluation de notre système d'AS avec le plan de test suivi d'une synthèse et de la conclusion du chapitre.

Première partie

Volet théorique : État de l'art

Chapitre 1

Généralités sur l'analyse de sentiments

1.1 Introduction

Nous aborderons dans ce chapitre les détail sur le sujet de l'analyse des sentiments notamment : termes, applications, approches et les défis.

1.2 Notions de base

Lorsqu'on aborde le domaine de l'analyse de sentiments, l'une des premières questions à se poser pourrait être la suivante : Quelle est la diférence entre un sentiment et une opinion et la relation avec l'emotion ?

Pour y répondre, nous allons tout d'abord définir le sentiment , puis l'émotionet enfin l'opinion.

1.2.1 Sentiment

Larousse [1] définit le sentiment comme étant « un état affectif durable lié à certaines émotions ou représentations ».

Le sentiment est définit comme «la composante de l'émotion qui implique les fonctions cognitives de l'organisme et la manière d'apprécier. Le sentiment est à l'origine d'une connaissance immédiate ou d'une simple impression ». [24]

1.2.2 Émotion

L'émotion est une expérience psychophysiologique complexe et intense avec un début brutal et une durée relativement brève.[4]

[1] https ://www.larousse.fr/

Robert Plutchik, était un leader d'opinion dans l'étude des émotions, il a conçu la théorie psycho-évolutionnaire de l'émotion (*Fig.*1.1), ce qui permet de catégoriser les émotions en émotions primaires et les réponses afférantes .[18]
Il a soutenu que les émotions primaires sont un développement évolutif et que la réponse à chacune de ces émotions est celle qui est susceptible d'offrir le plus haut niveau de survie possible :

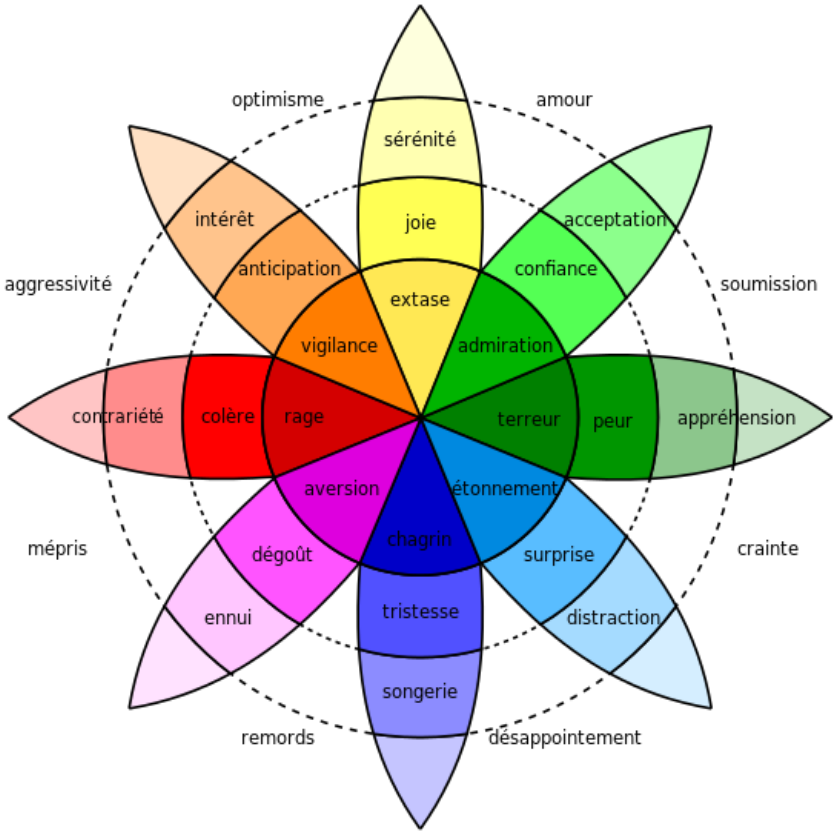


FIG. 1.1 : Roue des émotions de Plutchik

1.2.3 Opinion

Selon LAROUSSE, l'opinion est « Jugement, avis, sentiment qu'un individu ou un groupe émet sur un sujet, des faits, ce qu'il en pense», ou encore comme «Ensemble des idées d'un groupe social sur les problèmes politiques, économiques, moraux, etc.»

Une opinion définit par le quintuple suivant, dont les composants sont liés les uns aux autres : [15]

$$(e_i, a_ij, s_ijkl, h_k, t_l)$$

- e_i : entité i cible de l'opinion .
- a_{i_j} : aspect j de l'entité i.

- s_{ijkl} : sentiment exprimé sur l'aspect j de l'entité i par la source k dans le temps l.
- h_k : source de l'opinion .
- t_l : moment de l'opinion .

1.2.4 Type d'opinion

Les types d'opinion selon [15] :

1.2.4.1 Opinion ordinaire

ce type est tout simplement appelé opinion dans la littérature , on peut cependant distinguer deux types d'opinions :

- **Opinion directe :**désigne une opinion exprimée directement sur une entité ou un aspect d'une entité (exemple : L'écran de ce téléphone est impressionnant).
- **Opinion indirecte :** désigne une opinion exprimée indirectement sur une entité ou un aspect d'une entité basé sur d'une autre entité (exemple : Après avoir changé de type de carburant, la voiture roulait difécilement).

1.2.4.2 Opinion comparative

L'opinion comparative exprime une relation de similitude ou de diférence entre plusieurs entités, il existe deux types d'opinions comparatives :

- **Comparaison évaluée :** dans ce type de comparaison, il existe une préférence évidente entre les entités (exemple : la BMW est plus rapide que la Renault4)
- **Comparaison non évaluée :** dans ce cas, il existe une diférence entre les entités, cependant, on ne peut déterminer laquelle le détenteur de l'opinion préfère (exemple : La vitesse de cette BMW est diférente de la Renault4)

1.3 Analyse des sentiments SA

1.3.1 Définition

Dans la littérature, sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining,sont des termes utilisés pour désigner des technologies d'analyse automatique des discours, écrits ou parlés, afin d'en extraire des informations subjectives comme des jugements, des évaluations ou des émotions .[12]

La définition de AS comportant les domaines d'application ainsi que sa relation avec le TALN est «l'analyse des sentiments est le domaine d'étude qui analyse les opinions,

les sentiments,les évaluations, les attitudes et les émotions des gens vers des entités telles que des produits, des services, des organisations, des particuliers, des problèmes, des événements, des sujets, et leurs attributs».[22]

1.3.2 Tâches de SA

Les tâches de SA peuvent être différentes la détection de la subjectivité, la classification de l'orientation et les type d'analyse ;

1.3.2.1 Détection de subjectivité

Les documents qui expriment le point de vue de l'auteur sont subjectifs, et ceux qui sont factuels sont objectifs. La détection de la subjectivité consiste à classer les textes comme étant subjectifs ou objectifs.[14]

1.3.2.2 Classification de l'orientation

La classification de subjectivité (Subjectivity classification) est la tâche qui distingue les phrases exprimant des informations objectives des phrases exprimant des vues et opinions subjectives.[42]

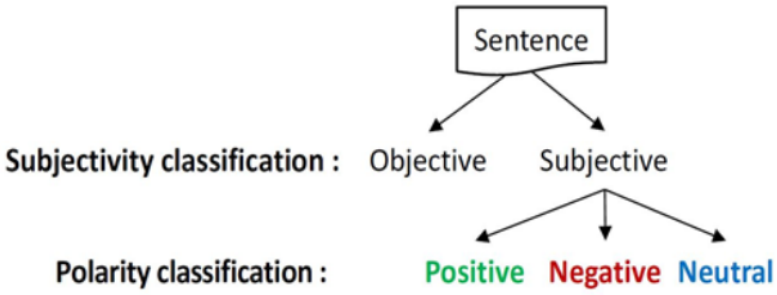


Fig. 1.2 : Classification de l'orientation

1.3.3 Types de SA

Il existe de nombreux types d'analyses de sentiments allant des systèmes qui se concentrent sur la classification de la polarité (positif, négatif, neutre) aux systèmes qui détectent des émotions (en colère, heureux, triste, etc.) ou identifient des intentions (par exemple, intéressé, pas intéressé). Dans la section suivante, nous aborderons les types les plus importants.[36]

1.3.3.1 Analyse fine des sentiments

Au lieu de parler de phrases positives, négatives ou neutres,on considère les catégories suivantes :

«Très positive , Positive,Neutre , Négative ,Très négative»

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

1.3.3.2 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.

1.3.4 Niveaux de SA

L'étude de l'analyse de sentiment peut s'effectuer à différents niveaux. La décomposition de niveau de granularité permet de simplifier l'analyse de sentiments du document global. [42]

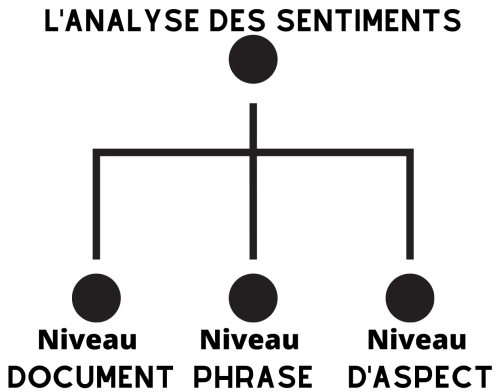


Fig. 1.3 : Niveux d'analyse de sentiment

1.3.4.1 Niveau Document

L'analyse des sentiments au niveau du document suppose que chaque document exprime des opinions sur une seule entité (ex : un seul produit).

1.3.4.2 Niveau Phrase

Le niveau phrase est une direction principale dans la fouille d'opinion ,[27] , il est considéré comme une étape intermédiaire dans un processus globale pour déterminer l'orientation sémantique du document entier.

La polarité du document est obtenue en synthétisant les scores de polarité des phrases subjectives avec un poids représentant l'importance de la phrase.
La subjectivité d'une phrase est obtenue par la présence de mots subjectifs d'un dictionnaire de langue, et aussi par un ensemble de règles de langage qui indiquent l'intensification .[9]

1.3.4.3 Niveau aspect (entité)

Le sentiment global que comporte un document n'est pas suffisant, et on cherche les aspects (caractéristiques ou parties) d'un objet (produits, sujet, ...) qui sont ciblés par les commentaires. [42]
En se basant sur ce niveau d'étude, il est possible d'avoir une structuration des opinions sur les entités et leurs différents aspects.
Dans l'exemple suivant cité dans [15] "The iPhone's call quality is good, but its battery life is short" l'évaluation concerne deux aspects ; la qualité d'appel (call quality) sur laquelle l'opinion est positive, et la durée de vie de batterie (battery life) dont un sentiment négatif est exprimé, et cela pour une même entité iPhone.

1.4 Domaines d'application de SA

L'importance de SA est présente dans plusieurs domaines ainsi plusieurs applications ont vu le jour dans ce contexte.

Quelques applications sont mentionées brièvement ci-dessous :

1.4.1 Domaine du commerce

L'analyse de sentiments est appliquée dans le but d'extraire le sentiment du marché.

Les vendeurs veulent avoir des informations sur l'opinion des clients à travers les commentaires sur les produits, pour utiliser cette information dans leur stratégie de marketing afin de combattre leurs concurrents. Les clients, à leurs tours, veulent avoir les opinions des autres clients pour guider leurs choix (produits, hôtels, vacances, ...) [42].

1.4.2 Domaine de santé

Le domaine de la santé est l'un des domaines les moins explorés dans les travaux d'analyse de sentiments.

Dans ce domaine, il est question de savoir que pensent les gens envers leurs médecins, leurs médicaments prescrits, leurs maladies et les traitements qu'ils subissent ?
SentiHealth-Cancer (SHC-pt) est un outil d'analyse de sentiments qui permet la détection de l'état émotionnel des malades de Cancer dans des communautés brésiliennes à travers leurs postes Facebook en langue portugaise .[38]

1.4.3 Détection de spam d'opinion

Tout le monde peut mettre n'importe quoi sur Internet, ce qui augmente les risques de spam sur le Web. Les internautes peuvent écrire ou envoyer des spams pour induire les utilisateurs en erreur.

1.4.4 Domaine politique

De nos jours, les médias et les populations s'intéressent bien avant le début des élections à connaitre l'élu du peuple et le promis à un siège important du gouvernement.

L'avis populaire étant longuement caché dans la presse et dans les cafés voit maintenant le jour avec la viralité (diffusion rapide) des réseaux sociaux.

1.5 Problèmes de SA

Le domaine de SA est au départ lié au traitement du langage naturel TALN,ce domaine presente quelques problème d'analyes :

1.5.1 Détection de la subjectivité

Il s'agit de différencier le texte avec opinion et sans opinion,i est utilisé pour améliorer les performances du système en incluant un module de détection de subjectivité pour filtrer les faits objectifs,mais cela est souvent difficile à faire.
Considérez les exemples suivants :
-Je déteste les histoires d'amour.
- Je n'aime pas le film "Je déteste les histoires".
Le premier exemple présente un fait objectif tandis que le deuxième exemple représente l'opinion sur un film particulier.

1.5.2 Sentiment implicite

Une phrase peut avoir un sentiment implicite même sans la présence de tout sentiment porteur de mots, considérez les exemples suivants :
- Comment peuton s'asseoir à travers ce film ?
- Il faut s'interroger sur la stabilité d'esprit de l'auteur qui a écrit ce livre.
Les deux phrases ci-dessus ne portent pas explicitement de mots avec un sentiment négatif

bien que les deux soient des phrases négatives.
Ainsi, l'identification de la sémantique est plus importante en AS que la détection de syntaxe.

1.5.3 Dépendance du domaine

Il existe de nombreux mots dont la polarité change d'un domaine à l'autre. Considérez les exemples suivants :
- L'histoire était imprévisible.
- La direction de la voiture est imprévisible.
- Allez lire le livre.
Dans le premier exemple, le sentiment véhiculé est positif alors que le sentiment véhiculé dans le second est négatif. Le troisième exemple a un sentiment positif dans le domaine du livre mais un sentiment négatif dans le domaine du film (où le réalisateur est invité à aller lire le livre).

1.5.4 Identification d'entité

Un texte ou une phrase peut avoir plusieurs entités,il est extrêmement important de connaître l'entité vers laquelle l'avis est dirigé. Considérez les exemples suivants : Samsung est meilleur que Nokia Ram a battu Hari au football. Les exemples sont positifs pour Samsung et Ram respectivement mais négatifs pour Nokia et Hari.

1.5.5 Négation

La gestion de la négation est une tâche difficile en SA, elle peut être exprimée de manière subtile même sans l'utilisation explicite d'un mot négatif.
Une méthode souvent suivie pour gérer explicitement la négation dans des phrases comme « Je n'aime pas le film », consiste à inverser la polarité de tous les mots apparaissant après l'opérateur de négation (comme pas), mais cela ne fonctionne pas pour "je n'aime pas le jeu mais j'aime la mise en scène".

1.6 Conclusion

Dans ce chapitre, nous avons présenté la revue de littérature sur l'analyse des sentiments. Ceci comprend un survole théorique sur les concepts de base et leurs caractéristiques.

Chapitre 2

Généralités sur l'apprentissage automatique

2.1 Introduction

Le domaine de l'apprentissage automatique concerne la question de savoir comment construire des programmes informatiques qui s'améliorent automatiquement avec l'expérience.

Au cours des dernières années, de nombreuses applications d'apprentissage automatique ont été développées, allant des programmes d'exploration de données permettant de détecter les transactions frauduleuses par carte de crédit aux systèmes de filtrage des informations permettant d'apprendre les préférences de lecture des utilisateurs aux véhicules autonomes qui apprennent à conduire sur des autoroutes publiques.

Dans le même temps, la théorie et les algorithmes qui fondent ce domaine ont considérablement évolué.

Dans ce chapitre, nous discuterons les techniques d'apprentissage automatique les plus efficaces avec les fondements théoriques de l'apprentissage, les domaines de l'application et les procédés, avec l'explication des algorithmes et des méthodes d'évaluation les plus connus.

2.2 Définition

Tom Mitchell dans son livre « Machine Learning » a défini l'apprentissage automatique par « Le domaine de l'apprentissage automatique concerne la question de savoir comment construire des programmes informatiques qui s'améliorent automatiquement avec l'expérience.[2]

La recherche en apprentissage automatique fait partie des recherches sur l'intelligence artificielle, cherchant à fournir des connaissances aux ordinateurs par le biais de données, d'observations et en interaction avec le monde, cette connaissance acquise permet aux ordinateurs de généraliser correctement à de nouveaux paramètres. [23]

L'objectif visé est de rendre la machine ou l'ordinateur capable d'apporter des solutions à des problèmes compliqués, par le traitement d'une quantité astronomique d'informations. Cela offre ainsi une possibilité d'analyser et de mettre en évidence les corrélations qui existent entre deux ou plusieurs situations données et de prédire leurs différentes implications.

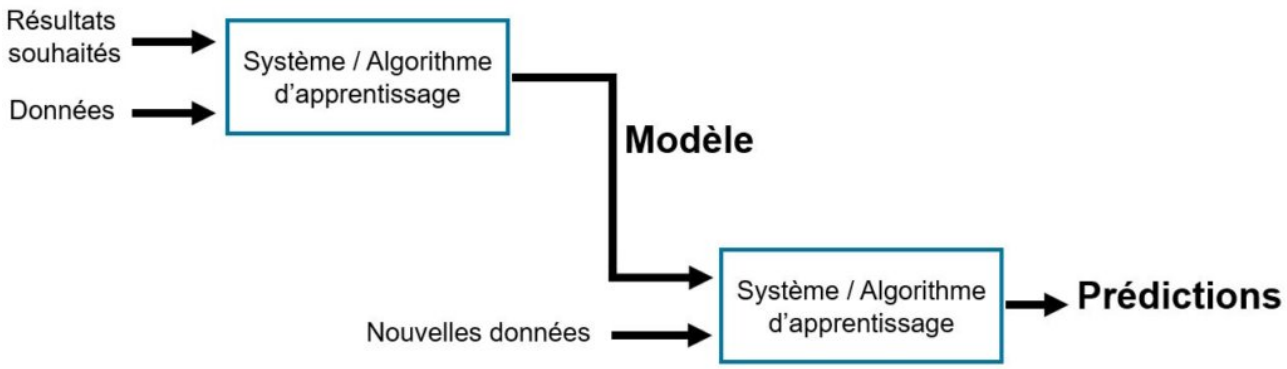


Fig. 2.1 : Processus de ML

2.3 Domaines d'application

L'apprentissage automatique est utilisé n'importe où, de l'automatisation des tâches banales à l'offre d'informations intelligentes, les industries de tous les secteurs tentent d'en tirer profit. Vous utilisez peut-être déjà un appareil qui l'utilise. Par exemple, un tracker de fitness portable comme Fitbit, ou un assistant domestique intelligent comme Google Home,mais il existe beaucoup plus d'exemples de ML en cours d'utilisation.

2.3.1 Vision industrielle

La vision assistée par ordinateur est un ensemble de technologies (matériel et logiciel) qui permet d'extraire des données sémantiquement pertinentes à partir d'images.

Ces technologies permettent aux machines, aux robots et aux applications de voir et de comprendre le monde tel que nous le voyons.

L'équivalent, en termes d'IA, des yeux humains et de la capacité de notre cerveau à traiter et analyser les images perçues,la reproduction de la vision humaine par des ordinateurs constitue d'ailleurs l'un des grands objectifs de la computer vision.

2.3.2 Reconnaissance vocale

La technique de reconnaissance vocale a fait d'importants progrès ,elle fournit des systèmes capables de transcrire en texte la langue parlée.

C'est la traduction des mots prononcés dans le texte. Il est utilisé dans les recherches vocales et plus encore. Les interfaces utilisateur vocales incluent la numérotation vocale, le routage des appels et le contrôle de l'appliance.

2.3.3 Accélération des sciences empiriques

De nombreuses sciences à forte intensité de données utilisent maintenant des méthodes d'apprentissage automatique pour faciliter le processus de découverte scientifique, il est utilisé pour apprendre des modèles d'expression des gènes dans la cellule à partir

de données à haut débit, pour découvrir des objets astronomiques inhabituels a partir de données volumineuses collectées par l'enquête Sloan et pour caractériser les schémas complexes d'activation cérébrale qui indiquent différents états cognitifs.

2.3.4 Contrôle de robot

Les méthodes d'apprentissage automatique ont été utilisées avec succès dans un certain nombre de systèmes robotiques, Par exemple, plusieurs chercheurs ont démontré l'utilisation de ML pour acquérir des stratégies de contrôle du vol en hélicoptère stable et de la voltige aérienne en hélicoptère.

2.4 Différents procédés d'apprentissage

ML impliquede nombreux systèmes d'apprentissage qui définissent ses différents modes de fonctionnement.il s'agit de :[10]

2.4.1 Apprentissage Supervisé

L'algorithme d'apprentissage reçoit un ensemble d'entrées avec les sorties correctes correspondantes, et l'algorithme apprend en comparant sa sortie réelle avec des sorties correctes pour trouver des erreurs. Il modifie ensuite le modèle en conséquence.

Grâce à des méthodes telles que la classification, la régression, la prédiction et l'augmentation du gradient, l'apprentissage supervisé utilise des modèles pour prédire les valeurs de l'étiquette sur des données supplémentaires non étiquetées. L'apprentissage supervisé est couramment utilisé dans les applications où les données historiques prédisent des événements futurs probables.

2.4.2 Apprentissage Non Supervisé

Le système doit dans l'ensemble de données cibler les données selon leurs attributs disponibles, pour les classer en groupe homogènes d'exemples.

La similarité est généralement calculée selon une fonction de distance entre paires d'exemples. C'est ensuite à l'opérateur d'associer ou déduire du sens pour chaque groupe et pour les motifs (patterns en anglais) d'apparition de groupes, ou de groupes de groupes, dans leur « espace ». Cette méthode est souvent source de sérendipité.

2.4.3 Apprentissage Semi-Supervisé

Il est utilisé pour les mêmes applications que l'apprentissage supervisé. Mais il utilise à la fois des données étiquetées et non étiquetées pour la formation - généralement une petite quantité de données étiquetées avec une grande quantité de données non étiquetées

(car les données non étiquetées sont moins coûteuses et nécessitent moins d'efforts pour les acquérir).

Ce type d'apprentissage peut être utilisé avec des méthodes telles que la classification, la régression et la prédiction. L'apprentissage semi-supervisé est utile lorsque le coût associé à l'étiquetage est trop élevé pour permettre un processus de formation entièrement étiqueté. Les premiers exemples de cela incluent l'identification du visage d'une personne sur une webcam.

2.4.4 Apprentissage par renforcement

Suppose que lors de ses pérégrinations, un agent (entité qui agit de façon autonome) reçoit des récompenses ou des punitions en fonction des actions qu'il exécute. Il s'agit alors d'établir automatiquement, à partir des retours d'expérience, des stratégies d'action des agents qui maximisent l'espérance de récompenses.

Ce type d'apprentissage est souvent utilisé dans le cadre de la robotique, de la théorie des jeux et des véhicules autonomes.

2.5 Différents algorithmes ML

Il existe plusieurs algorithmes dans les différents les type d'apprentissage automatique($Fig.2.2$), le choix de l'algorithme dépend fortement du besoin et de la tâche a résoudre. Ces algorithmes sont souvent combinés pour obtenir diverses variantes d'apprentissage.[10]

FIG. 2.2 : Carte heuristique des algorithmes de l'apprentissage automatique

2.5.1 Algorithmes d'apprentissage supervisé

Les algorithmes de l'apprentissage supervisé sont divers et différents, néanmoins ils se rejoignent dans l'exigence d'un ensemble d'exemples étiquetés afin de faire apprendre les modèles. Dans le suivant, on va présenter les plus utilisés .[10]

2.5.1.1 Naïve Bayes

La classification naïve bayésienne est un type de classification bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses. Elle met en œuvre un classifieur bayésien naïf, ou classifieur naïf de Bayes, appartenant à la famille des classifieurs linéaires. Le modèle probabiliste pour un classifieur est le modèle conditionnel

$$p(C|F_1, ..., F_n)$$

Où C est une variable de classe dépendante dont les instances ou classes sont peu nombreuses, conditionnée par plusieurs variables caractéristiques $F_1, ..., F_n$.
Lorsque le nombre de caractéristiques n est grand, ou lorsque ces caractéristiques peuvent prendre un grand nombre de valeurs, baser ce modèle sur des tableaux de probabilités devient impossible. Par conséquent, nous le dérivons pour qu'il soit plus facilement soluble.

À l'aide du théorème de Bayes, nous écrivons :

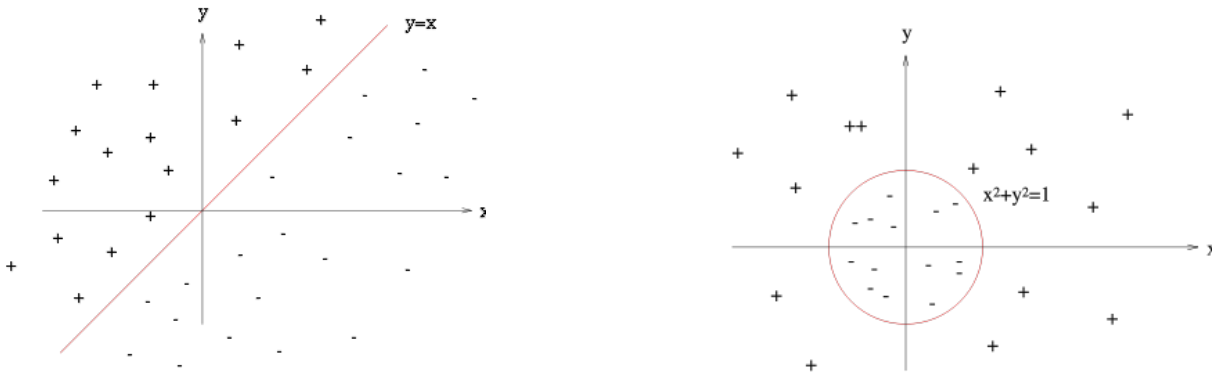
$$p(C|F_1, ..., F_n) = \frac{p(C)p(F_1, ..., F_n|C)}{p(F_1, ..., F_n)}$$

En langage courant, cela signifie :

$$postrieure = \frac{antrieure * vraisembalnce}{vidence}$$

2.5.1.2 Machine a vecteurs support SVM

L'algorithme de machine à vecteur de support est utilisé à des fins de classification et de régression. En d'autres termes, avec des données de formation étiquetées (apprentissage supervisé), l'algorithme produit un hyperplan optimal qui catégorise de nouveaux exemples. Dans l'espace de deux dimentional cet hyperplan est une ligne divisant un plan en deux parties où dans chaque classe se trouvait de chaque côté.[21]



(a) Separatrice lineaire. (b) Separatrice-non-lineaire

Fig. 2.3 : Types de séparation

2.5.1.3 Arbres de décision - Decision Tree

Les arbres de décision sont un modèle classificateur dans lequel chaque nœud de l'arbre représente un test sur l'attribut de l'ensemble de données, et ses enfants représentent les résultats. Les nœuds foliaires représentent les classes finales des points de données. Il s'agit d'un modèle de classificateur supervisé qui utilise des données portant des étiquettes connues pour former l'arbre décisionnel, puis le modèle est appliqué aux données d'essai. Pour chaque nœud de l'arbre, la meilleure condition ou décision de test doit être prise [5], exemple dans la figure 2.4.

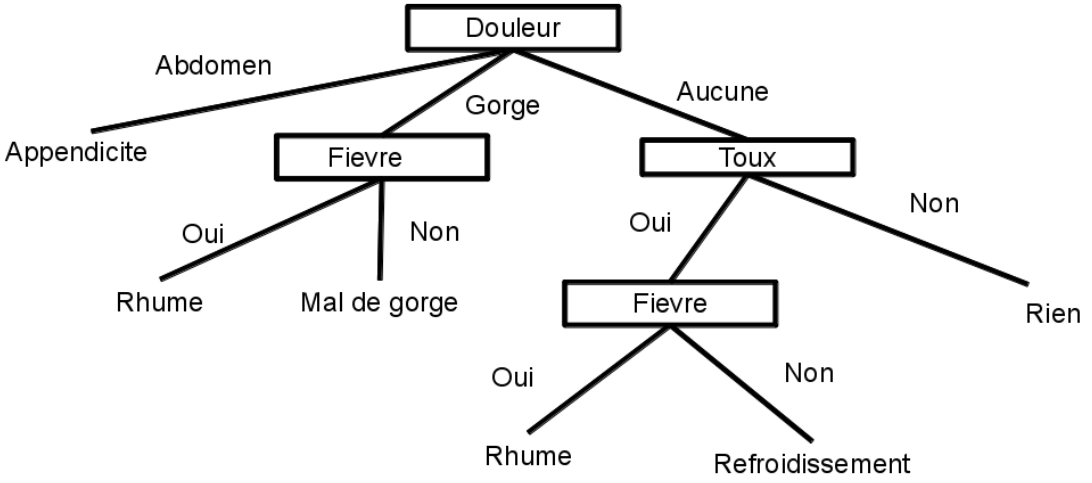


FIG. 2.4 : Arbre de décision

2.5.1.4 Les réseaux de neurones (Neural Networks)

Les réseaux de neurones sont inspirés des neurones du système nerveux humains. Ils permettent de trouver des patterns complexes dans les données. Ces réseaux de neurones apprennent une tâche spécifique en fonction des données d'entrainement.

Les réseaux de neurones se composent de nœuds (les cercles dans l'image). Dans ces réseaux, on retrouve le tiers d'entrée (Input Layer) qui va recevoir les données d'entrées. L'Input Layer va propager les données par la suite aux tiers cachés (Hidden Layers). Finalement le Tiers de sortie (le plus à droite) permet de produire le résultat de classification. Chaque tiers du réseau de neurones est un ensemble d'interconnexions des noeuds d'un tiers avec ceux des autres tiers.[32]

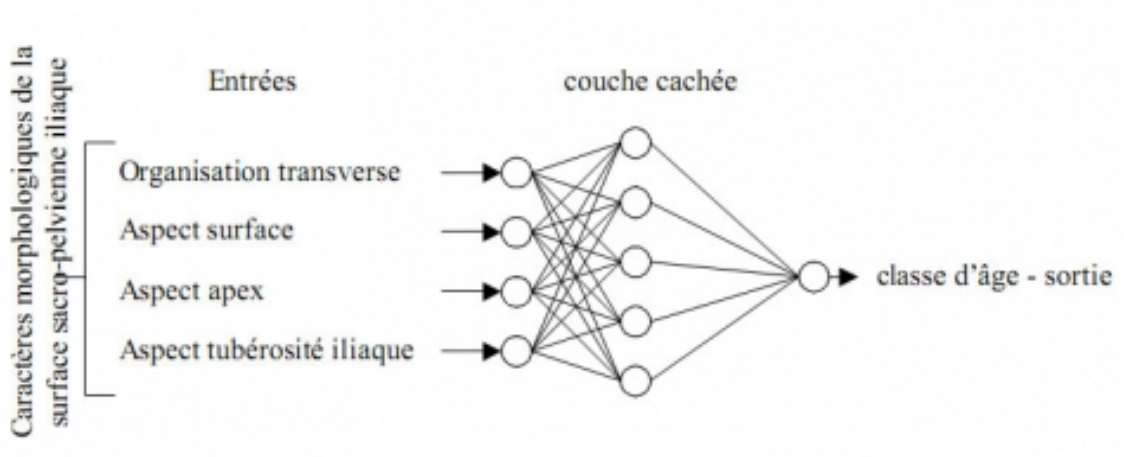


FIG. 2.5 : Schéma d'un perceptron multi-couches illustrant l'estimation de l'âge au décès à partir de l'observation de critères osseux de la surface sacro-pelvienne iliaque

2.5.2 Algorithmes d'apprentissage non-supervisé

Des modèles d'apprentissage non supervisés sont utilisés lorsque nous n'avons que les variables d'entrée (X) et aucune variable de sortie correspondante. Ils utilisent des données d'apprentissage non étiquetées pour modéliser la structure sous-jacente des données.

2.5.2.1 K-means

Le partitionnement en k-moyennes (ou k-means en anglais) est une méthode de partitionnement de données et un problème d'optimisation combinatoire.

Étant donnés des points et un entier k, le problème est de diviser les points en k groupes, souvent appelés clusters, de façon à minimiser une certaine fonction.
Nous considérons la distance d'un point à la moyenne des points de son cluster, la fonction à minimiser est la somme des carrés de ces distances. Il existe une heuristique classique pour ce problème, souvent appelée méthodes des k-moyennes, utilisées pour la plupart des applications. Le problème est aussi étudié comme un problème d'optimisation classique, avec par exemple des algorithmes d'approximation.
Les k-moyennes sont notamment utilisés en apprentissage non supervisé où l'on divise des observations en k partitions. Les nuées dynamiques sont une généralisation de ce principe, pour laquelle chaque partition est représentée par un noyau pouvant être plus complexe qu'une moyenne.[5]

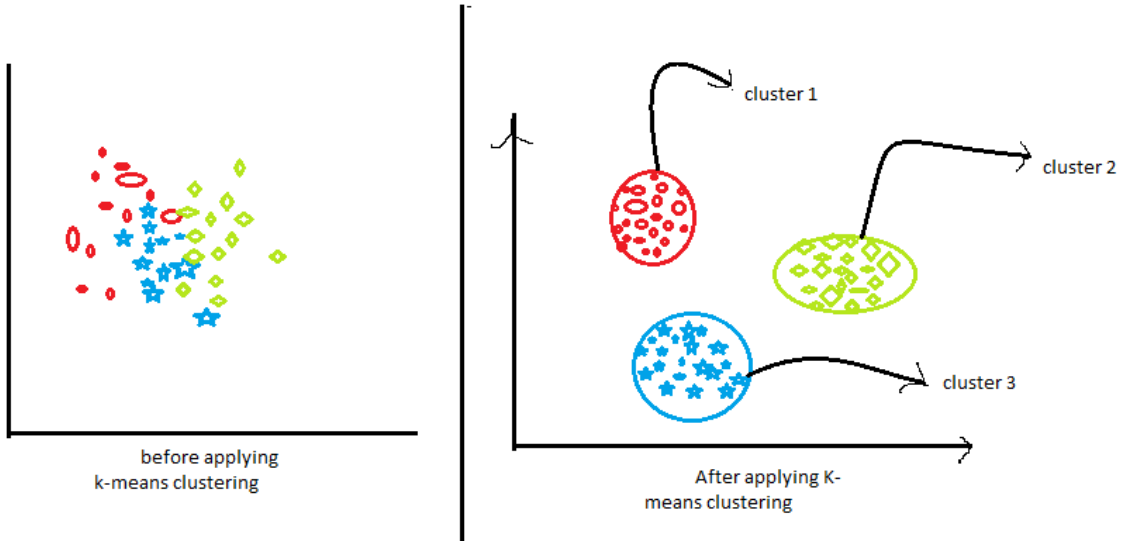


Fig. 2.6 : K-means clustering

2.6 Évaluation

2.6.1 Matrice de confusion

Est une matrice (tableau) qui peut être utilisée pour la technique de mesure du rendement pour la classification de l'apprentissage automatique. Il s'agit d'une sorte de matrice

(tableau) qui vous aide à résumer, décrire ou évaluer le rendement du modèle de classification sur un ensemble de données de test pour que les valeurs vraies soient connues.

Chaque colonne de la matrice de confusion représente les instances d'une classe réelle et chaque ligne représente les instances d'une classe prévue. , mais cela peut aussi être l'inverse, c.-à-d. colonne pour les classes prédites et ligne pour les classes réelles.

La matrice de confusion visualise l'exactitude d'un classificateur en comparant les classes réelles et prévues.[40]

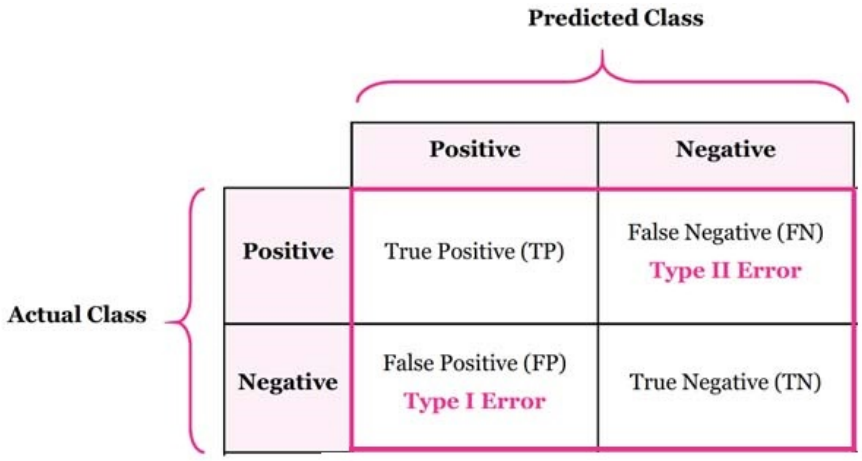


Fig. 2.7 : Matrice de confusion

- **TP** représente le nombre de messages exprimant une opinion positive et classés positifs par le classifieur.
- **TN** représente le nombre de messages exprimant une opinion négative et classés négatifs par le classifieur.
- **FP** représente le nombre de messages exprimant une opinion négative et classés positifs par le classifieur.
- **FN** représente le nombre de messages exprimant une opinion positive et classés négatifs

2.6.2 Mesures d'évaluation

Comme mentionné par Wang and Li (2019), on définit plusieurs mesures basées sur la matrice de cpar le classifieur.onfusion, afin de quantifier la performance d'un classifieur selon différents points de vue :

- Accuracy - Exactitude.
- Precision.
- Recall - Rappel.
- F1-Score / F-mesure.

2.6.2.1 Accuracy - Exactitude

Cette métrique calcule les Performances globales du modèle de classification indépendamment des classes, elle est donnée par le ration entre le nombre total de messages correctement classés par le classifieur sur le nombre total de messages.

$$Exactitud = \frac{TP + TN}{TN + FN + TP + FP}$$

2.6.2.2 Precision

La précision est le rapport des observations positives correctement prédites au total des observations positives prévues. La question que cette réponse métrique est de toutes les valeurs étiquetées comme positives, combien sont réellement ? La haute précision est liée au faible taux de faux positifs.

$$Precision = \frac{TP}{TP + FP}$$

2.6.2.3 Recall / Rappel

Le rappel est le rapport des observations positives correctement prédites à toutes les observations de la classe réelle - « OUI ». La question de rappel des réponses est : De toutes les valeurs positives qui existent vraiment, combien en avons-nous étiquetées ?

$$Precision = \frac{TP}{TP + FN}$$

2.6.2.4 F1-Score / F-mesure

Il est possible d'augmenter la valeur de la précision, mais au détriment du rappel et vice-versa, cette métrique combine la précision et le rappel pour en donner un compromis. Elle est calculée comme suit :

$$F1{-}Score = 2 * \frac{Recall{*}Precision}{Recall + Precision}$$

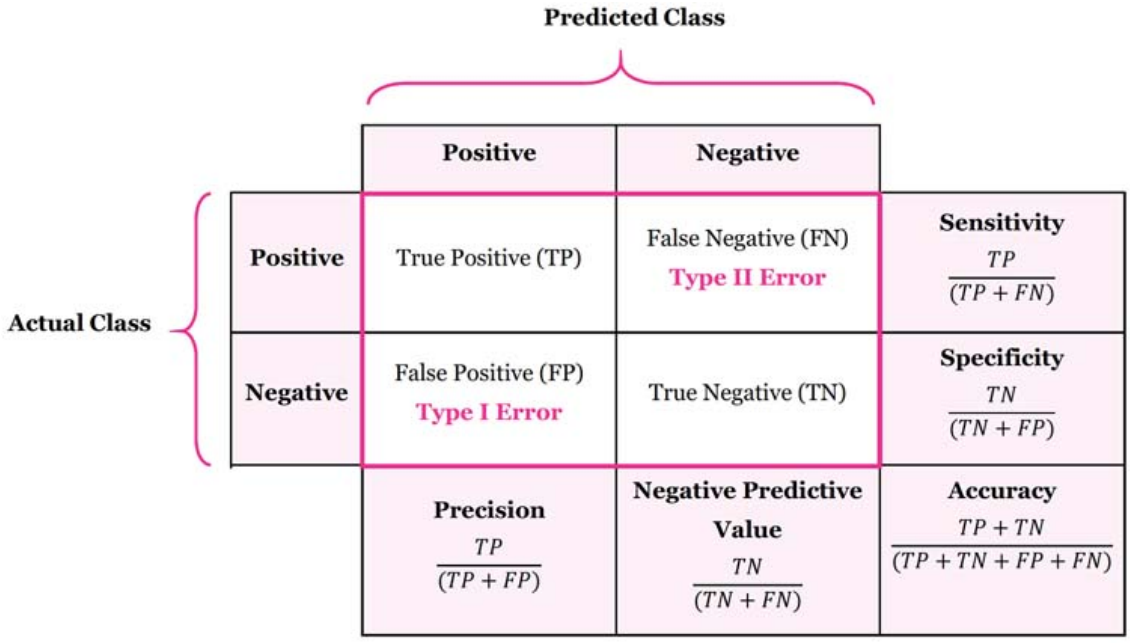


FIG. 2.8 : Mesures d'évaluation basées sur la matrice de confusion

2.6.3 Démarche d'évaluation

Le but de l'évaluation est d'estimer au mieux les performances d'un classifieur sur de nouvelles données, c'est-à-dire lorsque le classifieur sera utilisé en pratique.

L'évaluation doit donc reposer sur une démarche permettant de capturer la capacité du classifieur à se généraliser à de nouvelles données. C'est pourquoi, plutôt que d'utiliser un seul et même jeu de données pour la phase d'apprentissage et la phase d'évaluation, on préfère, via des techniques d'échantillonnage, construire et évaluer un classifieur sur des données indépendantes tirées d'une même population. On parle alors d'évaluation par validation croisée, dont les deux principales variantes sont :

- La validation croisée simple.
- La K-validation croisée.

2.6.3.1 Validation croisée simple

La validation croisée (cross-validation) est une méthode statistique d'évaluation qui consiste à diviser le corpus en deux ensembles ; un d'apprentissage, utilisée pour entrainer un modèle, et un ensemble de teste pour tester ce modèle. Dans la validation croisée, chaque segment du corpus va être utilisé dans l'apprentissage et dans le teste. La forme la plus connu est la validation croisée à k-plis (k-fold cross validation) dans laquelle l'ensemble de données ou le corpus est divisée en k segments de tailles équivalents (ou proches). À chaque rond de validation (apprentissage et teste), k-1 segments sont utilisés pour entrainer le modèle et le segment restant sert pour validation. L'opération sera répéter k fois ; à chaque itération un segment différent est utilisé pour le teste, Ensuite, le résultat de performance est calculé en une agrégation des résultats des différentes itérations, telle que la moyenne.[8]

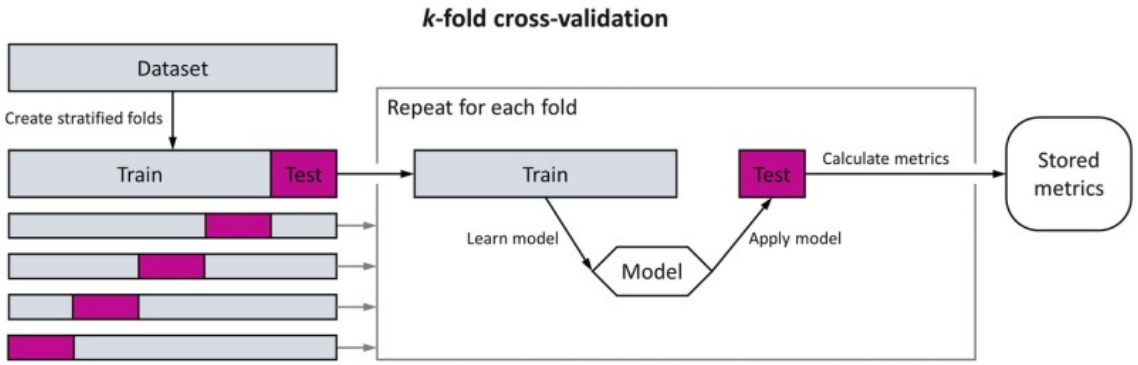


FIG. 2.9 : Validation croisée

2.7 Conclusion

Dans ce chapitre nous avons passé en revue tout ce qui tourne sur l'apprentissage automatique (concepts, domaines, procédés, algorithmes et mesures d'évaluation).

Chapitre 3

Approches et langages d'analyse de sentiment

3.1 Introduction

Les techniques d'analyse des sentiments peuvent être grossièrement divisées en trois approches : une les approches non-supervisées basées sur les lexiques de mots, les approches supervisées basées sur les méthodes d'apprentissage automatique, et les méthodes hybrides combinant des caractéristiques des deux approches .[34]

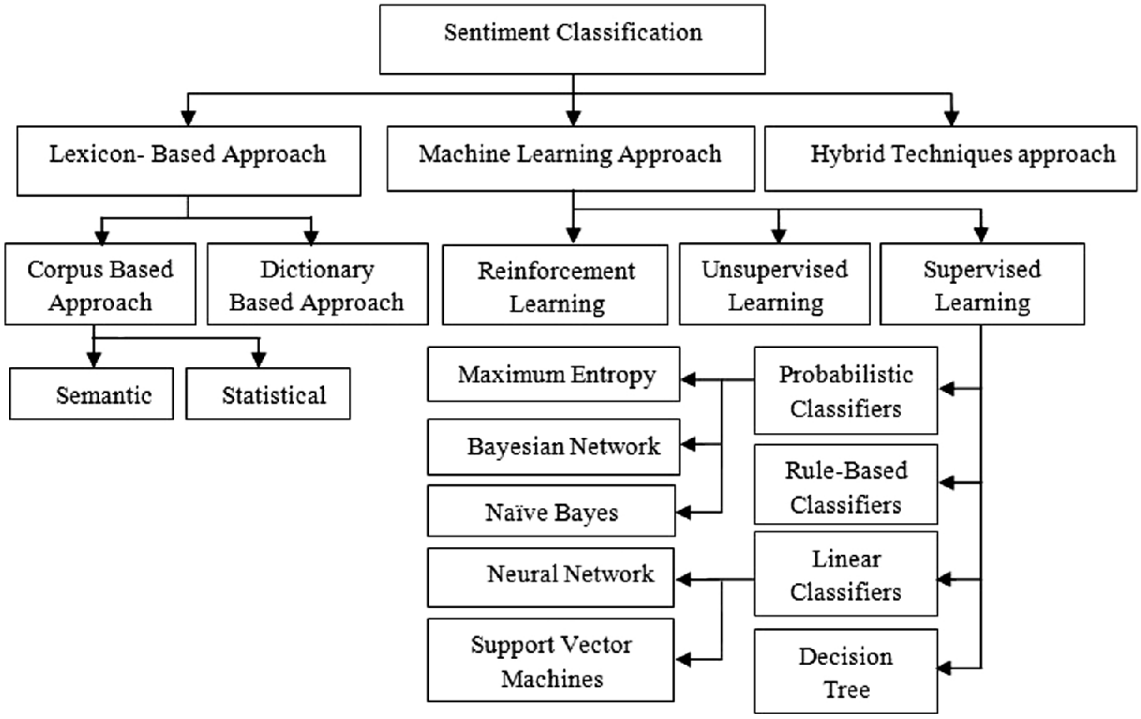


FIG. 3.1 : Approches de SA

3.1.1 Approche basée sur les lexiques

L'approche basée sur le lexique est basée sur les hypothèses que l'orientation du sentiment contextuel est la somme de l'orientation du sentiment de chaque mot ou phrase.[3]

3.1.1.1 Construction du vocabulaire

Le vocabulaire est construit à partir de 3 méthodes.[34]

- **Méthodes manuelles :** La méthode basique, initiale et triviale est la construction manuelle, cela impliquera une ressource humaine qui va travailler manuellement sur cette création, mais cela prendra un temps énorme et mobilisera une ressource humaine conséquente. Sans oublier la nécessité d'avoir des experts spécialisés dans la linguistique, afin de comprendre et donner tous les mots d'opinion dans une base de données unique sans redondance.
- **Méthodes automatiques :**cette approche utilisées deux méthodes la première méthode fondée sur le corpus et la deuxième basées sur le dictionnaire.

1. **Méthode basée corpus :** Les approches fondées sur le corpus trouvent des modèles de mots de cooccurrence pour déterminer les sentiments des mots ou des phrases
2. **Méthode basée sur un dictionnaire :** utilisent des synonymes et des antonymes dans WordNet ou bien se basent sur un calcul de distance entre les mots d'opinion « bon » ou « mauvais » et le mot que l'on souhaite classifier, pour déterminer la polarité de ce dernier.

- **Méthode hybride :** la méthode basée corpus offre un vocabulaire de domaine. la taille de dernier est relativement réduite comparée à la méthode basée dictionnaire qui donne un ensemble plus grand de mots mais qui ne sont pas nécessairement du domaine analysé .[15] la combinition des deux méthodes précédentes, nous aurons un vocabulaire riche et couvrant le domaine de l'étude.

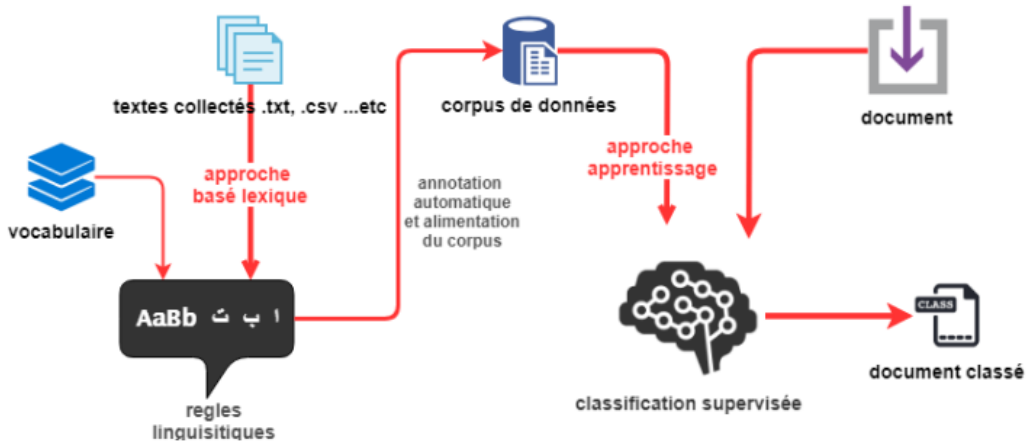


FIG. 3.2 : Approche hybride

3.1.1.2 Prétraitement

le prétraitement des données définit comme étant le processus de nettoyage et de préparation du texte pour la classification.
Initialement,Les textes du corpus sont bruts, ils ont besoin donc de plusieurs traitements avant de les utiliser dans la phase d'analyse, le processus de traitement additionnel est nommé la phase de prétraitement.
L'objectif de l'étape de prétraitement est de normaliser le texte sous une forme appropriée pour extraire les sentiments. [17]
Ci-dessous les étapes utilisées avec un exemple chacun.

1. **Morphosyntaxique POS :** Les catégories des mots comme les adjectifs, les adverbes, et aussi les noms et les verbes ($Fig3.3$) sont des bons indicateurs de sentiments dans les textes. [14]

   L'utilisation de ces types dont l'orientation sémantique est connue, permet de déterminer l'orientation sémantique globale des textes. ([7])

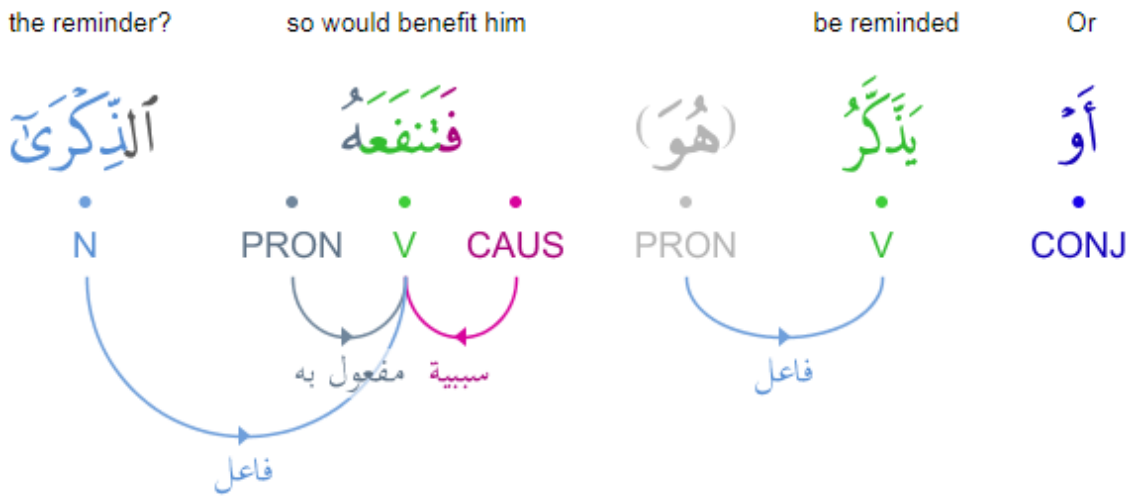


FIG. 3.3 : Morphosyntaxique POS

2. **Stemming :** Forcément (Forc)
3. **Raccourcissement exagéré des mots :** Foooort →fort
4. **Détection d'émoticônes :** Fort <3 ->Fort
5. **Détection d'hashtag et url :** #Foort visitez notre site www.site.com ->Fort visitez notre site

3.1.1.3 Calcul de la polarité

Une première méthode basée sur un lexique est proposées pour utiliser des mots d'opinion qui sont couramment utilisés pour exprimer des opinions positives ou négatives , La méthode compte fondamentalement le nombre de mots d'opinion positifs et négatifs qui sont près d'aspect (feature) de produit dans les Reviews.[6]

S'il y a plus des mots d'opinion positifs que de mots négatifs, l'opinion finale sur la fonction est positive, autrement elle est négative.

Cette méthode est simple et efficace, et donne des résultats raisonnables. Cependant, elle présente quelques lacunes majeures, elle ne dispose pas d'un mécanisme efficace pour traiter les mots d'opinion dépendants du contexte. Il y a beaucoup de ces mots comme déjà cités dans la phase de création.

Après avoir préparé la structure nous passons aux classifications ou nous utilisons essentiellement les dictionnaires ou corpus.
Le sentiment des textes est déterminé en appliquant une fonction qui tache à calculer la somme des scores des mots qui constituent le commentaire.

Tout en sachant que les scores des mots sont connus dès la phase de construction du corpus. **Exemple :** Splendide = + a, où « a » est la pondération ou le score donné à un mot pour exprimer **l'intensité du sentiment.**
Exemple dans le cas de rating : Aimer = +5, Bien = +1, contre = -1, haine = -5,

dégât = -8

Exemple Dans le cas négation et intensification

Aimer = +2, Ne pas aimer = -2, bon = +2, très bon = +4

3.1.1.4 Démarche

[34] Soit($C = W0, W1, W2 \ldots\ldots\ldots Wn$) un commentaire tel que chaque W_i a un score associé qui justifie sa polarité et son orientation (w) telle qu'elle est égale a 1 si positive, -1 si négative et 0 si neutre. Chaque phrase a un sentiment tel que

$$T(W_n) = \sum Oriontation(W_n)$$

Si T(W)>0	Si T(W)=0	Si T(W)<0
Positive	Neutre	Negative

Tab. 3.1 : Types de Polarité

l'approche basée lexique ne besoin pas d'annotation manuelle, mais ne donne pas toujours de très bons résultats surtout en termes de rappel, donc plusieurs comentaires porteurs d'opinions positive ou négative sont classés comme neutre, cela est dû à la simplicité des algorithmes de catégorisation de cette approche et aux mots qui apparaissent dans les commentaires du corpus exprimant une opinion positive ou négative et qui ne sont pas dans le vocabulaire.

Cette méthode a réussi à classifier les commentaires, mais au-dela d'un seuil de masse de commentaire avec plus grande complexité ça devient difficile, alors il est nécessaire de voire des méthodes plus avancées afin d'obtenir une précision assez élevée.

3.1.2 Approche basé sur l'apprentissage automatique ML

La classification des commentaires selon la nature de sentiments ou d'opinions qu'ils expriment est résolue en utilisant l'approche tendance qui est l'apprentissage automatique.

Ainsi, l'analyse de sentiments adopte une approche qui découle directement de la fouille de Texte.

Cette approche est très utilisée en raison de ses résultats jugés satisfaisants et surtout de ses avantages à l'image de la spécialisation au domaine de l'analyse de sentiments.

Donc afin de déterminer si un texte ou un commentaire appartient à une classe donnée exemple (positifs, négatifs) des algorithmes comme le Naïve Bayes ou encore le SVM, à savoir l'apprentissage automatique supervisé, peuvent être utilisés.

Ces algorithmes sont entrainés sur un ensemble d'exemples de documents dont la classe est connue au préalable dit corpus d'entrainement. Ou bien des algorithmes de l'apprentissage non supervisé, a savoir K-means.

De ce qui suit nous allons présenter les étapes et les méthodes utilisées dans cette approche ainsi que quelques travaux basés sur l'apprentissage automatique.

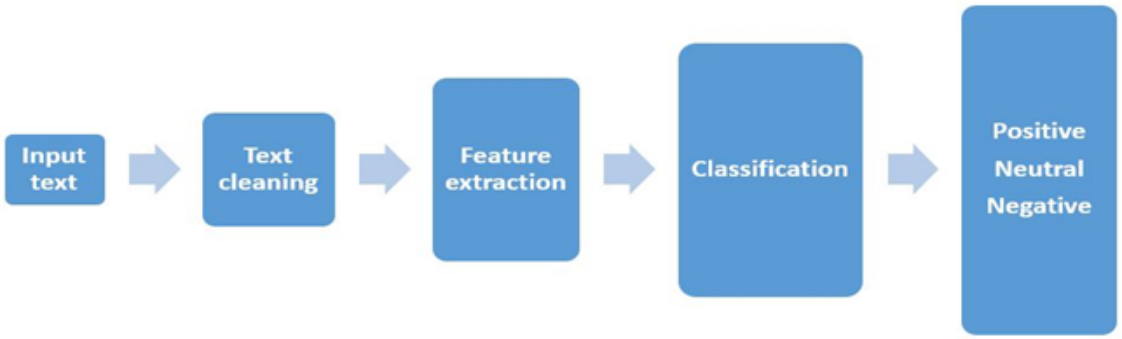


FIG. 3.4 : Processus de l'approche apprentissage automatique

L'ensemble du processus comporte plusieurs étapes, nettoyage de texte en ligne, suppression des espaces, extension des abréviations, création de raccourcis, suppression des mots vides, traitement de la négation et enfin sélection des fonctionnalités.

Les caractéristiques dans le contexte de l'extraction d'opinions sont les mots, les termes ou les expressions qui expriment fortement l'opinion sous forme positive ou négative. Cela signifie qu'ils ont un impact plus important sur l'orientation du texte que d'autres mots du même texte,on a détailler le processus dans ce qui suit.

3.1.2.1 Prétraitement

Cette phase est commune dans les deux approches, néanmoins elle diffère dans ses détails La phase de prétraitement est très importante pour modifier le format des phrases en supprimant des mots des vides ou des caractères spéciaux ou bien en modifiant carrément le format des mots.

3.1.2.2 Annotation

En donnant une étiquette positive, négative, neutre à chaque commentaire, le classifieur ou bien l'algorithme d'apprentissage dans le cas d'apprentissage automatique va pouvoir construire un modèle de classification robuste.

Dans l'approche apprentissage automatique, nous avons un corpus étiqueté qui sera éclaté en corpus d'entrainement et un corpus de test. Le corpus de test est aussi annoté pour vérifier par la suite la précision de l'algorithme et s'il a vraiment prédit les bons éléments dans les bonnes classes. [13]

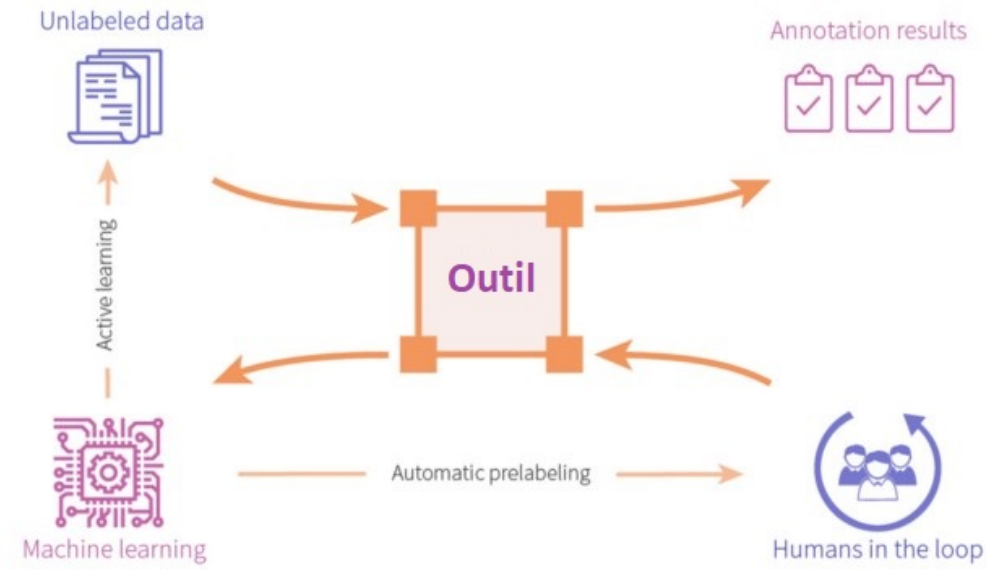


FIG. 3.5 : Annotation

- **Annotation manuelle :** Cette annotation est faite par un petit ensemble d'annotateurs humains (des arabophones dans le cas de la langue arabe) avec des compétences scientifiques vérifiant un certain niveau d'anakyse.[26]
- **Annotation automatique :** L'annotation automatique se fait avec une approche basée lexique, cette méthode vient remédier au coût de temps de l'annotation manuelle mais ne donne pas toujours une très bonne annotation et le taux d'erreur est souvent élevé, ce qui ne permet pas de construire un bon modèle de classification.[13]

3.1.2.3 Rreprésentation du texte

cette phase permet d'extraire l'information du texte pour le traitement ultérieur par des modèles de machine Learning. En d'autres termes.

1. **sac de mots (BOW :bag of words) :**Il est appelé un « Sac de mots, parce que toute information sur l'ordre ou la structure des mots dans le document est ignorée. Le modèle est uniquement concerné par la question de savoir si les mots connus se produisent dans le document, et non ou exactement dans le document.[29]

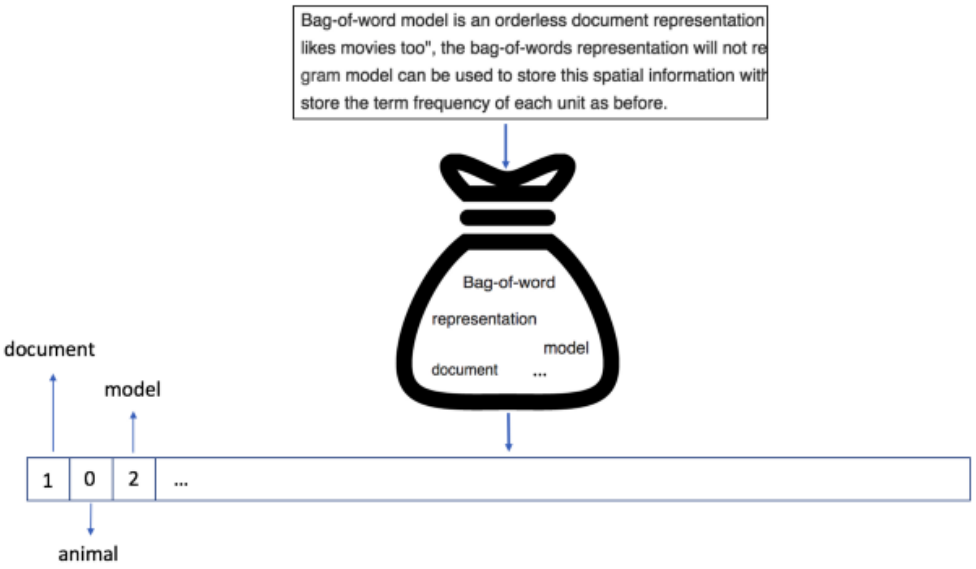


FIG. 3.6 : Bag of words

2. **Alimentation du Bag of words (BOW)** : Inspiré des techniques de Recherche d'Information, le poids peut être estimé de plusieurs façons, on cite dans ce qui suit les plus utilisées.

   - **Fréquence** :Dans cette représentation, le poids w_ij (importance) d'un terme t_i est estimé selon sa fréquence d'apparition le document d_i ainsi le montre la formule suivante :

     $$W_ij = f(t_j; d_i) = TF(t_j; d_i)$$

     Où $TF(t_j; d_i)$ est le nombre d'apparitions de t_j dans le document d_i. Cette représentation, contrairement au modèle booléen, accorde de l'importance à la fréquence d'apparition du mot dans le document en partant du principe que plus un mot est fréquent, plus il a de l'importance.

   - **Term Frequency/Inverse Documents Frequency TF-IDF** Inverse document Frequency (idf) : utilisé pour calculer le poids des mots rares dans tous les documents du corpus. Les mots qui apparaissent rarement dans le corpus ont un score IDF élevé. Il est donné par l'équation ci-dessous.

     $$W_ij = f(t_j; d_i) = TF(t_j; d_i) * IDF(t_j)$$

     $$IDF(t_j) = \log \frac{n}{n(T)}$$

     Où : $TF(t_j; d_i)$ C'est la même fonction utilisée dans le calcul de fréquence. IDF (tj) le nombre total de documents « N » diviser par le nombre de documents qui contiennent le terme t_j « df_j ».

   - **Word Embedding (WE)** Word embedding est l'une des représentations les plus populaires du vocabulaire du document,il est capable de capturer le contexte d'un mot dans un document, similitude sémantique et syntaxique,

relation avec d'autres mots aussi comme le nom collectif d'un ensemble de techniques de modélisation et d'apprentissage de la langue dans le traitement du langage naturel NLP où les mots ou les phrases du vocabulaire sont mappés à des vecteurs de nombres réels dans un espace de faible dimension par rapport à la taille du vocabulaire.
ainsi, le word embeddings est capable de trouver que :King—man + woman = Queen, ou Paris—France + England = London.
Sans avoir recours à une ressource sémantique ou à une intervention humaine.

Word2vec est l'une des techniques populaires dans WE elle possède deux architectures neuronales, appelées CBOW et Skip-Gram, parmis lesquelles l'utilisateur peut choisir. CBOW reçoit en entrée le contexte d'un mot, c'est à dire les termes qui l'entourent dans une phrase, et essaye de prédire le mot en question. Skip-Gram fait exactement le contraire : elle prend en entrée un mot et essaye de prédire son contexte. Dans les deux cas, l'entraînement du réseau se fait en parcourant le texte fourni et en modifiant les poids neuronaux afin de réduire l'erreur de prédiction de l'algorithme. [19]

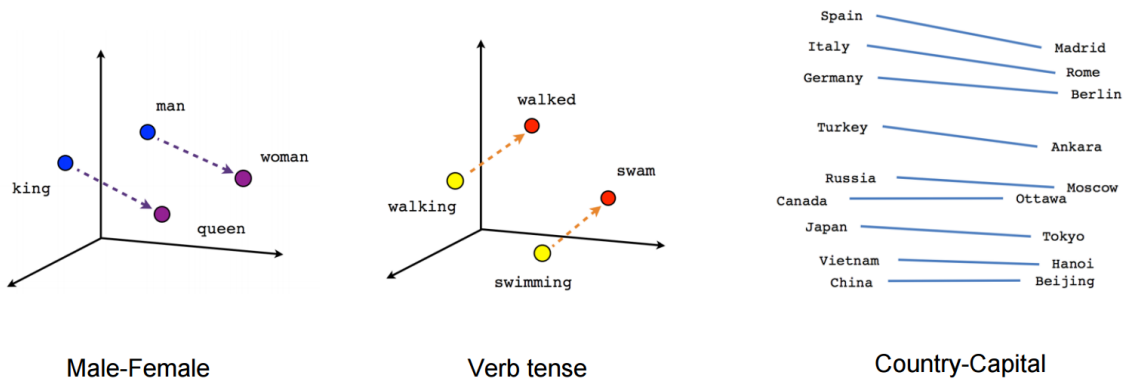


Fig. 3.7 : Technique Word2vec

3.1.3 Approche hybride

Une première façon de combiner les méthodes est d'utiliser les dictionnaires afin de préparer le corpus avant de classer les textes a l'aide d'outils d'apprentissage automatique. Une autre façon de faire est d'utiliser les techniques d'apprentissage automatique dans le but de construire les dictionnaires d'opinion nécessaires a l'approche basée lexique.

[1] présentent une méthode ayant pour objectif de définir l'orientation sémantique des adjectifs pour la construction du dictionnaire d'opinion.
Ils extraient tout d'abord tous les adjectifs du corpus a l'aide d'un analyseur syntaxique, puis utilisent un algorithme de clustering afin de classer les adjectifs selon leur polarité. Une dernière façon d'utiliser conjointement les deux approches est de construire plusieurs types de classifieurs et de combiner leurs résultats, soit par des systèmes de vote, soit par un algorithme d'apprentissage. [37]

3.2 Langue arabe

[11] L'arabe est l'une des six langues ofcielles des Nations Unies et la langue maternelle d'environ 300 millions de personnes dans 22 pays diérents,4 dont l'arabe standard. et dialectes. L'arabe a trois variétés principales illustrées à la figure .

L'arabe classique également appelé arabe coranique, est utilisé dans les textes religieux et dans de nombreux manuscrits arabes anciens.

MSA(Modern Standard Arabic) est la langue formelle de communication comprise par la majorité des arabophones, comme il est couramment utilisé à la radio, dans les journaux et à la télévision. L'arabe dialectal ou parlé est utilisé dans la conversation quotidienne et récemment utilisé à la fois à la télévision et à la radio.

3.3 Dialecte Algérien

Dialecte algérien (DALG) est un dialecte maghrébin, principalement utilisé dans la communication informelle, y compris les médias sociaux.[16]
DALG n'est pas utilisé dans l'enseignement scolaire ou dans les nouvelles télévisées.

Il est plus utilisé dans la vie quotidienne, la musique, les séries diffusées, etc. en Algérie. Ce dialecte est considéré comme une langue de faible variété, ce qui signifie que DALG est faible normalisé . Cependant, DALG est parlé seulement par 70-80% de la population algérienne (dont le nombre total est de 40 autres 20-30% de la population utilisent berbère.

DALG s'est aussi enrichi de l'influence de la langue des pays qui ont colonisé la population algérienne. Parmi ces langues sont le turc, l'italien et plus récemment le français. Par conséquent, DALG est considéré comme un mélange entre différentes langues où MSA est en première position.

3.4 Problématiques du Traitement Automatique DALG

Est un sous problème du traitement automatique du langage naturel TALN. En tant que comportement humain, les langues naturelles sont très difficiles à traiter automatiquement, vue l'énorme quantité de connaissance humaine véhiculée implicitement dans un texte.

Ce problème de complexité sera plus délicat avec les langues à morphologie complexe dont l'Arabe. Des problèmes apparaissent lors des tentatives d'application des techniques de TALN sur la langue arabe, et des travaux commencent à prendre en charge ces problèmes par différentes motivations et objectifs.

3.4.1 Translitération

La translitération est un processus de passage d'un texte écrit en un script ou alphabet donné vers un autre La translitération de l'arabizi vers l'arabe fait cependant face à un ensemble de problématiques [41] :

1. **Traitement des voyelles** : les voyelles (a, i, o, u, e, y) peuvent être remplacées par les différentes lettres arabes (أ,ا,ي,و,ة) ou encore par aucune lettre. Cela dépend de leurs emplacements dans le mot.

2. **Ambiguïté entre plusieurs lettres** : une lettre arabizi peut correspondre à plusieurs lettres arabes. Par exemple, la lettre 't' peut correspondre aux deux lettres arabes (ط et ت par T).

3. **Ambiguïté reliée au contexte :** dans certains cas, plusieurs translitérations peuvent correspondre au même mot. Par exemple le mot (matar 'مطر') pourrait être translitéré en la pluie , ou encore(' مطار' matar) aéroport.[20]

4. **Ambiguïté reliée au code switching :** certains mots d'autre langues tels que le français ou l'anglais peuvent être pris pour des mots en arabizi. Par exemple le mot 'men' en anglais signifie homme en Dalg (من).

3.4.2 Synthèse des travaux qui traitent DALG

Certains travaux ayant traité le problème d'AS dans le dialecte Algérien, on a désigné ce qui suit :

Dans cet article [30] a présenté un outil pour l'analyse des sentiments des messages écrits en dialecte Algérien. Cet outil est basé sur une approche qui utilise à la fois des lexiques et un traitement spécifique de l'agglutination. Cette approche a été expérimentée en utilisant deux lexiques de sentiments et un corpus de test contenant 749 messages. Les résultats obtenus sont encourageants et montrent une amélioration continue après chaque étape de l'approche considérée.

Dans [25], ils ont proposé une nouvelle approche d'analyse des sentiments basée sur le lexique pour aborder les aspects spécifiques de l'arabe algérien vernaculaire pleinement utilisé dans les réseaux sociaux. Un ensemble de données annoté manuellement et trois lexiques arabes algériens ont été créés pour explorer les différentes phases de notre approche.

[34] où ils ont proposé de faire une approche supervisée en utilisant des classificateurs tels que SVM et NB, et ils ont ajusté la phase de prétraitement pour minimiser le corpus suggérant des approches telles que la suppression des voyelles et la translittération, et ils ont réussi à améliorer les résultats de la classification .

[33] ont proposé un système qui opère en trois phases, la première consiste à la construction et le prétraitement manuel du corpus recueillis à partir des journaux arabes algériens. La seconde phase est le choix des caractéristiques pour la représentation des commentaires. Enfin la troisième phase est la réalisation du module de classification combinant quatre classificateurs SVM avec des fonctions noyaux différents. Ils ont utilisé deux stratégies nommé un contre un et un contre tous dont les résultats ont prouvé que la première stratégie est meilleure que la deuxième avec les commentaires des journaux en langue arabe.

Le travail	L'approche	Language	Les résultats
]30[	ML	DALG	F1 score=78%
]31[	ML	Tunis	accurcy=74%
]35[	Hybride	MSA/EGY	accurcy=85%
]34[	ML	DALG	F1 score=87%
]25[	Lexique	DALG	accurcy=79%
]33[	ML	DALG	accurcy=75%
]28[	ML	Maroc	accurcy=78%

Tab. 3.2 : Synthèse des travaux qui traitent les dialectes arabe

3.5 Conclusion

Ce chapitre présente un survol sur les principales approches et techniques orientées classification des sentiments en ML. Enfin, Il se termine par une synthèse des travaux qui traitent les dialectes arabes.

La partie suivante sera dédié à la conception et à la mise en oeuvre de notre Framework d'analyse de sentiments traitant le dialecte Algérien.

Deuxième partie

Volet pratique : Etat des lieux , Conception et Réalisation, tests et résultats

Chapitre 4

Etat des lieux

Dans le cadre de notre projet qui consiste à faire une analyse des sentiments en dialecte Algérien et qui sera appliqué sur les commentaires publiés sur le sujet Hirak, sur youtube et form-google. Les internautes arrivaient à exprimer et à révéler leurs opinons sur la politique actuelle de l'Algérie.

Les dernières statistique de l'utilisation des réseaux sociaux par le peuple algérien en janvier 2020 sont fournis à travers la figure 4.1 :

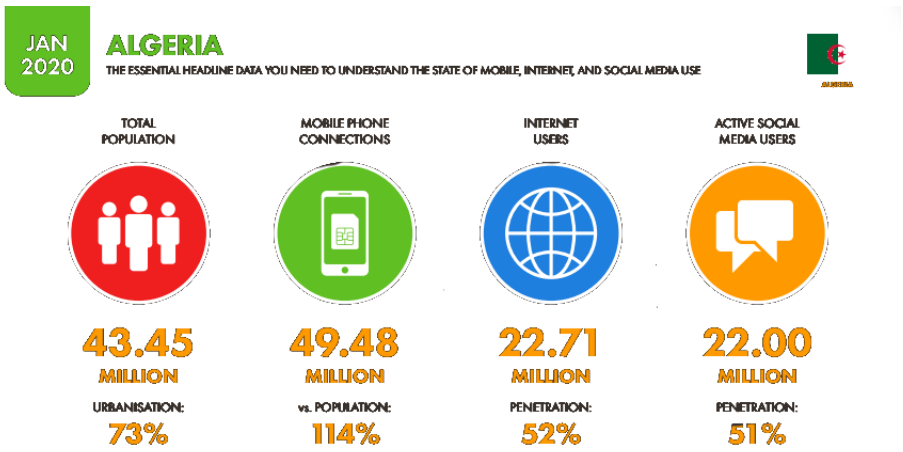


FIG. 4.1 : Statistiques de l'utilisation des réseaux sociaux par les Algériens

4.1 Algérien et les réseaux sociaux

Dans les figures 4.1 et 4.2 [1] ,tous ces chiffres et statistiques donnent une vue d'ensemble de la position et la diffusion dont bénéficient les sites des réseaux sociaux en Algérie, surtout sur Youtube . Ce dernier est devenu l'un des espaces virtuels concernant le militantisme contre la gouvernance actuelle et la lutte de la corruption à travers des publications qui font circulaient leurs opinions, perceptions et préoccupations sur ce qui se passe en Algérie.

[1] https ://datareportal.com/reports/digital-2020-algeria ?rq=Algeria%20

JAN 2020

MOST-VISITED WEBSITES (ALEXA)

RANKING OF TOP WEBSITES BY AVERAGE MONTHLY TRAFFIC ACCORDING TO ALEXA

#	WEBSITE	TIME / VISIT	PAGES / VISIT
01	GOOGLE.COM	12M 09S	14.6
02	YOUTUBE.COM	11M 44S	6.7
03	OUEDKNISS.COM	13M 45S	9.7
04	GOOGLE.DZ	5M 48S	8.8
05	FACEBOOK.COM	17M 48S	7.8
06	YAHOO.COM	4M 34S	4.4
07	BLOGSPOT.COM	3M 32S	2.7
08	ELBILAD.NET	4M 02S	2.1
09	WIKIPEDIA.ORG	3M 55S	3.0
10	GOOGLE.FR	5M 16S	4.6

#	WEBSITE	TIME / VISIT	PAGES / VISIT
11	ENNAHARONLINE.COM	7M 34S	3.7
12	ECHOROUKONLINE.COM	3M 53S	2.4
13	ELKHABAR.COM	4M 20S	2.5
14	BONGACAMS.COM	3M 04S	1.7
15	ALIEXPRESS.COM	12M 17S	9.0
16	LIVE.COM	4M 53S	5.0
17	LIVEJASMIN.COM	2M 20S	1.5
18	ELKHADRA.COM	1M 39S	1.4
19	AMAZON.COM	9M 41S	9.0
20	MEDIAFIRE.COM	2M 33S	2.3

Fig. 4.2 : Statistiques de l'utilisation des média sociaux par les Algériens

4.2 Définition de terme mouvement populaire

Mouvement populaire est liée à l'idée de mouvement sociale. Qui est essentiellement un groupe de personnes partageant certaines activités et ils utilisent un discours qui vise à changer une société ,défiant le pouvoir politique existant comme associé .

La création du mouvement Hirak et sa capacité d'influencer et de provoquer le changement est liée au pouvoir social. La "HIRAK " dénote un mouvement total et global de divers catégories et segments de la société, qui recherchent un changement qualitatif dans la nature gouvernance politique et ordre social.

4.3 Février 22

Le 22 février 2019, l'une des manifestations populaires les plus importantes et les plus impressionnantes au monde, des millions d'Algériens sont descendus dans la rue après la prière du vendredi pour instauré la démocratie et le changement.
Des manifestations dont l'objectif principal était d'exiger la dimission du président Abdelaziz Bouteflika, au vu de sa candidature Anticonstitutionnelle pour un cinquième mandat présidentiel. C'est Les manifestations dont ont été témoins la plupart des villes algéiennes, auxquelles ont participé différents groupes d'âge et classes sociales, ont tenté debrosser l'image de la nouvelle Algérie qui cherche à soutenir les manifestants en adoptant la paix et en n'utilisant pas la violence comme tendance et comme manifestation de celle-ci, l'ont qualifié de "mouvement populaire".

Le mouvement populaire en Algérie a représenté une remise en cause de l'état de peur et du mur du silence historique appliqué à l'Algérie, et il est lié en particulier aux événements de la « décennie noire ».Ce qui est devenu une référence pour établir la légitimité du régime au pouvoir de l'ancien président Abdelaziz Bouteflika.

4.3.1 Chronologie du mouvement du 22 février

le figure ci-dessus représenter le chronologie des plus importantes Les événements du mouvement 22 février.

Fig. 4.3 : Chronologie du mouvement du 22 février(départ de Hirak)

4.4 Classification des opinions de Hirak

Après les élections présidentielles du 12 décembre, le mouvement a montré une grande division entre un partisan qui considère qu'il s'agit de la seule solution pour se débarrasser de l'ancien régime, dans le cadre de la solution constitutionnelle et de la préservation sur la sécurité et la stabilité de l'Algérie, et chez les opposants qui y voient une opportunité de renouveler le régime lui-même avec l'émergence d'un troisième groupe qui ne soutient ni les élections ni le mouvement et considère qu'il a dévié de son cours et de ses principes d'origine. La table 4.1 détaillé les point de vie de chaque catégorie .

Catégorie 1	Catégorie 2	Catégorie 3
-Vous insistez pour protester et continuer le Hirak -Refuse de reconnaître la légitimité du président élu Tebboune	-Ils pensent qu'il est possible d'interagir avec le nouveau président dans le cadre de contrôles spécifiques -Avec la continuité du mouvement comme garantie centrale pour faire pression sur le président et faire vivre les revendications You can simply impress your audience and add a unique zing.	-Ne soutient ni les élections ni le mouvement et considère qu'il a dévié de son cours et de ses principes d'origine

Tab. 4.1 : Catégories des opinions

4.4.1 linguistique de Hirak

Pendant le mouvement algérien, le dictionnaire linguistique s'est doté d'un certain nombre d'"insultes politiques" et d'un vocabulaire qui a coïncidé avec leur apparition depuis sa création. L'utilisation de ces termes dans les médias sociaux et médiatiques s'est accrue . Parmi les termes qui sont apparus et ont accompagné le mouvement et sont encore en circulation jusqu'à présent, nous prenons plusieurs exemples [39] comme le mote **"alrakmaja"** en frençais "surfer sur la vague" ce que signifie Il a été utilisé pour décrire ceux qui ont rejoint le mouvement populaire, après que leur position était de soutien à l'ancien président, le mot "**kachirist**" qui tire son nom d'un type de nourriture "kachire"signifie la trahison et de division dans les rangs, **"alshyat"** dérivé du mot "shitta" en Dalg, (brosse de nettoyage de chaussures)Un signe soumission et de gain ,**"bousbaa lazrag"**(doigt bleu)un terme ironique pour tous ceux qui ont voté après le Hirak et beaucoup des mots comme (Zwaf,Menjal,Makist,Forchita, Ronjas,dhobab) plupart d'entre eux sont classés comme des insultes(négative).

Ce que nous avons évoqué plus haut montre que dans le domaine que nous avons choisi, il existe une diversité de mots qui peuvent exprimer une intention et des sentiments s'ils sont présents dans les commentaires .

4.5 Conclusion

Ce chapitre présente un survol sur les dernières statistiques de l'utilisation des Algériens des réseaux sociaux, la chronologie du mouvement du 22 février, la classification des opinions de Hirak et la diversité du dictionnaire linguistique.

Le chapitre 4 suivant est consacré à la conception complète de notre outil d'AS.

Chapitre 5

Conception de notre outil d'AS

5.1 Introduction

L'objectif de notre projet de fin d'études est la conception et la mise en oeuvre d'un outil d'analyse de sentiments des commentaires publiés en Algérien concernant le domaine politique. Cependant, L'extraction des données d'un big-social data comme Youtube n'est pas une tâche très simple. Nous devons en tenir compte des obstacles rencontrés.

Nous avons choisi de construire un dataset tournant sur le Hirak pour les raisons évoquées antérieurement en introduction et chapitre 3. Ce mouvement de militantisme est l'un des espaces où toutes les couches de la société participent et interagissent via les medias réels et sociaux, d'ici de nombreux enjeux de société en dérivent et de nombreux nouveaux termes s'ajoutent au dictionnaire du dialecte Algérien et tout ça pour améliorer la gestion des réseaux sociaux et mesurer l'opinion du peuple Algérien et des internautes pour une bonne prise de décision sur le futur de l'Algérie.

Notre approche d'analyse du sentiment sur le militantisme se base sur une ressource lexicale crée par extraction et collecte de données de Youtube et enrichie manuellement à partir des formulaires Google. Cette ressource linguistique aide à donner de l'information sur la polarité des mots afin de calculer le sentiment de chaque commentaire. Les commentaires reflètent les statuts postés sur le réseau social Youtube.

Ce chapitre étudie la manière dont les données seront traitées, analysées et principalement comment les classifier ? Avant de passer à cette chose, parlons des hypothèses de départ.

5.2 Hypothèse de départ

Un ensemble d'hypothèses sur la détection de la polarité des commentaires en dialecte algérien est détaillé comme suit :

- **La nature** des commentaires à traiter.
- **Le type d'opinion** utile pour le tratement traiter.
- **Le niveau** d'analyse de sentiments.

5.2.1 Type d'opinion

la complexité d'analyse des opinions comparatives restent très dépendantes de la structure des phrases et de la langue utilisée ce qui constitue un obstacle pour le DALG ,nous avons choisi le type d'opinion régulières(ordinaire).

5.2.2 Niveau d'analyse

Nous avons précédemment expliqué précisément les différents niveaux d'analys (document, phrase et le niveau aspect).

Le choix du niveau dépend fortement de la structure et du type de la donnée a traiter, donc on a décider d'effectuer notre analyse de sentiments au niveau document pour des raisons suivantes :

- Très souvent un commentaire ne traite qu'un seul sujet
- Très souvent des commentaires ne comportent qu'une seule opinion.

5.2.3 Langue des documents

Des commentaires sont écrits en arabe et en français , et enfin en dialecte Algérien (arabic et arabizi) .

5.2.4 Approche d'analyse de sentiment

Plusieurs facteurs interviennent pour le choix de l'approche d'AS en Algérien, nous en citons :

- L'aptitude de construire des vocabulaires.
- La complexité du langage traité.
- Le comportement des Internautes

comme mentionné précédemment, cependant la nature du dialecte Algérien ne nous facilite pas AS, car c'est une langue non structurée, qui ne se base pas sur des règles linguistiques ni sur des règles orthographiques.

Prenons à titre d'exemple une phrase du dialecte algérien (maranich lahi) et (rani lahi)les deux phrase signifient en français(je suis occupé) malgré que la 1er phrase c'est la négation de 2em phrase et aussi pour l' écriture, elles s'écrivent différemment ,le Tableau ci-dessus est une illustration de ce cas :.

Mot	Équivalent
maranich	mraniche -mrnich-maranish
ngolo	ngoulo-nglo-ngolah

Tab. 5.1 : Différentes écriture en DALG

La diversité des accents algériens,rendont l'approche lexique est diffécile (la création des vocabulaires et des dictionnaires de mots de polarité) et écarte l'approche hybride,Notre approche de SA se base sur l'apprentissage automatique .

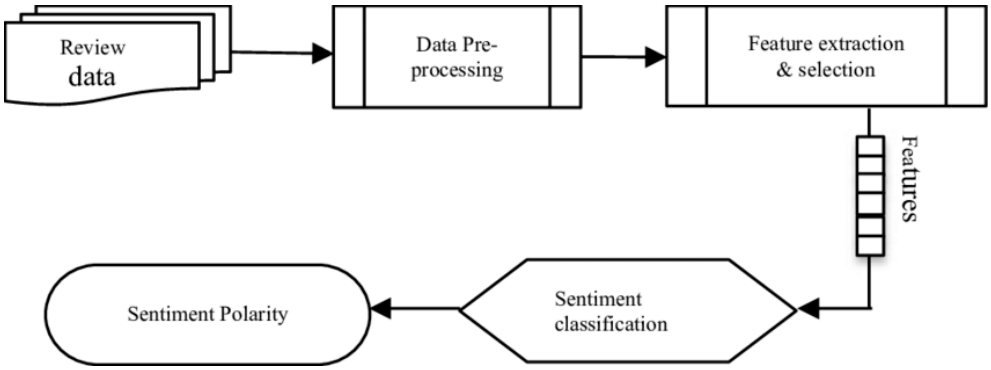


Fig. 5.1 : Apporoche apprentissage automatique

5.3 Architecture générale de la solution :

Cette section est consacrée au processus général de la méthodologie de SA des commentaires du big-social data Youtube :

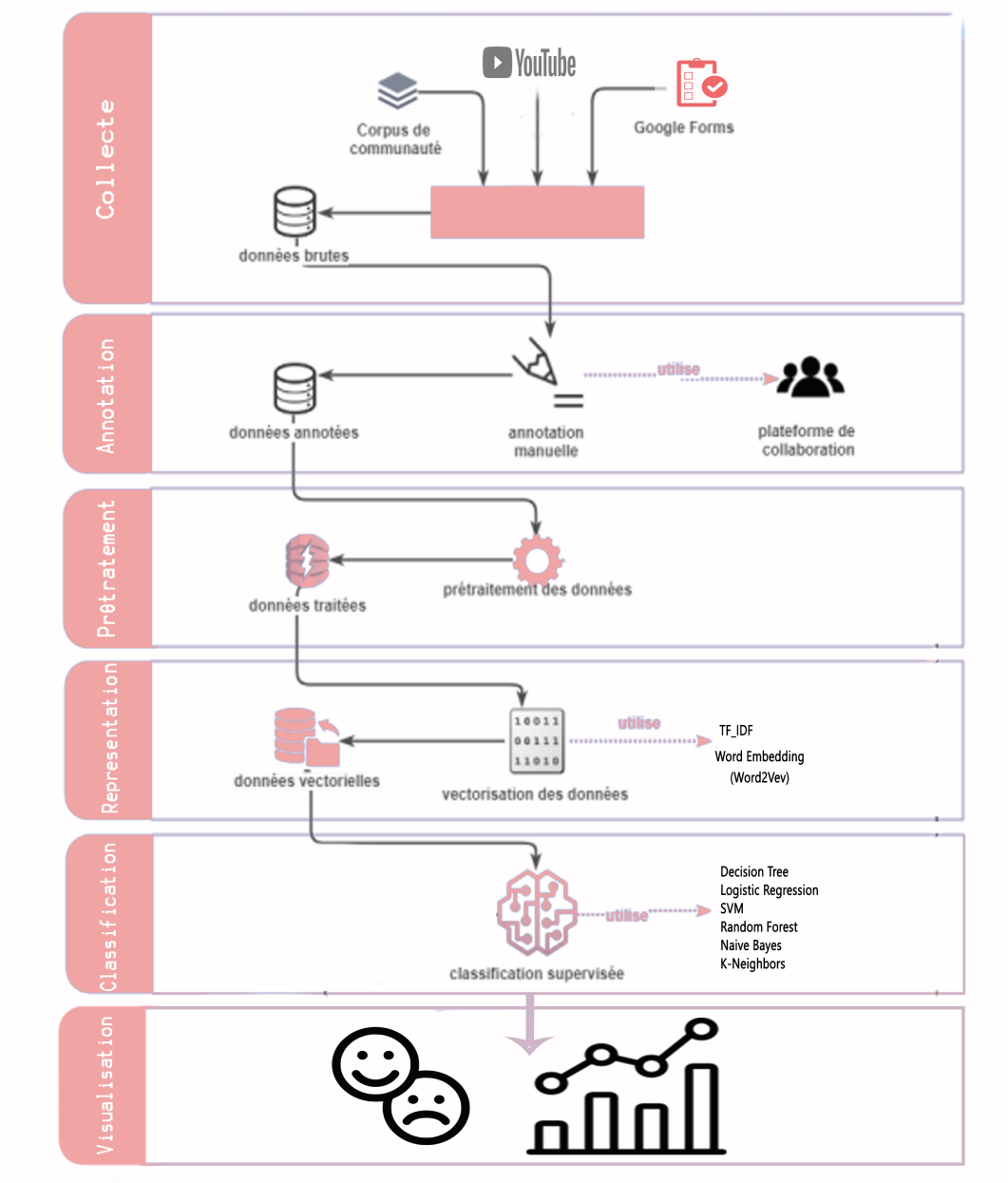


FIG. 5.2 : Architecture globale de notre système

5.3.1 Collecte de données :

La collecte ou l'extraction de données est la phase initiale du processus de l'AS, elle prépare la matière première pour effectuer cette tâche, et elle nous permet de créer des corpus afin de détecte la subjictivité des commentaire, en tenant compte des hypothèses de départ. Trois approches sont utilisées afin de collecter une masse importante de commentaires, dans le but de créer la collection des données (corpus) :

- **La 1ère approches :par Youtube**

  YouTube est un site web d'hébergement de vidéos sur lequel les utilisateurs peuvent

envoyer, évaluer, regarder, commenter et partager des vidéos. La récupération des commentaires sur YouTube se fait relativement à une vidéo et la collecte se focalise sur les vidéo liés au le Hirak . A partir de cette de ces dernières nous récupérons les commentaires postés.

- **La 2ème approche :** le crowdsourcing des commentaires, qui consiste littéralement à externaliser l'activité de collecte de commentaires vers la foule (crowd) c'est-à-dire vers un grand nombre d'utilisateurs.
- **La 3eme approche** : utiliser un corpus de communauté connu et précompilé . Toutes les données au format CSV.

5.3.1.1 Extraction de commentaires Youtube

L'extraction des commentaires Youtube sur le HIRAK s'effectue via API Youtbe qui reste la meilleure façon de réaliser cette tâche et obtension Access Token pour l'authentification.

Pour extraire des données via l'API :

1. Se Connecter à Google Developers Console.
2. Créer un nouveau projet.
3. Dans le nouveau tableau de bord du projet, cliquez sur Explorer et activer les API.
4. Dans la bibliothèque, accédez à l'API YouTube Data v3 sous les API YouTube.
5. Activez l'API.
6. Créer un justificatif d'identité.
7. Un écran apparaîtra avec la clé API.

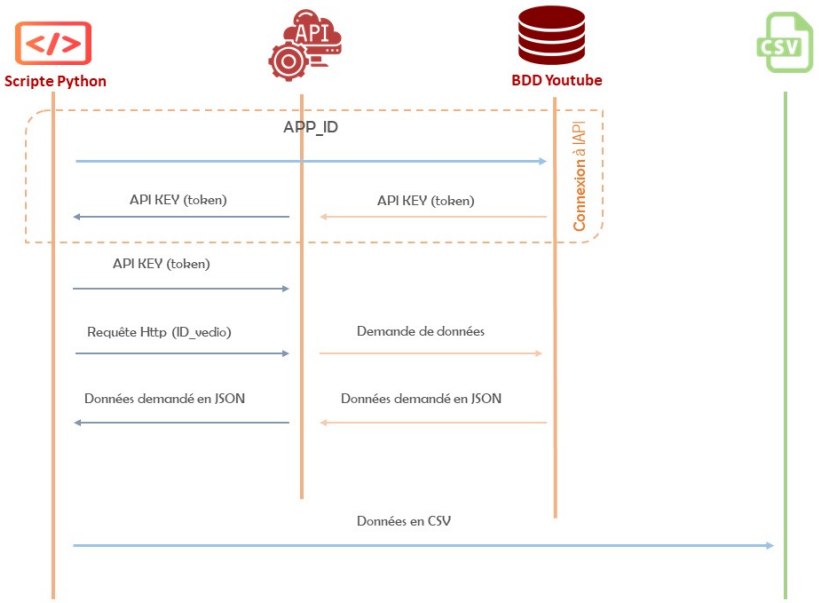


FIG. 5.3 : Diagramme de séquence de la collecte YouTube

Les champs « id » et « commentText » feront partie de la construction du corpus dans les prochaines étapes.

5.3.1.2 Financement participatif des commentaires

Connue sous le nom «crowdsourcing» ,cette méthode basée sur la contribution des utilisateurs et des internautes, pour arriver à collecter le maximum de commentaires et d'avis, est motivée par le manque des corpus contenant des commentaires en DAlG . Pour cela, un formulaire Google (*Fig*.5.4) a été conçu et crée. Il comprend simplement 03 champs, et ça afin d'inciter les utilisateurs à contribuer. Dans le but de le partager avec notre cercle d'amie et famille sur les réseaux : Le formulaire comporte les champs suivants :

- Le 1er et le 2em champ contiennent respectivement les commentaires (l'avis) en DALG avec les lettres arabes et arabizi
- • Le 3eme champ est réservé à l'annotation manuelle des commentaires via les libellés (positive/1, négative/0),

Le résultat est un fichier Excel contenant les trios champs « Avis(arabe) » , « Avis(arabizi) »et « polarité » (Fig 5.4)

FIG. 5.4 : Champs du formulaire google

5.3.2 Annotation

L'une des phases les plus importantes dans le processus d'analyse de sentiments basés sur l'apprentissage automatique est la création d'un corpus contenant des commentaires labellisés ou étiquetés qui vont servir par la suite comme corpus d'entrainement pour le modèle de classification supervisée et d'un corpus de test afin d'évaluer ce dernier.
Ainsi, notre système requiert un ensemble de documents dont la classe est déjà connue (étiqueté) pour construire un modèle qui pourra prédire la classe d'un nouveau message donné.
La qualité du modèle dépend initialement de la qualité des données annotées en entrée.
Pour créer un bon classificateur, il est préférable d'avoir un maximum de documents labellisés en entrée, ces lignes doivent être bien annoté, vu qu'on opte pour une annotation manuelle le processus est vraiment pénible et prend beaucoup de temps et demande une ressource humaine conséquente.
La variété des dialectes en Algérie ne facilite pas la tâche de l'annotation. Pour cela nous avons pensé a créer une plateforme de collaboration, dédiée a l'annotation manuelle dans le but de faciliter le processus. Les fonctionnalités :

- Affichage des documents annoter ligne par ligne avec possibilité d'avancer et reculer pour passer d'un document a un autre
- Création des étiquettes pour l'annotation
- Fusion et sauvegarde des différents résultats.
- conversion vere fichier CSV

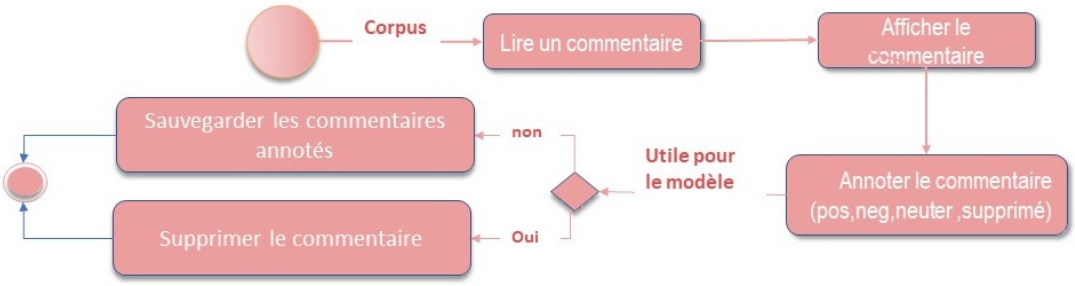


FIG. 5.5 : Diagramme d'activité du processus d'annotation

5.3.3 Prétraitement

La phase de prétraitement (Fig. 5.6) suit directement la collecte, cette étape est très importante et critique dans le processus de l'analyse de sentiments, puisque les données récupérées sont issues des sources différentes, avec une différence d'écriture remarquable, surtout en tenant compte du contexte algérien.

Les commentaires sont souvent dupliqués et peuvent contenir des caractères spéciaux ou des mots indésirables.

Tout le nettoyage se fait dans l'optique d'avoir un corpus avec un vocabulaire minimal d'une part, d'autre part, avec un maximum de mots informatifs, parlants, qui portents de l'information pertinente concernant la polarité.
Le détailler de tous les processuses de prétraitement par la suite :

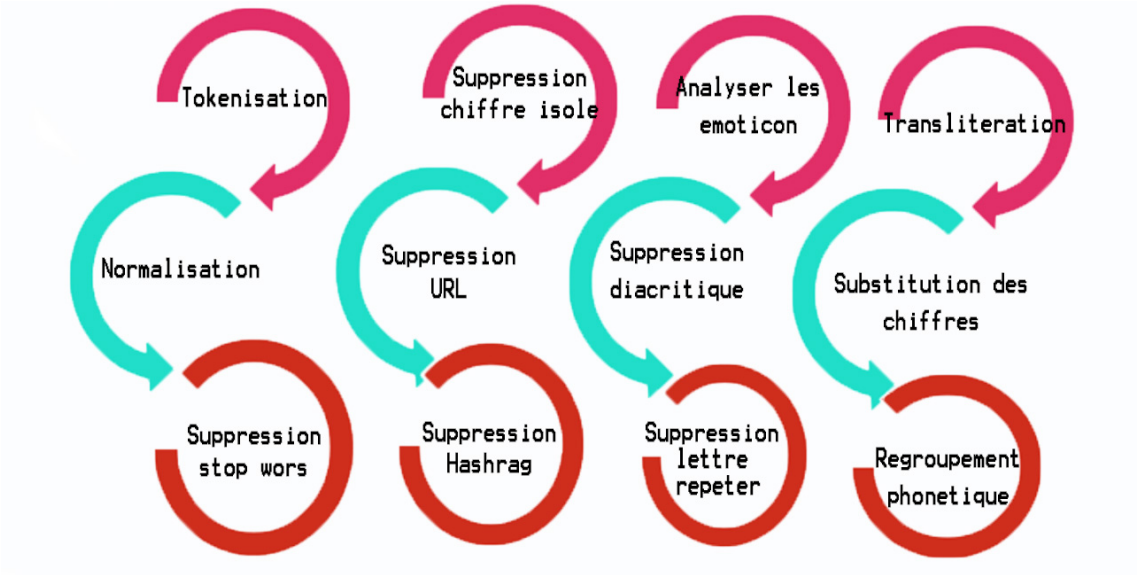


FIG. 5.6 : Processes de prétraitement

La phase du prétraitement est identifiée comme la phase la plus importante dans le processus de SA, puisque c'est la que l'on prépare nos commentaires afin de construire un classifieur robuste.
Comme nous pouvons le constater, sur les réseaux sociaux les internautes écrivent d'une

manière anarchique et les commentaires peuvent contenir des fautes d'orthographe, des liens,des Hashtags, des gifs, des stickers, des caractères spéciaux etc., que doit impérativement enlever afin d'avoir un corpus minimal, prêt et unifié.

La création de ce module nécessite un travail consistant et une maitrise avancée d'élément linguistique d'un langage ou d'un dialecte.

En effet, nous gardons les étapes qui sont départagées entre les langages réguliers et irréguliers comme

- **Tokenisation :** Ce module permet le découpage d'une chaine de caractère (Message ou commentaire) en mots dits « Token ». La tokenisation est encore plus importante dans l'analyse des sentiments que dans d'autres domaines de la NLP, car les informations sur les sentiments sont souvent peu représentées et inhabituellement représentées par des phrases.

  Entrée :[chofo had l'article tfahmo bzf swal7]
  Sortié :['chofo',' had', 'l'article' ,'tfahmo ' ,'bzf','swal7']

- **Élimination des liens URL :** Tout lien URL fourni n'aura pas vraiment d'effet sur le fait que le commentaire soit positif ou négatif, nous allons donc commencer par les éliminer d'abord de commentaire avant de poursuivre avec le texte restant.

  Entrée :[chofo had l'article https :// tfahmo bzf swal7]
  Sortié :[chofo had l'article tfahmo bzf swal7]

- **Substitution des hashtags :** L'hashtag est très représentatif en AS puisqu'il peut réorienter la polarité d'une phrase facilement. On l'a carrément supprimé de notre corpus et gardé juste le mot qui l'accompagne.
  Entrée :[mazal wa9rin #hirak_dz]
  Sortié :[mazal wa9rin hirak dz]

- **Suppression des lettres répétées :** on a supprimé les lettres répétées plus de deux fois dans un mot en les réduisant en une seule lettre. l'importance de cette phase est la réduction de la taille de notre corpus, malgré que portent un sentiment plus fort mais il est utile dans l'approche lexique, les mots dans lesquels a été écrit avec différents nombres de lettres mais sont les mêmes, seront stockés comme les mêmes mots.
  Entrée :[yetnahaaaaaaaaw gaaaaaaaaaaaaaaaaa3]
  Sortié :[yetna7aw ga3]

- **Suppression des mots vides :** les mots vides ou stop word en anglais sont les mots qui n'ont aucun effet sur la polarité de la phrase et ces mots sont plutôt fonctionnels ; conjonction, préposition, articles, etc. qui servent dans la structure d'un texte plus que dans son contenu.

  À cet effet, pour des langues comme l'anglais, le français ou l'arabe standard moderne, il existe des listes de mots outils bien connues. Ces listes sont disponibles gratuitement sur internet ou via des outils comme NLTK. (Schütze, Manning, and Raghavan 2008) À propos de DALG nous n'avons trouvé aucun ressources à l'exception de l'arabe standarde.

Entrée :[fi hirak kayn bzaf nas ghi ytab3o wsay]
Sortié :[hirak nas ytab3o]

- **Suppression de lettres isolées et de chiffres :** L'objectif de cette étape est similaire a celui de la suppression des mots vides, sauf que cette partie ne nécessite pas une liste au préalable, ces mots ne contribuent pas au corpus et n'ont pas de sens pour être analysés et représentés puis calculés et pris en compte. Et comme ils ne sont pas significatifs, eux et les chiffres, ils seront tous deux supprimés à cette étape. **Entrée :**[f kol jem3a nkhorjo b 3lamatna]
Sortié :[kol jem3a nkhorjo 3lamatna]

- **Élimination des diacritiques** Les signes diacritiques sont aussi éliminés dans cette étape du fait que ces signes ne sont généralement pas utilisés dans les textes arabes, et leur présence provoque une ambigüité. Aussi la suppression du caractère tatweel ' qui' n'a aucune influence sur le sens des mots (Alayba, Palade, England, Iqbal, 2017).

- Remplace les emoticon par leur polarité : L'analyse des emojis présente cependant certains biais, notamment celui de la subjectivité de la personne qui interprète le sens des commentaires , dans notre cas ne le néglige pas mais on remplacé pas leur polarité qu'il a été tiré d'une source mondiale qui l'a évalué avec des scours.
Entrée :[tahya dzayer <3 <3]
Sortié :[tahya dzayer positive positive]

- **Translittération :** Il est important pour nos données d'avoir un corpus unifié, donc nous ne pouvons pas permettre d'avoir une nature mixte de langues dans notre corpus. C'est à ce moment-là que nous sommes parvenus à la conclusion de l'opération de translittération, qui consiste essentiellement à transformer les lettres écrites arabes en lettres écrites latines tout en préservant la façon dont les lettres doivent être prononcées.

Entrée :[جيش شعب خاوا خاوا ( )]
Sortié :[jaych chaab khawa khawa]

- **Substitution des chiffres par des lettres :**Même si cela peut sembler déraisonnable, mais exprimer notre dialecte Algérien peut produire des mots si complexes que parfois les gens utilisent des chiffres au lieu de lettres. Par exemple, ils utilisent le nombre 9 pour exprimer(ق) et le nombre 3 pour exprimer(ع).. **Entrée :**[cha3b krah wbaghi yet7arar]
Sortié :[chaab krah wbaghi yetharar]

- **Regroupement phonétique :** Après avoir unifié les mots contenants des chiffres, nous passons au regroupement phonétique afin de maximiser son utilisation dans la catégorisation des commentaires, donc le résultat de ce regroupement constituera une ressource qui sera utilisée par la suite pour corriger les nouveaux textes. Le but etant de regrouper les mots qui se prononcent de la même manière citons par exemple :
('nshallah', nchallah, 'nchaalla)–>NXL

5.3.4 Représentation

Pour classifier les commentaires (texte), ils doivent être transformés (représentés) par des vecteurs et c'est la phase d'extraction des features. Au cours de ce processus, nous présenterons nos données en utilisant les deux méthodes de représentation :

- TF— IDF
- Plongement de mots(Word Embedding)

5.3.5 Classification

Ensemble de données (dataSet) prétraité et représenté d'une façon vectorielle (statistique), nous procédons ensuite a une série de tests pour déterminer le meilleur algorithme de classification supervisée qui donne les meilleurs résultats en fonction de la représentation choisie.

Pour ce faire nous avons choisi les algorithmes d'apprentissage automatique parmis les plus utilisés dans le domaine de NLP et de SA, à savoir :

- SVM Support Vector Machine
- NBM Naive Bayes
- DT Arbres de décision
- Regression logistic
- kNN
- Random Forest

En effet, il est difficile de prédire l'association qui donnera de meilleurs résultats avant d'avoir essayé les différentes combinaisons (Représentation – Classification), car la qualité de la prédiction est très liée au corpus et à la nature des textes.
Objectif classification de trouver le meilleur résultat possible avec la meilleure combinaison entre ces variations.

5.4 Conclusion

Dans ce chapitre, nous nous sommes concentrés sur le processus général de la méthodologie de notre système d'AS en expliquant le rôle de chaque sous-processus. Nous avons exploité les techniques de prétraitement les plus communes pour réduire le bruit dans le texte et avoir une meilleure et plus efficace analyse. Après avoir prétraiter les commentaires, l'étape qui suit est la classification des commentaires en fonction du sentiment exprimé : positif ou négatif .

Le chapitre suivant va porter sur les outils et les bibliothèques utilisés pour l'implémentation et la mise en œuvre de notre système.

Chapitre 6

Réalisation, tests et résultats

6.1 Introduction

Nous présentons, dans ce chapitre, les corpus que nous avons collectés, les expérimentations que nous avons faites, et les résultats obtenus. Dans nos expérimentations, nous avons traité les commentaires qui portent sur le Hirak de l'Algérie et qui sont écrits en arabe standard et/ou dialectal.

6.2 Environnement et outiles de développement :

Python :

On a choisi Python comme langage de programmation étant donné sa robustesse et son aptitude à être un langage multi-plateforme (Windows, Unix et MacOS). De plus, c'est un outil open source. Il est particulièrement répandu dans le domaine du TALN et du machine learning puisqu'il procure une souplesse de modulation et des fonctions existantes relatives au domaine du traitement automatique de la langue, de l'apprentissage automatique et de la classifcation. C'est avec ce langage de programmation que nous avons implémenté toutes les fonctions, méthodes et étapes de notre processus de fouille d'opinions.

Anaconda :

Est une plateforme de distribution Python recensant plus de 20 millions d'utilisateurs dans le monde, basé sur un écosystème totalement open-source. Anaconda contient Jupyter, qui est essentiellement une combinaison entre un IDE et un serveur pour exécuter vos Notebooks. Jupyter prend en charge aujourd'hui plus de 40 langages informatiques.
Ces fameux Notebooks, très appréciés dans la communauté des Data Scientists, contiennent à la fois du code et des éléments de présentation, tels que des images ou du texte, réunis en un seul endroit.

6.3 Bibliothèques utilisées

Bibliothèque	Déscription	Utilisation
NumPy	Ajoutant la prise en charge de grands tableaux et matrices multidimensionnels, ainsi qu'une grande collection de fonctions mathématiques de haut niveau pour fonctionner sur ces tableaux	Manipulation de données numériques.
pandas	Pour la manipulation et l'analyse des données. En particulier, il propose des structures de données et des opérations pour manipuler des tableaux numériques et des séries chronologiques.	manipulation et l'analyse des données
scikit learn	Destinée a l'apprentissage automatique, elle comprend notamment des fonctions pour l'apprentissage supervisé. Elle est conçue pour s'harmoniser avec d'autres bibliothèques libres Python,	Catégorisation des documents en dialecte algérien
GENSIM topic modelling for humans	Pour représenter des documents sous forme de vecteurs sémantiques	Répresentation des textes
NLTK	Pour travailler avec des données en langage humain, avec des corpus, de la catégorisation du texte, de l'analyse de la structure linguistique	Utilisé en prétraitement : tokenisation, suppression de mots vides
matplotlib	Pour créer des visualisations statiques, animées et interactives en Python.	visualisation de données
RegEx	Est une séquence spéciale de caractères qui utilise un modèle de recherche pour trouver une chaîne ou un ensemble de chaînes. Il peut détecter la présence ou l'absence d'un texte en faisant correspondre un motif particulier	nettoyage les texes

Tab. 6.1 : Bibliothèques utilisées

Et Bibliothèques utilisées pour les commentaires latine .

Aaransia : Son utilisation est le translittération des langues et dialectes .

Phonetics : son utilisation est le regroupement phonétique .

6.4 Réalisation

6.4.1 Collecte de données :

6.4.1.1 Utilisation Google API :

Une clé API est un identifiant unique qui permet d'authentifier les requêtes associées à le projet à des fins d'utilisation et de facturation. on doit associer au moins une clé API à projet ($Fig6.1, Fig6.2$).

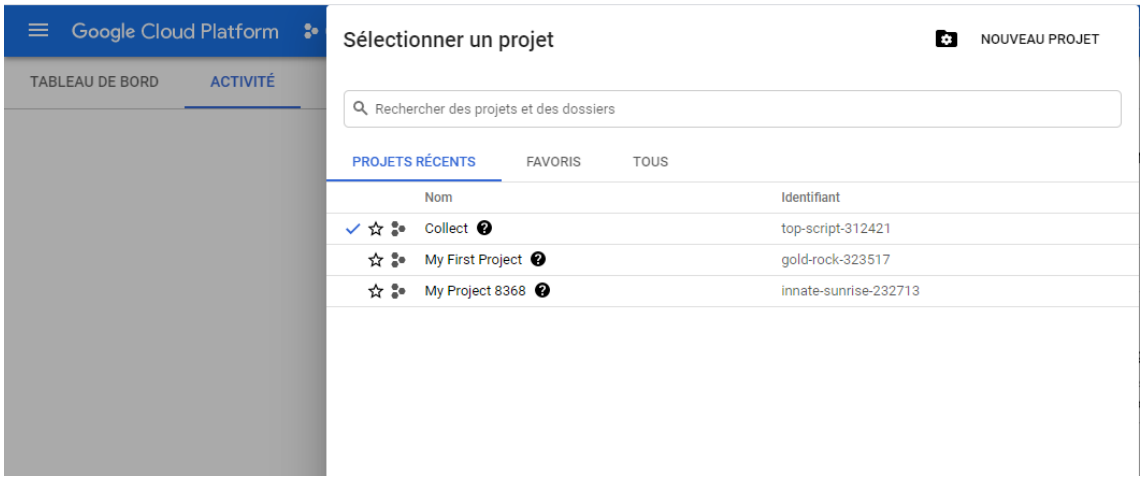


FIG. 6.1 : Platforme console google et creation nouvelle projet

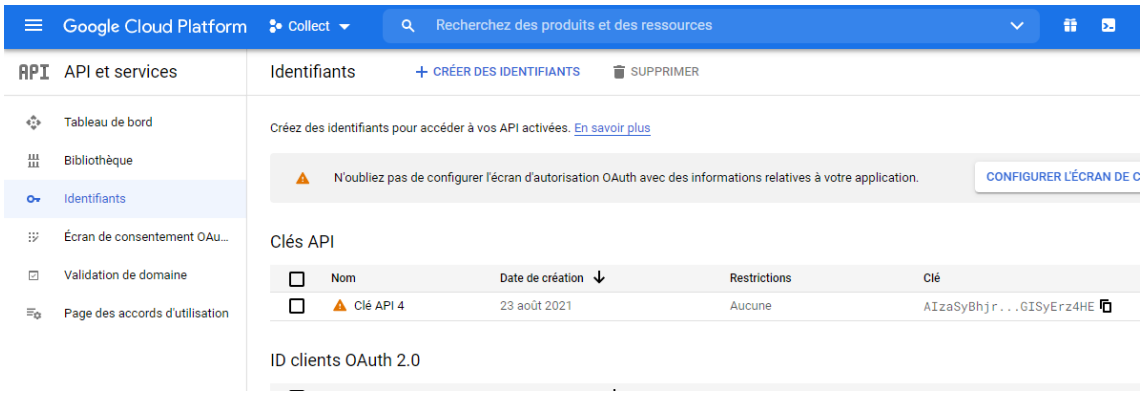


FIG. 6.2 : Prendre clés API

6.4.1.2 Extraction des commentaires :

La sélection des vidéos a été basée sur le contenu de la vidéo et le nombre d'interactions des utilisateurs.

Les figures ci-dessous représentent une capture d'écran de code d'extraction les commantaires de Youtube .

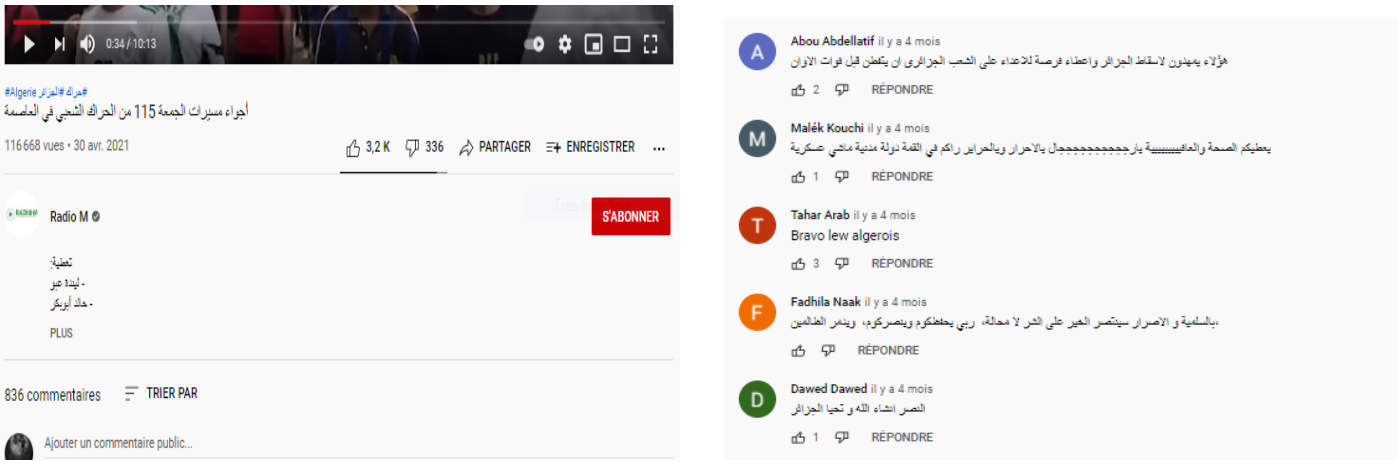
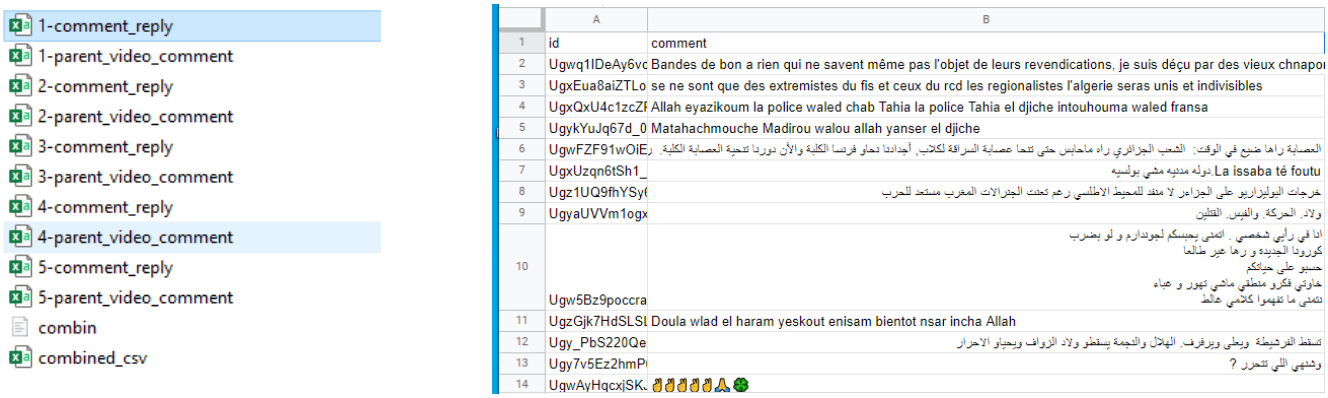


FIG. 6.3 : Vedio séléctionné de Youtube

```python
class VideoComment:
    def __init__(self,  videoId, key ):
        self.comments = defaultdict(list)
        self.replies = defaultdict(list)
        self.params = {
                    'part': 'snippet,replies',
                    'videoId': videoId,
                    'textFormat': 'plainText',
                    'key': key
                }

    def load_comments(self, mat):···

    def get_video_comments(self):···


    def create_df(self):···
```

FIG. 6.4 : Code source d'extraction des commentaires de Youtube

La récupération se fait sur les commentaire et leur réponses (reply comment) de video Youtube.
la structure d' extraction Reply comment diffère de celle du Comment. On les a regroupés dans une seule structure représentée par deux colonnes ID et Comment.



(a) Fichies csv extraits (b) Résulta de Combinaison

FIG. 6.5 : Résultats d'extraction de commentaires des videos Youtube

6.4.2 Collecte à travers Google Form :

En réalisant ce formulaire, pour but recueillir un échantillon d'avis de personnes de confiance et créer un corpus robuste, le contenu de leur commentaire serait une réponse directe sur leur opinion liés au Hirak, contrairement à ce qui a été collecté sur YouTube, et aussi pour déterminer la polarité qui nous aider dans la phase d'annotation.

Nous avons eu de gros problèmes pour publier ce formulaire et même pour le remplir, que ce soit par des amis ou via les réseaux sociaux, d'autant plus que les Algériens sont très sensibles à ce sujet. Même pendant la période où il a été publié, la situation était pleine d'événements politiques et de questions lourdes qui ont perturbé l'opinion publique, donc leur nombre était un peu faible.

FIG. 6.6 : Description de formulaire Google

Voici la liste des commentaires collectés du formulaire google et séparés en champ arabic et champ arabizi.

	A	B	C	D
1	Horodateur	تعليق بالدارجة (بالأحرف العربية)	تعليق بالدارجة (بالأحرف اللاتينية)	اذا كان التعليق اجابيا ا
2	2021/05/04 12:4	الحراك لم يكن له نتائج بالدرجة التي كان عليهت	7irakk ta3biiir 3and rafed cha3b llwa9iii3 ta3iis mota3aich	0
3	2021/06/18 12:3	الجيش شعب خاوه خاوه	Jich cha3b 5awa 5awa	1
4	2021/06/18 12:4	الشعب وعا و مابقاتش هاذيك النية تاع زمان لي نحبها للشعب	Ness karhet w jabet 7a9ha Les hommes	0
5	2021/06/18 12:4	البلاد بلادنا و نديرو رايبا	Leblad bladna w ndiro rayna	1
6	2021/06/18 12:4	يتنحاو قاع	Yatnahaw ga3	1
7	2021/06/18 12:5	حنا شعب تاع هدره برك مكان والو فالصح	hna cha3b ta3 hadra brk makan walo fsah	0
8	2021/06/18 12:5	شعب الجزائري يوقفو مع بعضاهم في وقت الشدة	Cha3b dzayri yo9fo me3a ba3dahahom fi wa9t cheda	1
9	2021/06/18 1:00	الحراك بدا حاجة مليحة بصح دركا رادي ضدو، الشعب تاعنا ما يحجبوش العجب لازم ير	Je pense dorka c'est le temp bch nakhdmo 3la rwa7na w r	1
10	2021/06/18 1:08	مدنية ماشي عسكرية	Madania mashi asskaria	0
11	2021/06/18 1:08	العربي المزابي القبايلي الشاوي خاوه خاوه	3arby chawi 9bayli mzabi khawa khawa	1
12	2021/06/18 1:08	حنا في حنا و اللي علطنا معاه بسامحنا	7na fi 7na o li ghlatna m3ah : ysama7na	1
13	2021/06/18 1:43	شحال من واحد خارج غير باش يسرق	Ch7al mn wa7d 5arj bch ysra9	0
14	2021/06/18 2:17	الحراك هذا يمثل فئة معينة كما حابذا نحتارمو رايهم ونوجهم لازم يحتارمو الاعلبية لي م	hirak hada ymetel fia mo3ana kima habina nhtarmou rayh	0
15	2021/06/18 4:35	رانا كامل خاوه خاوه مكاش فرق بيداتنا كامل كونتر العصابة.	Rana Kamel 5awa 5awa makach far9 binatna Kamel contr	1
16	2021/06/18 4:40	الحراك، الضرب اللول كنا متاحدين حتى لي ماخرجش كان خودا دوكا دخلت التفرقة بيناتنا	صعيب يراف باش تبنيو الثقة بين الشعب إذا كنا ماشي متفاهمين.	0
17	2021/06/18 6:16	أحسن حاجة صرات هو كي دردا الحراك، يعني صرا شوية تغيير ولو 1 بالمئة، و لا عد	Ahsan 7aja srat howa ki darna l7irak, ya3ni sra shwiya tag	1

FIG. 6.7 : Extrait des commentaires collectés via le formulaire

A	B
data	sentiment
لا فائدة من الحراك الا الفوضى و تعطيل مصالح الناس	positive
كان قضية شعب و وطن مبعدا كل واحد حوس على مصلحتو و بان فالتالي بلي مع الوقت الحراك ما ولاش عندو أهمية	negative
عرفنا بلي الشعب واعي و يوقف مع بعضاه و مهما كان مكانش لي راه يشتتنا كان هدفنا واحد	positive
شعب الجزائر مسلم والي العروبة ينتسب حدا مسلمين اكرمنا الله بلغة الضاد ولا عزاء للزواف .	positive
الشعب الجزائري يد وحده في كل الظروف	positive
الحراك ورالنا بلي الشعب الجزايري دايمن يد وحده	positive
تاثرت بالحراك بزاف وفرحت بيه لاخاطر عبر الشعب و الشباب علي رايهم	positive
الشعب الجزائري إخوه ويد واحده	positive
الشعب الجزائري في وقت الشده بيان، اليد في اليد تتسقم البلاد، بصح خاصنا نبداو نغيرو من رواحنا باش نغيرو النظام و	positive
واه نوقفو مع بعض	positive
حراك الشعب الجزايري كان ظاهره وحاجة مليحة الشعب نوعي برا البوفوار دارلنا كونطر خطة لعب على الوقت ولقادا	positive
7irakk ta3biiir 3and rafed cha3b llwa9iii3 ta3iiis mota3aich lakin lam yakoun nata2ijou	negative
Jich cha3b 5awa 5awa	positive
Ness karhet w jabet 7a9ha Les hommes	negative
Leblad bladna w ndiro rayna	positive
Yatnahaw ga3	positive

Fig. 6.8 : Fichier csv final de Formulaire Google

6.4.2.1 Résultats de collecte des donnéés

Dans cette phase, trois corpus différents sont exploités :
Le 1er est un corpus conçu et crée pour les commentaires collectés de Youtube.
Le 2em est un corpus aussi conçu et crée pour les contributions des amis et utilisateurs qui partagent les pages du Hirak sous forme de formulaires Google prêts pour l'annotation manuelle.
Le 3em est un corpus public librement disponible sur Internet et déjà annoté. Ce corpus va être utilisé comme jeux de données pour les tests d'expérimentation.
Les résultats des statistiques sont dans le tableau ci-dessus :

	Youtube	Form Google	Dataset
Nombre de Commentaires	8965	143	49864
Nombre des mots	33347	2215	2525704
Nombre de vocabulaires	12962	1400	177176

Tab. 6.2 : Statistiques sur les données collectées

6.4.3 Annotation :

DOCCANO est une plateforme open source de collaboration qui facilite le processus d'annotation. Elle a été employée pour annoter les données collectées de Youtube.

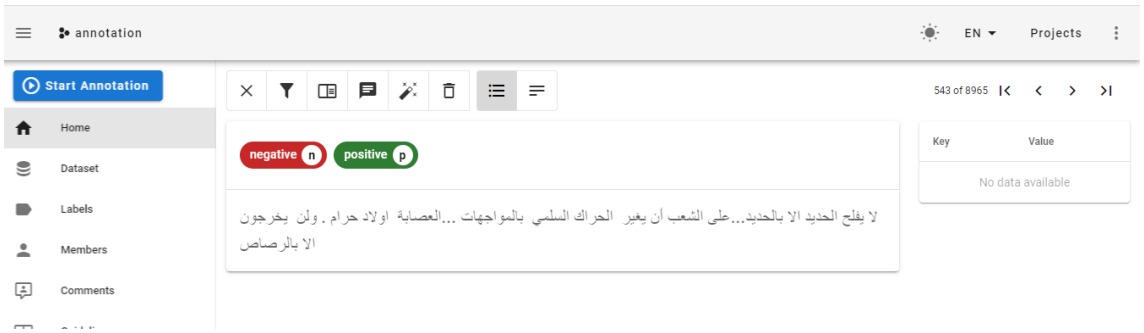


FIG. 6.9 : Annotation sur plateforme Doccano

6.4.3.1 Résultats d'annotation

Le nombre de commentaires collectés sur YouTube est Presque de 9K. Au cours du processus prétraitement et d'annotation, de nombreux commentaires ont été épurer comme présence d'instances dupliquées, les TAG d'amis ou de stickers et les commentaires hors sujet (invitation pour subscribe, Prière ; qoraan karim , appel au secours ..)

Cependant, 2227 commentaires sont utiles et significatifs et le reste sont des commentaires non significatifs. Cette illustration représente une distribution des classes (*Fig*6.10).

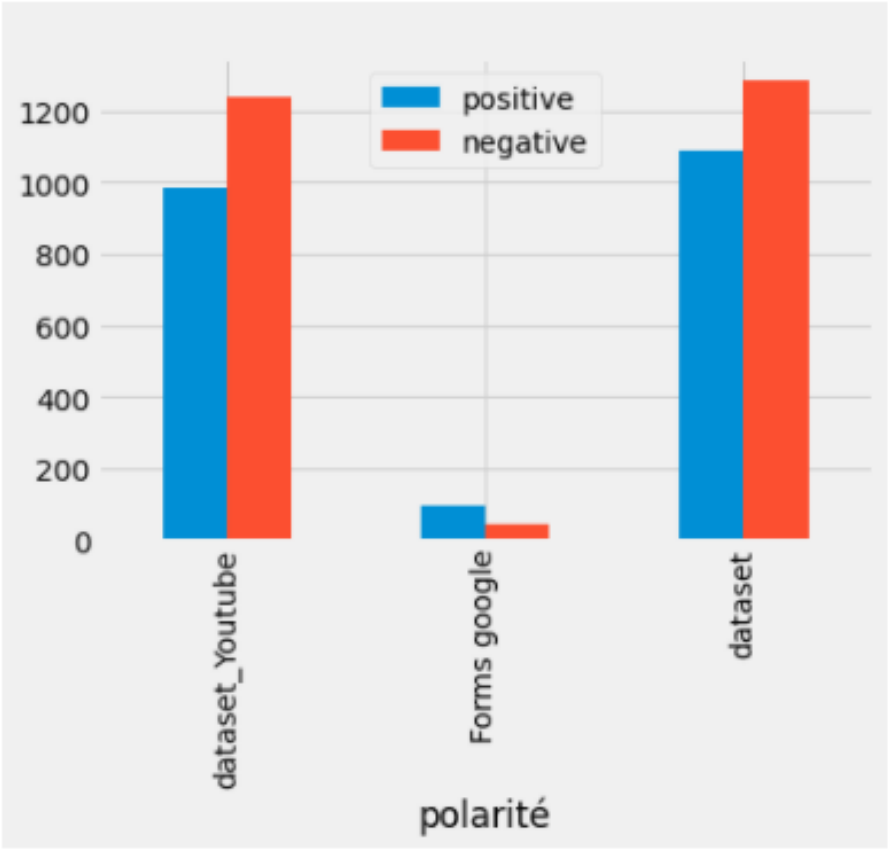


FIG. 6.10 : Distributions des classes dans chaque corpus

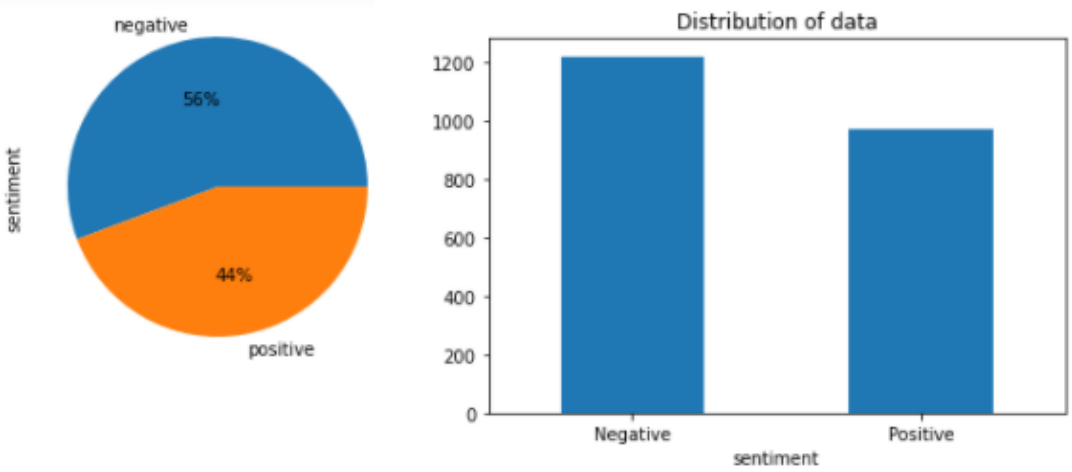


FIG. 6.11 : Distributions des classes dans tous les corpus

6.4.4 Prétraitement :

Les effets du prétraitement sur le nombre de mots dans notre dataset :

- Le prétraitment affine et supprime les mots non significatifs
- Le prétraitment réduit la taille du vocabulaire du corpus.
- Le prétraitment augmente la performance du modèle de l'analyse de sentiment.
- Regrouper les mots qui représentent la même instance, mais qui s'écrivent de plusieurs façons .

La figure (*Fig*.6.12) présente des statistiques sur le nombre de mots :

- S : sans prétraitement
- P : Données avec prétraitement
- LCF : LC+Phonétique (supprision des voyelles et regroupment les mots)

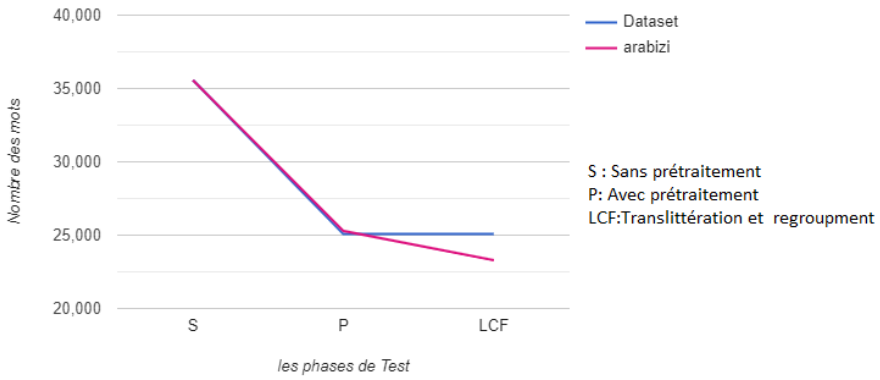


FIG. 6.12 : Evolution de la taille des mots dans certaines phases

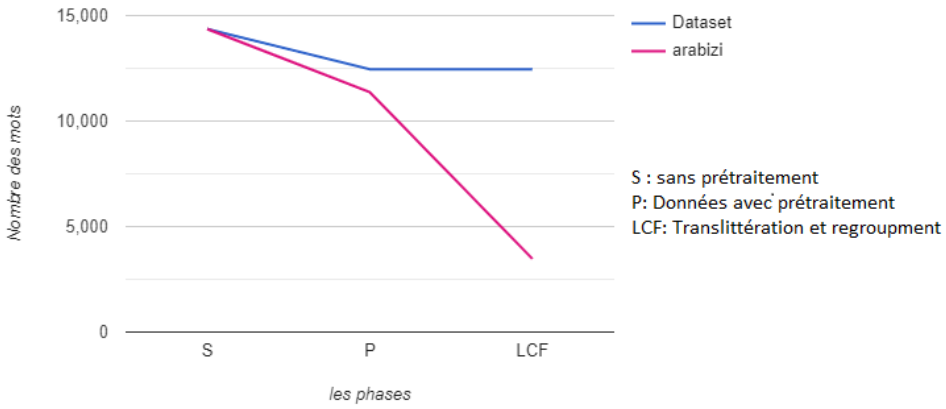


FIG. 6.13 : Evolution de la taille du vocabulaire dans certaines phases

6.5 Test et résultats

Dans section présente tous les résultats des tests obtenus après avoir utilisé les différents algorithmes de classification (Naïve Bayes, SVM et DT, RF,RL,KNN) les différentes méthodes de représentation ainsi que plusieurs types de prétraitement, nous allons donner par la suite des analyses et des explications à ces résultats de catégorisation.
Le but de cette partie est de vérifier que notre modèle détecte bien les opinions dans le dialecte Algérien et prédit bien leurs classes de polarités.

6.5.1 Plan de tests :

Notre stratégie consiste à utiliser plusieurs phase de traitement de données, représentations et algorithmes :

- **Les données** :
- S : Données sans prétraitement
- P : Données avec prétraitement
- A :Données contenant seulement les lettres Arabe
- L :Données contenant seulement les lettre latine (Arabizi)
- LF : L+Phonétique (supprision des voyelles et regroupement les mots)
- AC Transliteration Dataset complies à Arabe
- LC : Transliteration Dataset complies à Arabizi
- LCF : LC+Phonétique (supprision des voyelles et regroupment les mots)

6.5.1.1 Présentation

Nous l'avons testé pour voir laquelle donne le meilleur résultat en fonction des données en main, les méthodes que nous avons testées sont : • Méthode 1 : TF-IDF (ffréquence du terme – fréquence inverse du document) • Méthode 2 : Word Embedding

6.5.1.2 Algorithme de classification

Les algorithmes sur lesquels nous avons testé sont :

- Algorithme 1 : **Naive Bayes**
- Algorithme 2 : **SVM**
- Algorithme 3 : **KNN**
- Algorithme 4 : **Régression logistique RL**

- Algorithme 5 : **Arbre de décision DT**
- Algorithme 5 : **Random Forest**

6.5.1.3 Types de classification :

Nous allons effectuer des tests sur nos données Binaire (positive et négatives)

6.5.1.4 Métrique d'évaluation

Afin d'évaluer notre solution nous allons utiliser des métriques d'évaluations [partie 1], pour cela nous avons décidé d'utiliser l'exactitude

- l'exactitude : représenté par la proportion de prédiction correcte parmi toutes les prédictions, elle est une mesure très simple et très « intuitive ».

En utilisant la méthode de validation Croisée (cros_validation), nous discutons les résultats de nos expériences et les différentes méthodes de représentation avec notre jeu des données. Le tableau et les graphes rapporte les résultats obtenus par les modèles machine learning alimenté par les représentations Word Embedding et TF-IDF. Toutes ces variations et combinaisons sont effectuées dans le but d'atteindre un modèle robuste.

	NB		KNN		RL		SVM		DT		RF	
	TF_IDF	WE	TF_IDF	WE	TF_IDF	WE	TF_IDF	WE	TF_IDF	WE	TF_IDF	WE
S	76%	/	62%	53%	73%	57%	74%	60%	65%	57%	73%	57%
P	77%	/	65%	56%	75%	60%	75%	61%	64%	56%	71%	65%
A	79%	/	64%	54%	77%	54%	75%	54%	69%	59%	76%	63%
L	55%	/	46%	55%	56%	56%	55%	61%	57%	48%	55%	61%
LF	48%	/	41%	54%	45%	56%	45%	56%	59%	53%	48%	56%
AC	71%	/	41%	57%	73%	58%	71%	61%	71%	61%	71%	62%
LC	71%	/	61%	57%	74%	57%	72%	57%	67%	60%	70%	60%
LCF	72%	/	62%	51%	74%	57%	77%	58%	66%	58%	70%	60%

FIG. 6.14 : Résultats de classification avec différent algorithmes

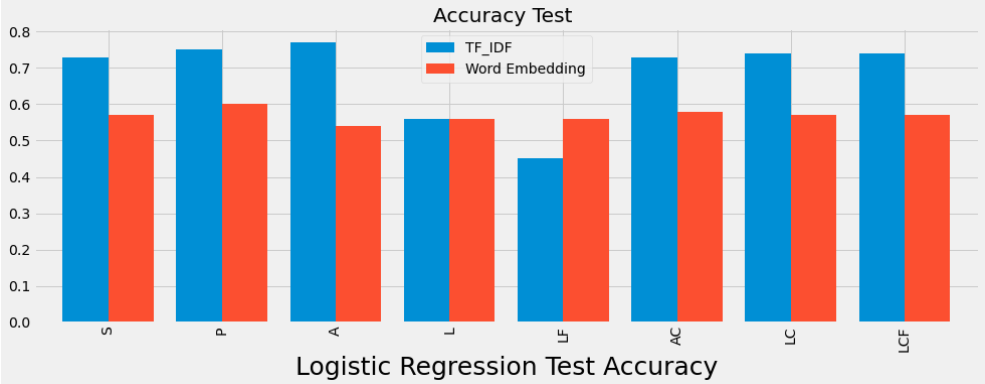


FIG. 6.15 : Résultats de Regression logistic pour toutes les phases de test avec les algorithmes de représentation

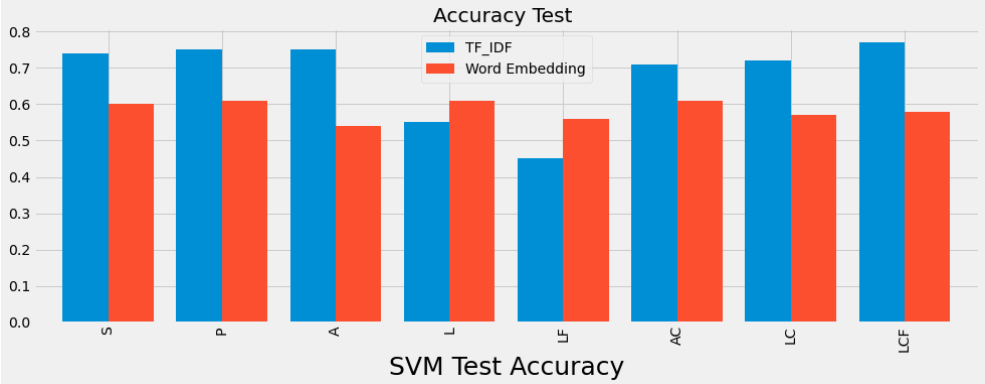


FIG. 6.16 : Résultats de SVM pour toutes les phases de test avec les algorithmes de représentation

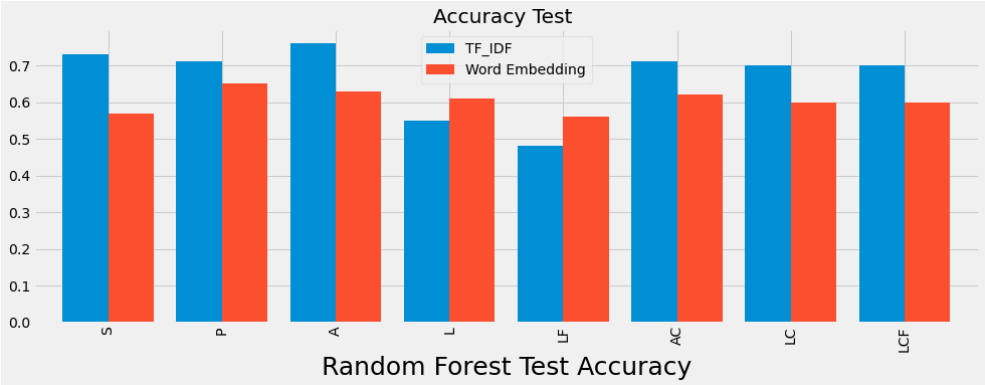


FIG. 6.17 : Résultats de Random Forest pour tous les phases de test avec les algorithms des représentation

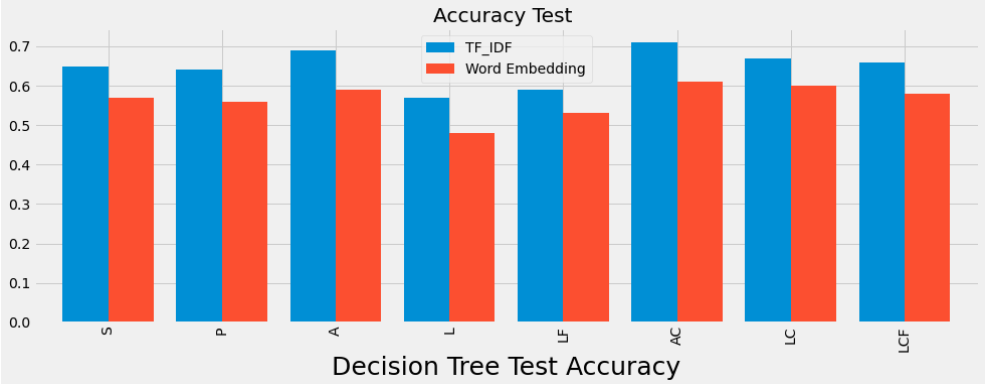


FIG. 6.18 : Résultats de Decision Tree pour tous les phases de test avec les algorithms des représentation

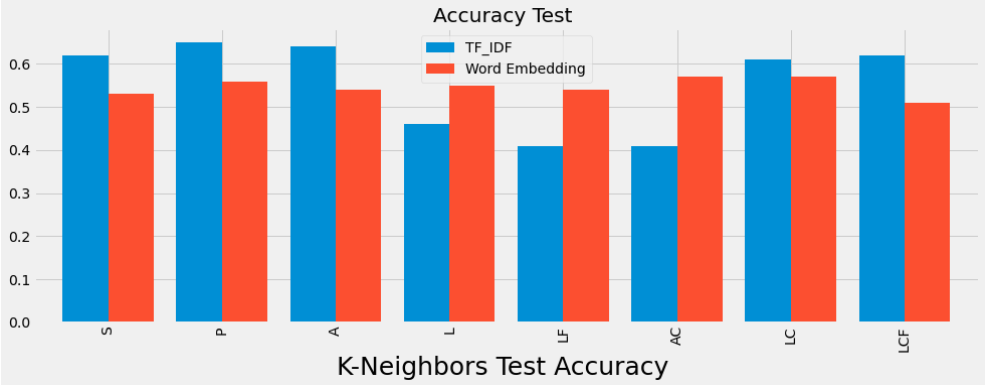


FIG. 6.19 : Résultats de KNN pour tous les phases de test avec les algorithms des représentation

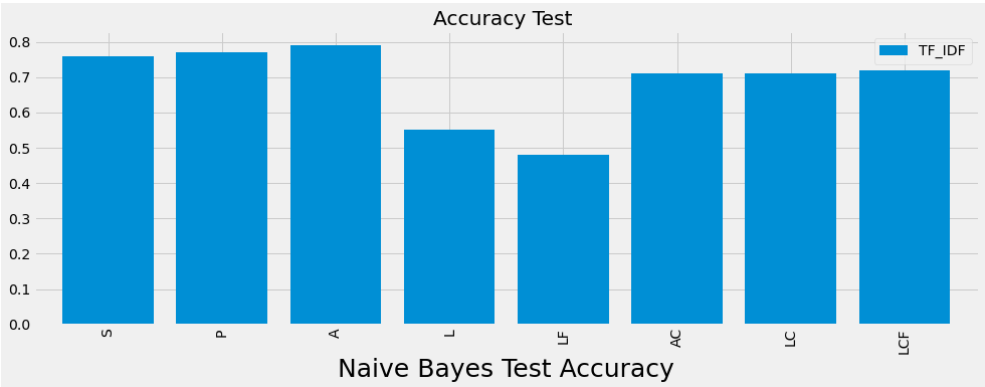


FIG. 6.20 : Résultats de KNN pour tous les phases de test avec les algorithms des représentation

6.5.2 Évaluation :

Dans l'état initiale de notre dataset **(phase S)** ,on remarque que Naïve Bayes donne un meilleur score de l'exactitude (accuracy 76%)avec la représentation TF_IDF , suivi par l'algorithme SVM (74%).

En effet, ce sont les classifieurs les plus utilisés par les chercheurs du domaine.

Naïve Bayes reste au premier rang et donné un bon résultat avec une exactitude 77% en fonction de **la phase de prétraitement (phase P)**et on regarde que RL a rejoint SVM avec percentage 75% également par la représentation TF_IDF ce que nous voyons dans la plupart des algorithms donner les scores les plus forts que la représentation Word Embedding .

Dans la phase de test qui gardé seulement les commentaires arabe **(phase A)** ,Naïve Bayes également donne de très bon résultat (79%) suivi par l'algorithmes RL,RF,SVM avec les valeurs varient de 75 % à 77% par la représentation TF_IDF .

Cependant , la phase AC qui traduise tous les commentaires l'arabizi à l'arabe à partire la fonction de transliteration , on remarque la diminution du resulta de RL à la valeur 73% et aussi les autres classifieurs SVM , RF ,DT,NB(accurcy 71%)

Dans l'état arabizi **(phase L)**,la représentation WE donne un meilleur score de l'exactitude 61% avec les classifieur SVM et RF tandis que les scores les rest d'algorthime avec les deux representateurs TF_IDF et WE étaient faibles .

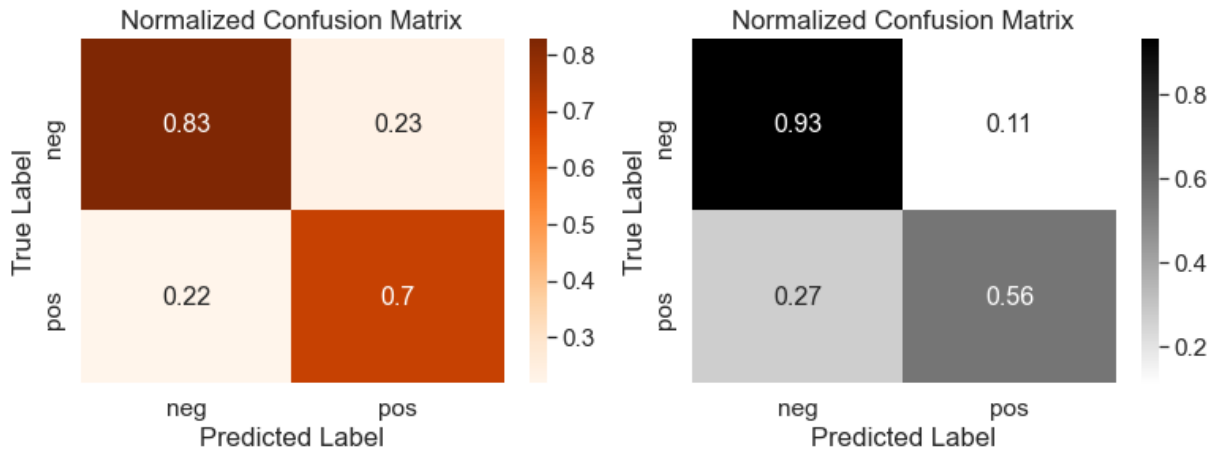
On remarque que Naïve Bayes donne un meilleur score de l'exactitude (accuracy 76%)avec la représentation TF_IDF , suivi par l'algorithme SVM (74%).

L'algorithme KNN a donné les scores les plus faibles, tandis que les scores de l'arbre de décision étaient au mieux modestes.

Et quand applique la phase LF (regroupement phénitique),on remarque que l'algorithme DT donné le meilleur resulta avec un accurcy faible (59%) par TD_IDf.

On Remarque que les meilleurs résultats sont donnés par le classifieur RL (74%) avec la representation Tf_idf et dans le WE avec accurcy 60% par les classifieurs DT et RF .

En fonction de la phase LCF ,on remarque la stabilité des résultats donnés par RF (60% avec WE) mais le meilleur score est donné aussi par le representation Tf_IDF avec l'algorithm SVM (77%) suivi par RL(74%) .



(a) Algorthime SVM (phase LCF)

(b) Algorthime NB(phase A)

Fig. 6.21 : Matrice de confusion des meilleurs résultats

6.5.3 Synthèse :

Après avoir testé sur jeux données différents, en utilisant de nombreuses combinaisons (**2** méthode de représentation X **6** Algorithme X **8** Jeux de données = **96** Test), nous avons beaucoup de résultats qui varient selon les méthodes, algorithmes et jeux données, le résultat le plus fréquent c'est les résultats qui ont des commentaires avec même nature : l'algorithme NB (79% accurcy) avec le TF-IDF donne les meilleurs scores avec les commentairs d'arabe seulement (la phases A). et aussi dans la phase LCF concerné les commentaires arabiz et qui applique la translitération et la regroupement phonétique à dataset complet ,l'algorithme SVM donné un bon résultat(77%) également avec la répresentation TF_IDF. les deux résultas sont des commentaires avec même nature (seulment arabe ou arabizi)

L'approche TF_IDF était une méthode de pointe. pendant de nombreuses années avant l'avènement des Word Embeddings l'algorithme word embadding (Word2Vec)fonctionnait mieux pour les tâches de classification mais dans notre travaille, nous avons remarqué l'inverse parce que l'ensemble de données que nous avons utilisé n'est pas assez grand pour permettre à l'algorithme d'apprendre les particularités du domaine .

Le prétraitement qui se base sur notre approche de groupement phonétique, et la translittération a amélioré les performances dans certain des algorithmes et des représentations. Nous constatons que la représentation TFIDF donne de meilleurs résultats que word Embadding.

Après plusieurs tests, KNN montre qu'il n'est vraiment performant en général, il donne les pires résultats dans pratiquement tous les tests .

6.5.4 Erreur d'analyse

Si nous analysons les commentaires mal classés nous trouvons beaucoup de commentaires ironiques ou sarcastiques, où des mots d'opinions positifs sont utilisés pour exprimer une opinion négative.

Par exemple : le commentaires « a7rar bravo » qui est de polarité négative, a été détecté comme positif par le classifeur.

Car le mot " bravo " qui est fortement utilisé dans les commentaires positifs et donc a une grande force polarité positive, et donc le classifeur à considérer que le commentaire était positif et ce malgré l'utilisation du mot Kachir qui est négative mais ayant une force de polarité plus faible que le mot bravo.

Aussi l 'erreur principale qui apparaît dans le processus de translittération est liée à la technique de regroupement phonétique . Dans certains cas, ces techniques renvoient un regroupment incorrect. Par exemple, le mot «الشعب» est translittéré comme « alcha3b » avec que ne resemble pas l'ecriture plus utilisé « cha3b » considéré comme un grand problème qui difficile pour le résoudre.

6.6 Conclusion

Dans ce chapitre, nous avons utilisé des techniques de domaines de catégorisation du texte et d'apprentissage automatique pour les besoins d'analyse des sentiments des commentairs ou nous nous sommes concentrés particulièrement sur la tâche de leurs classifications. Les expérimentations menées en passant par 03 collections de données et entrainant et testant deux modèles de ML nous ont permis de valider notre application et d'afficher les différents résultats. Aussi que, les représentations graphiques offertes par notre application permettent de mieux comparer et analyser les différents classificateurs.

Conclusion générale

Le but de ce mémoire de Master est générique et simple : classer le sentiment (positif/négatif) des commentaires extraits de Youtube et des formulaires Google en dialecte Algérien. Cependant, derrière ce simple objectif, différents défis ont été mis en évidence et traités.
Les défis les plus importants étaient :

- Arabizi : correspondant au fait d'utiliser l'écriture latine pour l'écriture mots arabes.
- Le manque de ressources dédiées au dialecte algérien : correspondant au manque de lexiques et corpus annotés.
- Le manque d'outils manipulant le dialecte.
- la complexité d'analyse sur le domaine politique (Hirak d'algérie).

nous avons distingué ces principales tâches pour développer un outil d'analyse de sentiment qui assure :

- la collecte et de sauvegarde des données,
- Le nettoyage des données.
- La représentation.
- La classification en classe souhaitée.

Nous avons présenté en détail les aspects théoriques et pratiques liés à l'implémentation de notre système d'analyse des sentiments. Il s'agit donc de la classification automatique des commentaires en trois classes : positive, négative ou neutre via les algorithmes traditionnels et en ML en passant par 03 collections de données (Dataset) pour la phase de prétraitement de données originales.

Le potentiel de l'utilisation du big-social data "Youtube" comme outil de mesure et de prédiction d'opinions pour le changement ou contre a été soigneusement démontré et les expérimentations menées ont permis de valider notre application. Les expériences faites ont montré que plus le niveau d'analyse est profond plus les résultats sont spécifiques.

Ce travail nous a permis de mettre plus en pratique nos connaissances théoriques sur les algorithmes traditionnelles et de ML ainsi que de les enrichir, et le plus important est

que nous fassions le premier pas vers l'apprentissage automatique, un des champs les plus importants de l'Intelligence Artificielle.

Parmi les orientations futures :

- Assemblage (hybridation/combinaison) des modèles de ML doit être une forme de technique d'algorithmes de méta-apprentissage dans laquelle nous proposons de combiner différents classificateurs afin d'améliorer la précision de la prédiction.
- Tester l'outil implémenté sur d'autres jeu de données et sur d'autres plateformes comme facebook, twitter ou autres
- Application d'autres classificateurs et utilisation d'autres fonctionnalités
- Création d'un lexique orienté Hirak

Bibliographie

[1] V. Hatzivassiloglou, K. McKeown in 35th annual meeting of the association for computational linguistics and 8th conference of the european chapter of the association for computational linguistics, **1997**, p. 174-181.

[2] T. M. Mitchell, *AI magazine* **1997**, *18*, 11-11.

[3] P. D. Turney, *arXiv preprint cs/0212032* **2002**.

[4] D. G. Myers, *Psychology : Seventh Edition New York NY : Worth Publishers* **2004**, *500*.

[5] M. Cord, P. Cunningham, *Machine learning techniques for multimedia : case studies on organization and retrieval*, Springer Science & Business Media, **2008**.

[6] X. Ding, B. Liu, P. S. Yu in Proceedings of the 2008 international conference on web search and data mining, **2008**, p. 231-240.

[7] A. Harb, G. Dray, M. Plantié, P. Poncelet, M. Roche, F. Trousset in INFORSID'08: INFormatique des Organisations et Systèmes d'Information et de Décision-Atelier FODOP'08, **2008**, p. 59-66.

[8] K. Ross, *Cross-Validation. In : Liu L Özsu MT eds. Encyclopedia of Database Systems. Boston : Springer* **2009**, 532-538.

[9] C. Zhang, D. Zeng, J. Li, F.-Y. Wang, W. Zuo, *Journal of the American Society for Information Science and Technology* **2009**, *60*, 2474-2487.

[10] T. O. Ayodele, *New advances in machine learning* **2010**, *3*, 19-48.

[11] N. Y. Habash, *Synthesis Lectures on Human Language Technologies* **2010**, *3*, 1-187.

[12] B. Liu et al., *Handbook of natural language processing* **2010**, *2*, 627-666.

[13] C. C. Aggarwal, C. Zhai in *Mining text data*, Springer, **2012**, p. 163-222.

[14] A. Kumar, T. M. Sebastian, A Kumar, T Sebastian, *Present and Future* **2012**, *4*, 1-14.

[15] B. Liu, *Synthesis lectures on human language technologies* **2012**, *5*, 1-167.

[16] K. Meftouh, N. Bouchemal, K. Smaïli in The third International Workshop on Spoken Languages Technologies for Under-resourced Languages-SLTU'12, **2012**, p. 1-7.

[17] E. Haddi, X. Liu, Y. Shi, *Procedia Computer Science* **2013**, *17*, 26-32.

[18] J Marcotte, *https ://commons.wikimedia.org/wiki/File : Plutchik-wheel$_f$r.svg*. **2013**.

[19] T. Mikolov, K. Chen, G. Corrado, J. Dean, *arXiv preprint arXiv :1301.3781* **2013**.

[20] A. Pasha, M. Al-Badrashiny, M. T. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, R. Roth in Lrec, *t. 14*, Citeseer, **2014**, p. 1094-1101.

[21] S. Shalev-Shwartz, S. Ben-David, *Understanding machine learning : From theory to algorithms*, Cambridge university press, **2014**.

[23] Y. LeCun, Y. Bengio, G. Hinton, *nature* **2015**, *521*, 436-444.

[24] B. Sablonnière, *La chimie des sentiments*, Odile Jacob, **2015**.

[25] M. Mataoui, O. Zelmati, M. Boumechache, *Research in Computing Science* **2016**, *110*, 55-70.

[26] L. Almuqren, A. Alzammam, S. Alotaibi, A. Cristea, S. Alhumoud in International Conference on Social Computing and Social Media, Springer, **2017**, p. 215-225.

[27] T. Chen, R. Xu, Y. He, X. Wang, *Expert Systems with Applications* **2017**, *72*, 221-230.

[28] A. Elouardighi, M. Maghfour, H. Hammia, F.-z. Aazi in 2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech), IEEE, **2017**, p. 1-8.

[29] Y. Goldberg, *Synthesis lectures on human language technologies* **2017**, *10*, 1-309.

[30] I. Guellil, F. Azouaou, H. Saâdane, N. Semmar, *Traitement Automatique des Langues* **2017**.

[31] S. Mdhaffar, F. Bougares, Y. Esteve, L. Hadrich-Belguith in Third Arabic Natural Language Processing Workshop (WANLP), **2017**, p. 55-61.

[32] C. Sammut, G. I. Webb, *Encyclopedia of machine learning and data mining*, Springer Publishing Company, Incorporated, **2017**.

[33] A. Ziani, N. Azizi, D. Schwab, M. Aldwairi, N. Chekkai, D. Zenakhra, S. Cheriguene in 2nd international conference on automatic control, telecommunications and signals, **2017**.

[34] A. Chader, D. Lanasri, L. Hamdad, M. C. E. Belkheir, W. Hennoune in KDIR, **2019**, p. 475-482.

[35] K. Elshakankery, M. F. Ahmed, *Egyptian Informatics Journal* **2019**, *20*, 163-171.

[36] M. Hadji, **2019**.

[38] F. J. Ramírez-Tinoco, G. Alor-Hernández, J. L. Sánchez-Cervantes, M. del Pilar Salas-Zárate, R. Valencia-García in *Current Trends in Semantic Web Technologies : Theory and Practice*, Springer, **2019**, p. 189-212.

[39] ultraalgeria, **2019**.

[40] R. Wang, J. Li in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, **2019**, p. 4135-4145.

[41] I. Guellil, A. Adeel, F. Azouaou, F. Benali, A.-E. Hachani, K. Dashtipour, M. Gogate, C. Ieracitano, R. Kashani, A. Hussain, *SN Computer Science* **2021**, *2*, 1-18.

[42] H. RAHAB.