RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
UNIVERSITE ECHAHID HAMMA LAKHDAR - EL OUED
FACULTÉ DES SCIENCES EXACTES

Département D'Informatique

Mémoire de Fin D'étude
Présenté pour l'obtention du Diplôme de
MASTER ACADEMIQUE

Domaine : **Mathématique et Informatique**
Filière : **Informatique**
Spécialité : **Systèmes Distribués et Intelligence Artificielle**

Thème :

Classification des commentaires à l'aide des fonctionnalités d'analyse des sentiments et de la pondération TF-IDF : cas l'extrémisme

Présenté par :

ADAIKA Raouia

BEN SAOUD Amal

Soutenue le 27 juin 2021 Devant le jury:

M. *Kholladi nadjoua houda* MCA Président

M. *Nadioui abd elhamid* MAA Rapporteur

M. *Med charaf eddine meftah* MCA Encadreur

Année universitaire : 2020/2021

Remerciement

Nous souhaitons Avant tout remercier le bon Dieu le Tout puissant, le Tout Miséricordieux, le Très Miséricordieux qui a tenu sa promesse et ne m'a jamais abandonné.

Comme nous tenons à exprimer toute notre gratitude et nos vives remerciements à notre encadreur de mémoire, Monsieur Meftah Mohammed Charaf eddine pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter notre réflexion et à réaliser le présent travail. Efforts qui resteront à jamais gravés dans notre mémoire. dieu seul puisse lui Exoser tous ses vœux et lui réserver tout le bonheur et la satisfaction qu'il mérite.

Nous adressons nos sincères remerciements au président et à tous Les membres du jury pour leurs patience et pour leurs judicieux conseils

DEDICACE

Ce projet de fin d'étude est dédié à mes chers grands parents et mes parents qui m'ont toujours poussé et motivé dans mes études, qu'ils en soient remerciés par cette trop modeste dédicace.

C'est un moment de plaisir de dédier cet œuvre à mon frère mohamed salim et mes sœurs nadjah et imane et douaa en signe d'amour, de reconnaissance et de gratitude pour le dévouement et les sacrifices dont ils ont fait toujours preuve à mon égard.

Je ne pourrais jamais oublier de dédier ce travail à

Mes étudiants pour dessiner un sourire sur mon visage

chaque matin Vous êtes le bonheur auquel j'aspire.....

Adaika raouia

DEDICACE

Je dédie cette recherche à chaque étudiant en sciences qui cherche à acquérir des connaissances et à apporter ses connaissances et son équilibre culturel

BEN SAOUD Amal

Résumé

Aujourd'hui, les réseaux sociaux pleins des textes dans lesquelles les internautes s'expriment en différents sujets, l'intérêt de leurs opinions est considérable, où la compréhension du contenu véhiculé par ces textes est un élément essentiel pour comprendre les tendances politiques, religieuses et même sociales des utilisateurs et pour ce là L'analyse des sentiments est une étape très importante pour atteindre la sécurité.

Dans ce travail, nous allons utiliser cinq algorithmes qui permettent d'analyser et classifier un ensemble de publications dérivées des réseaux sociaux. Les classes que nous avons définies sont : la classe positive, négative ou neutre.

Ce travail est utilise et comparer plusieurs algorithmes de classification pour déterminer l'état des commentaires de Twitter si elles est éxtrémistes ou pas.

Mots-clés : Analyse des Sentiments, fouille d'opinions, éxtrémistes, fouille de texte, détection émotionnelle, web social , éxtrémiste tweet .

Abstract

Today, the social networks full of the texts in which, the Internet-users express a different subjects, the interest of their opinions is considerable, where the comprehension of the content conveyed by these texts is an essential element To understand the users's political, religious and even social tendencies and for that sentiment analysis is a very important step for security.

In this work, we will use five algorithms that analyze and classify a set of publications derived from social networks. The classes that we have defined are: positive, negative or neutral. This experiment use and compare several comment classification algorithms at tweets to decide if it extremist or not .

Keywords: Sentiment Analysis , opinion mining , text mining, emotional detection, social web, extremist tweet.

ملخص

في هذه الأيام, نلاحظ أن الشبكات الاجتماعية مليئة بالنصوص التي يعبر فيها متصفحي الإنترنت عن أنفسهم في مواضيع مختلفة ,و في أرائهم أهمية كبيرة حيث يكون معنى المحتوى الذي تنقله هذه النصوص عنصرا أساسيا لفهم ميولات و توجهات المستخدمين السياسية و الدينية و حتى الاجتماعية وبهذا يعد تحليل المشاعر خطوة مهمة جدا لتحقيق الأمن . في هذا العمل سنستعمل خمس خوارزميات لتحليل و تصنيف مجموعة من التعليقات المستمدة من الشبكات الاجتماعية .

الفئات التي حددناها هي : إيجابية سلبية أو محايدة . يوظف هذا العمل و يقارن مجموعة من خوارزميات التصنيف على تعليقات التويتر و يحدد ما إذا كانت متطرفة أو لا .

الكلمات المفتاحية : تحليل المشاعر,التنقيب في الآراء ,التنقيب في النصوص ,الكشف عن العواطف ,الشبكات الاجتماعية ,التعليقات المتطرفة .

Table des matiéres

Chapitre III : Travaux connexes

Chapitre IV : Experimentation (Apprentissage , Test et Résultats)

Chapitre V : L'implémentation

Liste Des Figures

Liste des tableaux

LISTE DES ABRÉVIATIONS, ACRONYMES

TALN	**T**raitement Automatique Language Naturelle
IA	**I**ntelligence **A**rtificielle
UE	**U**nion **E**uropéenne
ML	**M**achine **L**earning
SVM	**S**upport **V**ector **M**achine
NB	**N**aive **B**ayes
RF	**R**andom **F**orest
DT	**D**ecision **T**ree
KNN	**K**-**N**earest **N**eighbhors
DBSCAN	**D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise
HTML	**H**yper **T**ext **M**arkup **L**anguage
TF-IDF	**T**erm **F**requency – **I**nverse **D**ocument **F**requency
Word2Vec	**W**ord **to V**ector
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
CD	**C**entrality **D**egree
C	**C**entrality
CB	**C**entrality **B**etweeness
PD	**P**restige **D**egree
SPAM	**S**piced **P**ork **A**nd **M**eat est une communication électronique non sollicitée
AFINN	A new word list for sentiment analysis on Twitter
ANEW	**A**ffective **N**orms for **E**nglish **W**ords
DL	**D**eep **L**earning

NLP	**N**atural **L**anguage **P**rocessing
API	**A**pplication **P**rogramming **I**nterface
CNN	**C**onvolutional **N**eural **N**etwork fr : réseau de neurones concurrentiels
LSTM	**L**ong **S**hort-**T**erm **M**emory
SOCMINT	**S**ocial **M**edia **I**ntelligence
ChCNN	**C**haracter **C**onvolutional **N**eural **N**etwork
ChSVM	**C**haracter **S**upport **V**ector **M**achine
RCNN	**R**egion Based **C**onvolutional **N**eural **N**etworks
ISIS	**I**slamic **S**tate of **I**raq and **S**yria
URL	**U**niform **R**esource **L**ocator
VIA	Définition : en : By way of ; through, fr : passant par .
RT	**R**E **T**weet
CSV	**C**omma **S**eparated **V**alues
RE	**R**egular **E**xpressions
Nltk	**N**atural **L**anguage **T**oolkit
SKLEARN	**S**cikits **L**earn
GENSIM	**G**enerate **S**imilar

Introduction Générale

Dès sa naissance, l'internet a ouvert d'innombrables possibilités à la société, il devenu un Outil incontournable d'échange d'information et partage les ressources , cette environnement numérique aider les gens communiquent, partagent de contenu dans les groupes de discussion, les blogs, les forums et autres sites pour expriment leur avis d'un grand nombre de sujets cela inclut l'expression d'opinions sociales et politiques. Toutefois, l'internet est souvent utilisé à mauvais escient par les extrémistes violents et les terroristes. En effet, les groupes extrémistes violents et terroristes recourent de plus en plus aux technologies de la communication en vue de lever des fonds, d'intimider, de former, de radicaliser, de recruter et d'inciter d'autres personnes à commettre des actes extrémistes violents et terroristes.

Pour ce la les gouvernements adoptent les mesures appropriées pour prévenir et lutter contre l'utilisation de l'internet à des fins d'extrémisme violent et de terrorisme (y compris sur les médias sociaux), tout en respectant la vie privée et les libertés d'expression, d'association, de réunion pacifique, de croyance ou de religion, ainsi qu'en respectant les impératifs d'une connectivité mondiale préservée et de la libre circulation de l'information en toute sécurité.

Parmi les mesures adoptées par le gouvernement un processus d'analyse des sentiments des postes, et des tweets ou d'avis afin de savoir ce que pensent les internautes. L'analyse des sentiments est une technologie d'analyse automatique des discours, écrits ou parlés et d'en faire ressortir les différentes opinions exprimées sur un sujet pr´ecis comme une marque, une actualité ou un produit. L'importance de l'analyse des sentiments est présente dans plusieurs domaines, `a savoir politique, marketing, gestion de la réputation, ...

L'analyse des sentiments relève de plusieurs disciplines en l'occurrence traitement automatique du la langage naturel (TALN) et de l'apprentissage automatique (Machine Learning).

Dans ce mémoire, notre objectif consiste à dévoiler les secrets de l'analyse des sentiments en adoptant une approche d'apprentissage automatique. Pour ce faire, nous avons implémenté cinq algorithme d'apprentissage sur des tweets appred dans le site kaggle. Nous avons considéré le modèle de représentation de données 'la pondération TF-IDF '. Les résultats obtenus en terme de précision révèlent que cette représentation mieux placée.

Organisation du manuscrit :

Notre travail structuré comme suit :

Dans le premier chapitre, nous allons parler de l'intelligence artificielle et de ses domaines, y compris l'apprentissage automatique, puis ses types et de ses importants algorithmes et leurs domaines d'application

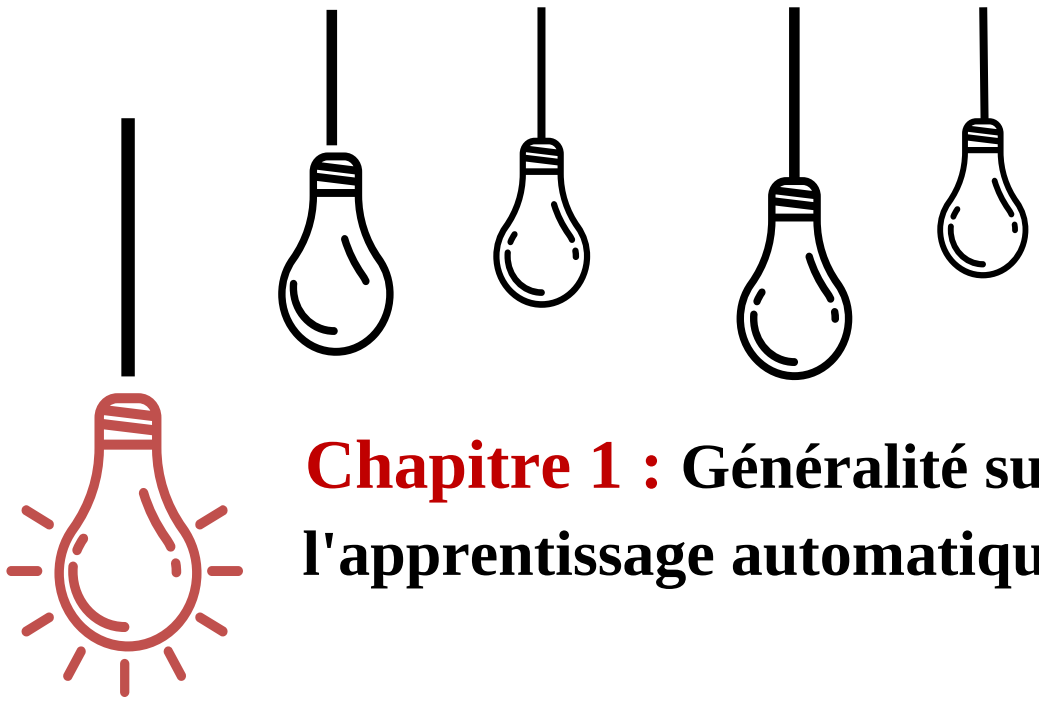
Dans le deuxième chapitre, nous introduit les caractéristiques principales des réseaux ,nous parlerons du modèle de traitement automatique des langues naturelles (TALN) et un de ses champs l'analyse des sentiments et opinion Mining. On présentera sa différents niveaux , leurs méthodes de classification et les outils d'analyse.

Et en le chapitre trois nous présentons les travaux liés à nos recherches ,ensuit on comparé les résultats de chacune des actions que nous avons mentionnées.

Ensuit , nous discuté sur l'approche appliquée pour collectant des données (dataset) et présenter et analyser les résultats des tests réalisés sur les données équilibrées et non équilibrée

A la fin le chapitre cinq présenter les outils de programmation et l'implémentation de notre application , présentation des interfaces et les résultats d'exécution.

Finalement, nous clôturons ce mémoire par une conclusion générale.



Chapitre 1 : Généralité sur l'apprentissage automatique

1. Généralité sur l'apprentissage automatique

1.1 Introduction

L'intelligence artificielle, c'est-à-dire le comportement et les caractéristiques spécifiques des programmes informatiques qui les font imiter les capacités mentales humaines et les modes de travail. L'une de ces caractéristiques les plus importantes est la capacité d'apprendre, de déduire et de réagir à des situations non programmées dans la machine.[1] L'apprentissage est réalisé à l'aide d'algorithmes capables de détecter des modèles et de générer des informations à partir des données qui leur sont présentées, pour les appliquer aux futurs processus de prise de décision et de prédiction, un processus qui évite d'avoir à planifier des étapes de manière spécifique pour chaque action potentielle seule.

1.2Intelligence artificielle

Avant de parler d'intelligence artificielle et des concepts qui y sont liés, il faut d'abord aborder les débuts de ce terme, comment il est apparu, quand et pourquoi?

1.2.1 Historique de l'Intelligence Artificielle

C'est en 1943 que Mc Culloch (neurophysiologiste) et Pitts (logicien) ont proposé les premières notions de neurone formel. Ce concept fut ensuite mis en réseau avec une couche d'entrée et une sortie par Rosenblatt en 1959 pour simuler le fonctionnement rétinien et tacher de reconnaître des formes. C'est l'origine du perceptron. Cette approche dite connexioniste a atteint ses limites technologiques, compte tenu de la puissance de calcul de l'époque, mais aussi théoriques au début des années 70. [2]

On peut considérer que le point de départ de l'intelligence artificielle a été donné avec les travaux d'Alan Turing qui posait le paradigme qu'une machine pouvait penser, notamment, dans son article « Computing Machinery and Intelligence», publié en 1950. Il propose un test, connu maintenant comme test de Turing, pour définir si la machine sait raisonner et peut se créer une autonomie. La notion même de l'IA a été ensuite conceptualisée et développée par John McCarthy de l'Université de Stanford et Marvin Minsky du MIT, pour ne citer qu'eux. [2]

1.2.2 Recherche précoce sur l'IA

Les premiers chercheurs en IA ont développé des algorithmes qui imitaient le raisonnement étape par étape que les humains utilisent lorsqu'ils résolvent des énigmes ou font des déductions logiques.À la fin des années 1980 et 1990, la recherche en IA avait également développé des méthodes très efficaces pour traiter des informations incertaines ou incomplètes, en concepts issus de la probabilité et de l'économie. [3]

La recherche d'algorithmes de résolution de problèmes plus efficaces est une priorité pour la recherche sur l'IA. [3]

1.2.3 Qu'est-ce que l'intelligence artificielle?

L'intelligence artificielle peut être définie comme «l'ensemble des théories et des techniques appliquées pour produire des machines capables de simuler l'intelligence», selon Larousse. Soit des ordinateurs, soit des logiciels performants généralement associés à l'intelligence humaine, amplifiés par la technologie:

- La capacité de penser
- La capacité de traiter de grandes quantités de données
- La capacité de distinguer les modèles et les modèles qui ne peuvent pas être détectés par l'être humain
- La capacité de comprendre et d'analyser ces modèles
- La capacité d'interagir avec les humains
- La capacité d'apprendre progressivement et d'améliorer continuellement ses performances [4]

1.2.4L'importance de l'intelligence artificielle

Certaines technologies associées à l'IA existent depuis plus de 50 ans, mais les progrès en terme de puissance de calcul, l'accès à une grande quantité de données et le développement de nouveaux algorithmes ont mené à des percées majeures dans le domaine de l'IA au cours des dernières années. [5]

L'intelligence artificielle est considéré comme un élément central de la transition numérique de la société et est devenue une priorité pour l'UE. Les futures applications de l'IA devraient mener à d'énormes changements - mais elle joue déjà un rôle dans notre quotidien [5]

1.2.5 Comment fonctionne l'intelligence artificielle

L'intelligence artificielle fonctionne en combinant et digérant un grand nombre de données. Elle s'appuie sur des programmes puissants d'exploitation. Mis à jour en permanence, ils témoignent d'une solide capacité à apprendre. L'IA est donc déterminée par 3 facteurs : la data, la puissance informatique et l'aptitude à apprendre. [6]

L'habilité à s'améliorer sans cesse rapproche d'ailleurs le plus la machine du modèle cognitif humain. On parle ainsi « d'intelligence ». Comme elle n'est pas organique mais fruit de la technique, on la qualifie « d'artificielle » [6]

1.3 Apprentissage automatique

1.3.1 Définition

L'apprentissage machine est une application d'intelligence artificielle (IA) qui permet aux systèmes d'apprendre et de s'améliorer automatiquement à partir de l'expérience elle-même sans être explicitement programmée. [7]

Ce processus d'apprentissage commence par des observations ou des données, comme des exemples, une expérience de première main ou des instructions, dans le but de rechercher des modèles dans les données et de prendre de meilleures décisions à l'avenir sur la base des exemples que nous fournissons. Votre objectif premier est de permettre aux ordinateurs d'apprendre automatiquement sans intervention ou assistance humaine et d'adapter les actions en conséquence. [7]

1.3.2 Types d'apprentissage automatique

L'apprentissage automatique ou Machine Learning (ML) est une branche de l'intelligence artificielle qui consiste à programmer des algorithmes permettant d'apprendre automatiquement à partir des données et d'expériences passées ou par interaction avec l'environnement. Ce qui rend l'apprentissage machine vraiment utile est le fait que l'algorithme peut "apprendre" et adapter ses résultats en fonction de nouvelles données sans aucune programmation à priori.

Il existe plusieurs façons d'apprendre automatiquement à partir des données dépendamment des problèmes à résoudre et des données disponibles. La figure 3.1 donne un sommaire des types d'apprentissage automatique les plus connus

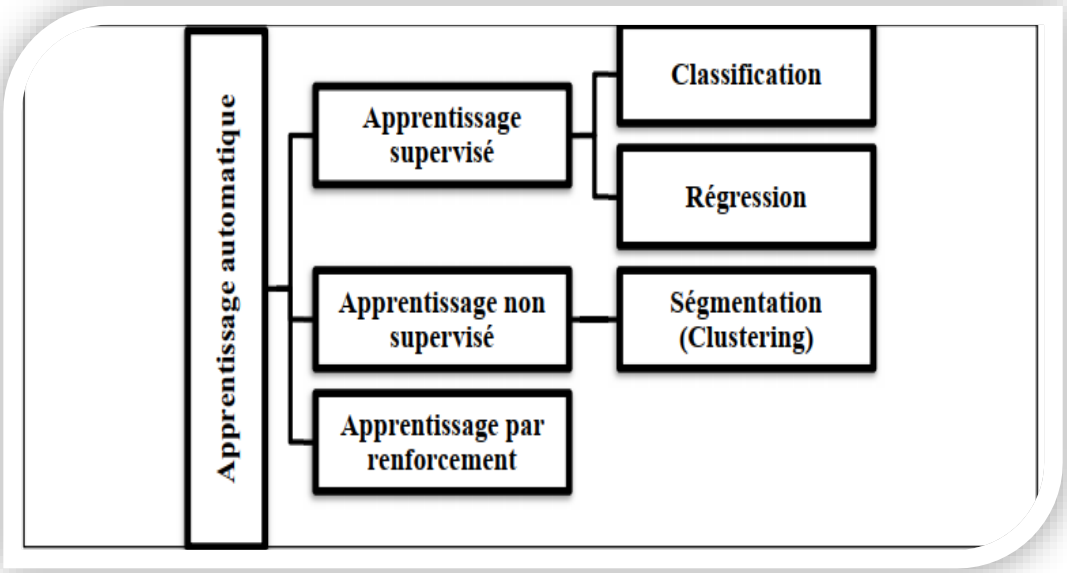


Figure 1.1: les grandes classes d'apprentissage automatique[8]

1.3.2.1 Apprentissage supervisé

L'algorithme est entrainé en utilisant une base de données d'apprentissage contenant des exemples de cas réels traités et validés. L'objectif est de trouver des corrélations entre les données d'entrée (variables explicatives) et les données de sorties (variables à prédire), pour ensuite inférer la connaissance extraite sur des entrées avec des sorties inconnues.
En apprentissage supervisé, on distingue entre deux types de tâches :Des tâches de classification ET Des tâches de Régression [8].

1.3.2.1.1 Algorithmes scientifiques supervisés

Nous leur montrons ce qui suit

a) Régression logistique:

Il s'agit d'un algorithme de classification utilisé lorsqu'une variable de réponse est catégorielle. L'idée de la régression logistique est de trouver une relation entre les propriétés et la probabilité d'un résultat donné. [11]

Par exemple, lorsqu'on doit prédire si un étudiant a réussi ou échoué un test alors que le nombre d'heures qu'il a passé à étudier est donné comme caractéristique, la variable de réponse a deux valeurs, réussite et échec. . [11]

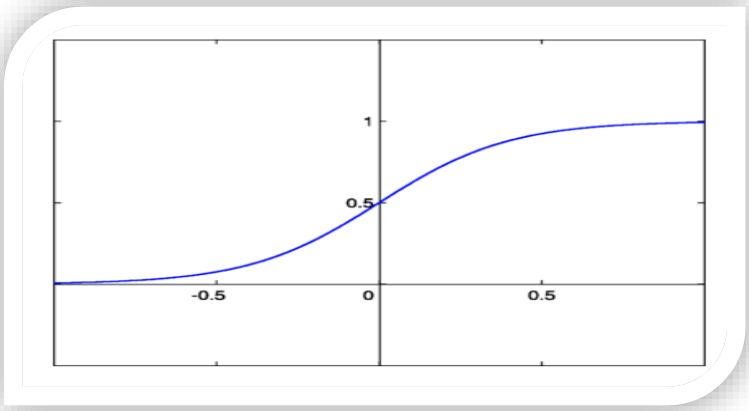


Figure 1.4 :Régression logistique [11].

b) Support Vector Machine (SVM)

Machine à Vecteurs de Support (SVM) est lui aussi un algorithme de classification binaire. Tout comme la régression logistique. Si on prend l'image ci-dessus, nous avons deux classes (Imaginons qu'il s'agit de e-mails, et que les mails Spam sont en rouge et les non spam sont en bleu). La régression Logistique pourra séparer ces deux classes en définissant le trait en rouge. le SVM va opter à séparer les deux classes par le trait vert. [12]

Sans entrer dans les détails, et pour des considérations mathématiques, le SVM choisira la séparation **la plus nette possible** entre les deux classes (comme le trait vert). C'est pour cela qu'on le nomme aussi **Large Margins classifier (classifieur aux marges larges)**. [12]

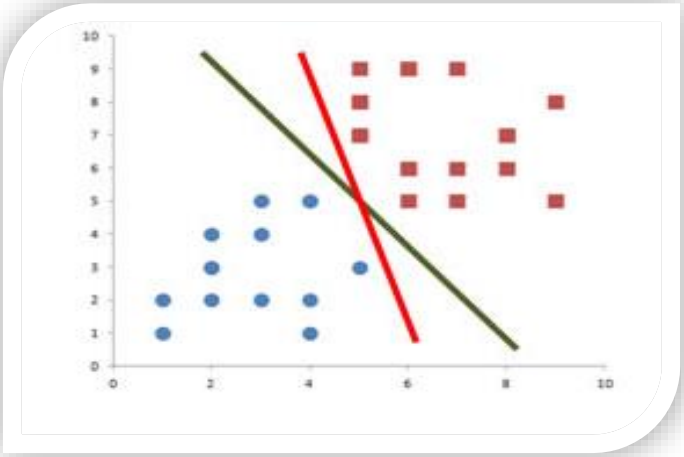


Figure 1.5 : Support Vector Machine (SVM) [12].

c) K-plus proches voisins (KNN)

Algorithme des k plus proches voisins L'algorithme des k-plus proches voisins, est une méthode d'apprentissage à base d'instances. Il ne comporte pas de phase d'apprentissage en tant que telle. Les documents faisant partie de l'ensemble d'apprentissage sont seulement enregistrés. Lorsqu'un nouveau document à classer arrive, il est comparé aux documents d'apprentissage à l'aide d'une mesure de similarité. Ses k plus proches voisins sont alors considérés : on observe leur catégorie et celle qui revient le plus parmi les voisins est affectée au document à classer. C'est là une version de base de l'algorithme que l'on peut raffiner. . [12]

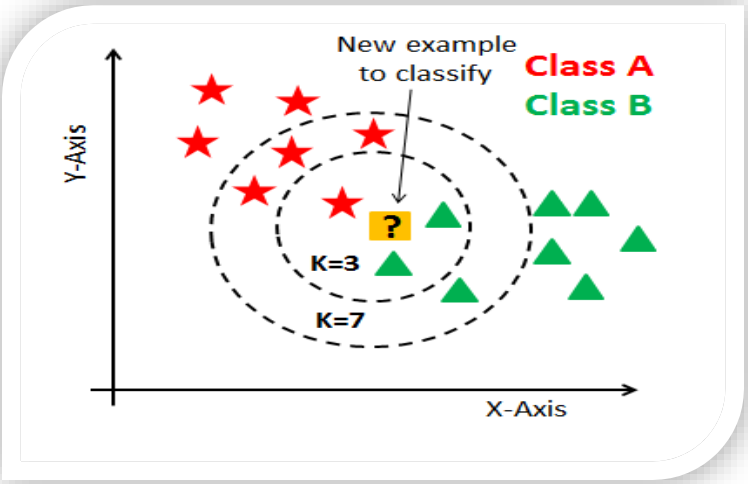


Figure 1.6 : Classification KNN [12] .

d) Naïve Bayes

Naive Bayes est un algorithme de classification qui suppose que les prédicteurs d'un ensemble de données sont indépendants. Cela signifie qu'il suppose que les caractéristiques ne sont pas liées les unes aux autres. Par exemple, si on lui donne une banane, le classificateur verra que le fruit est de couleur jaune, de forme oblongue et long et effilé. Toutes ces caractéristiques contribueront indépendamment à la probabilité qu'il s'agisse d'une banane et ne dépendent pas les unes des autres. Naive Bayes est basé sur le théorème de Bayes, qui est donné par :

$$P(A \mid B) = \frac{P(A)\, P(B \mid A)}{P(B)}$$

Figure1.7 : Bayes' Theorem[13]

Où :

P(A | B) = combien de fois se produit étant donné que B se produit

P(A) = quelle est la probabilité que A se produise

P(B) = quelle est la probabilité que B se produise

P(B | A) = combien de fois B se produit étant donné que A se produit . [13]

e) Classificateur d'arbre de décision

L'arbre de décision est un modèle prédictif modéré. Il peut apprendre à prédire la sortie en répondant à des questions basées sur les valeurs de l'entrée qu'il reçoit. [14**]**

C'est une méthode bien connue et populaire dans la science des données et est utilisée dans les opérations de prédiction et de classification ,Les arbres de décision utilisent différentes «propriétés» pour diviser les données à plusieurs reprises en sous-ensembles jusqu'à ce que les sous-groupes soient purs, ce qui signifie qu'ils partagent tous la même valeur cible.Pour mesurer la pureté d'un sous-ensemble[14**]**

Dans un arbre de décision, n'importe quel attribut peut être utilisé pour diviser les données, bien que l'algorithme sélectionne l'attribut et la condition en fonction de la pureté des sous-groupes résultants (la pureté d'un sous-ensemble est déterminée en s'assurant des valeurs cibles dans le sous-ensemble). [14**]**

Les arbres de décision présentent un certain nombre d'avantages et d'inconvénients qui doivent être pris en compte pour déterminer s'ils peuvent être utilisés dans une situation particulière[14**]**

Les fonctionnalités incluent:

1. Peut être utilisé pour la régression ou la classification
2. Peut-être affiché graphiquement
3. Haute capacité d'interprétation
4. Il est plus proche du processus décisionnel humain que les autres algorithmes de ML
5. Prédiction rapide
6. Les propriétés n'ont pas besoin d'être développées
7. apprend automatiquement les interactions de propriété
8. Il a tendance à ignorer les fonctionnalités qui ne sont pas pertinentes

Les inconvénients des arbres de décision sont les suivants:

1. Ses performances ne concurrencent généralement pas les meilleures méthodes d'apprentissage supervisé
2. De petites différences de données peuvent conduire à un arbre complètement différent
3. La division binaire itérative prend des décisions qui peuvent ne pas conduire à un arbre globalement idéal
4. Cela n'a pas tendance à bien fonctionner avec des classifications très déséquilibrées
5. Cela n'a pas tendance à bien fonctionner avec de très petits ensembles de données . [14]

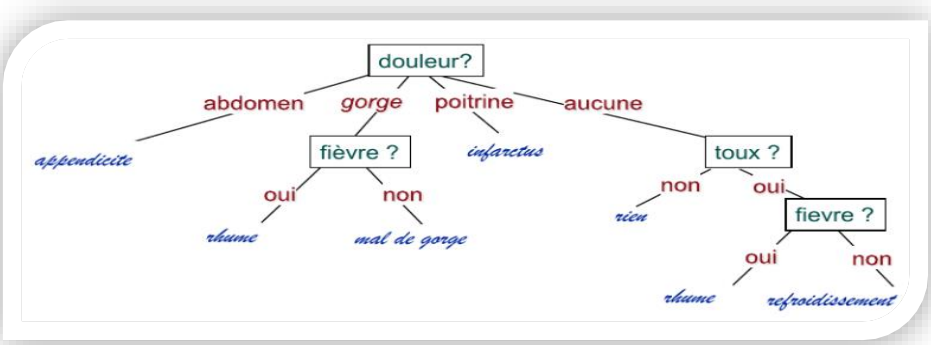


Figure 1.8 : Exemple pour Classificateur d'arbre de décision[15]

f) Classificateur de forêt aléatoire

Nous pouvons comprendre le fonctionnement de Algorithme Random Forest à l'aide des étapes suivantes

Étape 1 - Commencez par sélectionner des échantillons aléatoires parmi un ensemble de données donné.

Étape 2 - Ensuite, cet algorithme construiraun arbre de décision pour chaque échantillon. Ensuite, il obtiendra le résultat de la prédiction de chaque arbre de décision.

Étape 3 - Dans cette étape, le vote sera effectué pour chaque résultat prévu.

Étape 4 - Enfin, sélectionnez le résultat de prédiction le plus voté comme résultat final de la prédiction. .[**17**]

Le schéma suivant illustrera son fonctionnement -

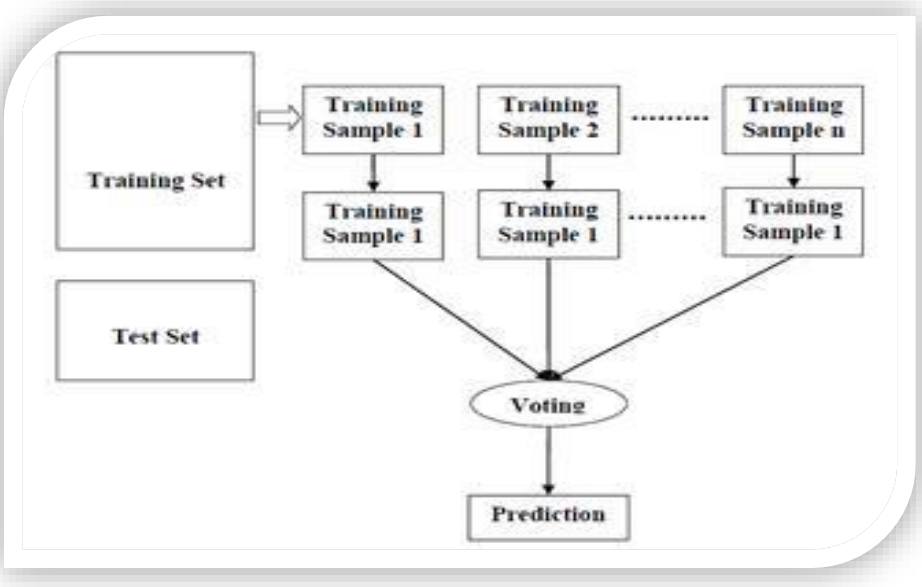


Figure 1.9: Algorithmes de classification - Forêt aléatoire.[**16**]

1.3.2.2 Apprentissage non supervisé

Pour ce type d'apprentissage la base de données d'apprentissage ne contient pas de variable cible (comme on l'a vu en apprentissage supervisé). Il y a seulement un ensemble de données collectées en entrée. L'algorithme doit découvrir par lui-même la structure en fonction des données. On utilise cette technique pour partitionner les données en groupes d'éléments homogènes. La distance est souvent la plus utilisée comme mesure de similarité entre les groupes [8]

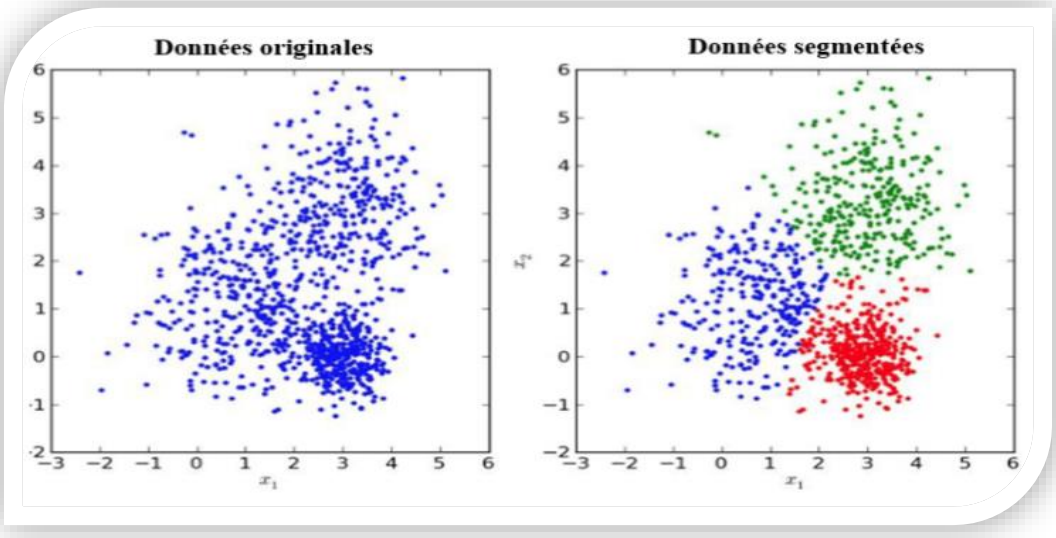


Figure 1.2: Exemple d'apprentissage non supervisé . [8]

1.3.2.2.1 Algorithmes scientifiques non supervisés

Nous leur montrons ce qui suit

a) Algorithme K-Means

L'algorithme de clustering K-Means est l'un des algorithmes de clustering les plus populaires parmi les data scientists.

L'un des avantages de l'algorithme K-Means est qu'il est très rapide ; En raison du petit nombre de calculs utilisés pour déterminer les clusters et leurs centres, la complexité de l'algorithme et le temps consommé pour l'ensemble du processus n'augmentent que linéairement avec l'augmentation du nombre total de points dans les données O(n).

Par contre, il y a quelques points négatifs :

• Comme on le sait, la détermination du nombre de groupes (N) se fait manuellement et sans l'aide de l'algorithme, mais il est important pour l'utilisateur que ce processus utilise l'algorithme ; Aider l'utilisateur à comprendre la nature des données, et obtenir quelques aspirations.

• L'algorithme sélectionne initialement les points centraux (X) au hasard, ce qui conduit à des incohérences et à des résultats irrécupérables ; Comme nous obtiendrons des résultats différents (centres de cluster) chaque fois que nous exécuterons l'algorithme sur les mêmes données [17]

b) Algorithme de clustering à décalage moyen

L'algorithme Mean-Shift est basé sur la méthode de la fenêtre glissante qui essaie de trouver des régions denses dans les points de données. Il s'agit également d'un algorithme centralisé, ce qui signifie que le but est de localiser les points centraux de chaque groupe en mettant à jour la fenêtre glissante, de mettre à jour les points en leur sein et de prendre leur moyenne comme centre du groupe. Et si nous avons plusieurs fenêtres de filtre au centre d'un seul groupe, ces fenêtres de filtre sont filtrées en post-traitement pour quasiment éliminer les doublons, conserver une fenêtre par groupe et déterminer le point central au moyen de la fenêtre[18]

c) Algorithme DBSCAN (Density-based spatial clustering of applications with noise)

L'algorithme DBSCAN est basé sur la densité comme l'algorithme Mean-Shift, mais avec deux fonctionnalités supplémentaires notables. [19]

L'algorithme DBSCAN offre des avantages concurrentiels par rapport aux autres algorithmes de clustering. Premièrement, il n'exige pas de l'utilisateur qu'il pré-détermine le nombre de groupes (N). Il prend également en compte les valeurs aberrantes, contrairement aux deux

algorithmes précédents, où la moyenne est affectée par les valeurs aberrantes même si elles sont très différentes de l'ensemble. L'algorithme se caractérise également par sa capacité à déterminer la forme et la taille des groupes d'une très bonne manière. [19]

Le principal inconvénient de l'algorithme DBSCAN est qu'il ne fonctionne pas bien lorsque les clusters sont de densité variable. En effet, lorsque nous définissons la plage ou le seuil de la distance epsilon ε et des points adjacents minimum (minPoint) nous trouverons des performances différentes en choisissant les points adjacents d'un groupe à l'autre lorsque la densité varie entre eux. Cette négativité se produit également dans les données à haute dimensionnalité, où la détermination de la distance epsilon ε devient un défi majeur [19]

1.3.3 Régression vs classification

L'identifier de type de sortie attendu de notre programme nous aideons à choisir l'algorithme d'apprentissage automatique. Cela répond également aux questions suivantes:

est-ce une valeur continue (un nombre)? Est-ce une valeur discrète (une catégorie)? Le premier cas s'appelle la régression et le second est la classification (Figure1.5) [9].

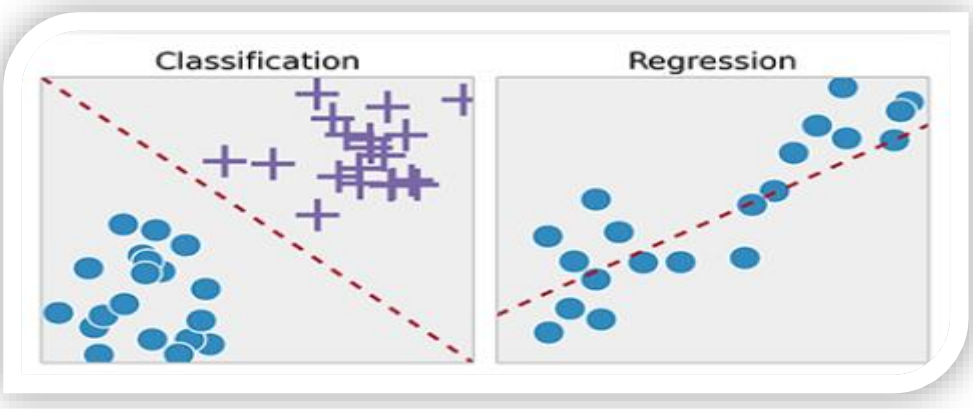


Figure 1.3: Illustration de la différence entre régression linéaire et classification linéaire[10].

1.4 Les données d'apprentissage, de validation et de test

Dans l'apprentissage automatique , une tâche courante est l'étude et la construction d' algorithmes qui peuvent apprendre et faire des prédictions sur les données . Ces algorithmes fonctionnent en faisant des prédictions ou des décisions basées sur les données, en construisant un modèle mathématique à partir des données d'entrée Les données utilisées pour construire le modèle final proviennent généralement de plusieurs ensembles de données . En particulier, trois ensembles de données sont couramment utilisés à différentes étapes de la création du modèle. Ensembles de formation, de validation et de test.[20]

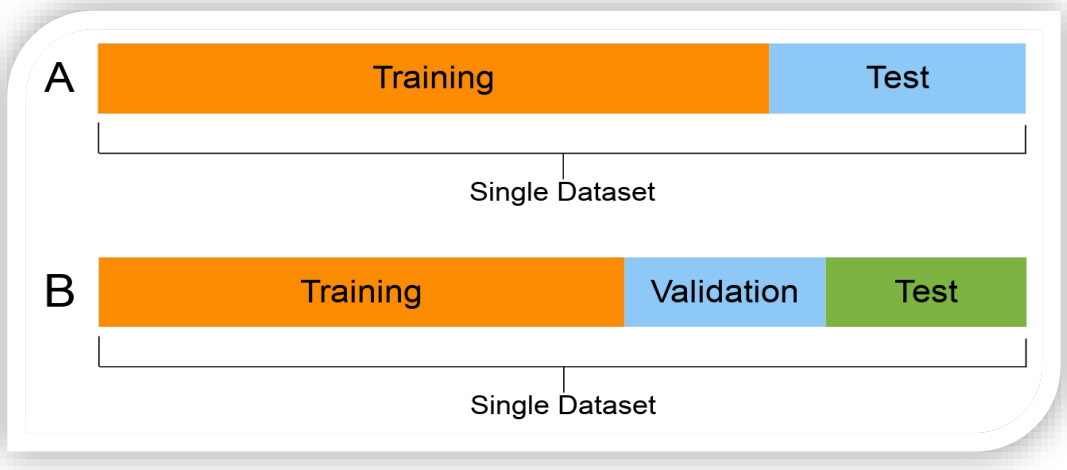


Figure 1.10 : Exemples de méthodes de segmentation d'ensembles de données. [20]

A Seuls un ensemble de données et un ensemble de test sont utilisés pour tester le modèle entraîné, Pour B, l'ensemble de données d'apprentissage utilise l'ensemble de validation pour tester le modèle entraîné, et l'ensemble de test évalue le modèle final.

1.4.1 Un ensemble de données d'apprentissage

Un ensemble de données d'apprentissage est un ensemble de données d'exemples utilisé pendant le processus d'apprentissage et utilisé pour ajuster des paramètres (tels que des poids), par exemple, un classificateur. [20]

Pour les tâches de classification, l'algorithme d'apprentissage supervisé examine l'ensemble de données d'apprentissage pour sélectionner ou apprendre les combinaisons optimales de variables qui généreront un bon modèle prédictif L'objectif est de produire un modèle entraîné (ajustement) qui se généralise bien sur de nouvelles données inconnues. le modèle ajusté est évalué à l'aide de « nouveaux » exemples provenant d'ensembles de données en attente (ensembles de données de validation et de test) pour estimer la précision du modèle dans la classification de nouvelles données. Pour réduire le risque de problèmes tels que le sur apprentissage, les exemples ne doivent pas être utilisés dans les ensembles de données de validation et de test pour entraîner le modèle. [20]

1.4.2 Ensemble de données de vérification

Un ensemble de données de validation est un ensemble de données d'exemples utilisés pour définir les hyper paramètres (c'est-à-dire la structure) d'un classificateur. Il est aussi parfois appelé l'ensemble de développement d'exemples d'hyper paramètres de réseaux de neurones artificiels, le nombre d'unités cachées dans chaque couche doit suivre la même distribution de probabilité que l'ensemble de données d'apprentissage. [20]

1.4.3 Ensemble de données de test

Une donnée de test est un ensemble de données indépendant des données d'apprentissage, mais qui suit la même distribution de probabilité que les données d'apprentissage. Si l'ajustement du modèle pour l'ensemble de données d'apprentissage s'adapte également bien à l'ensemble de données de test, un sur ajustement minimal s'est produit (voir la figure ci-dessous). [20]

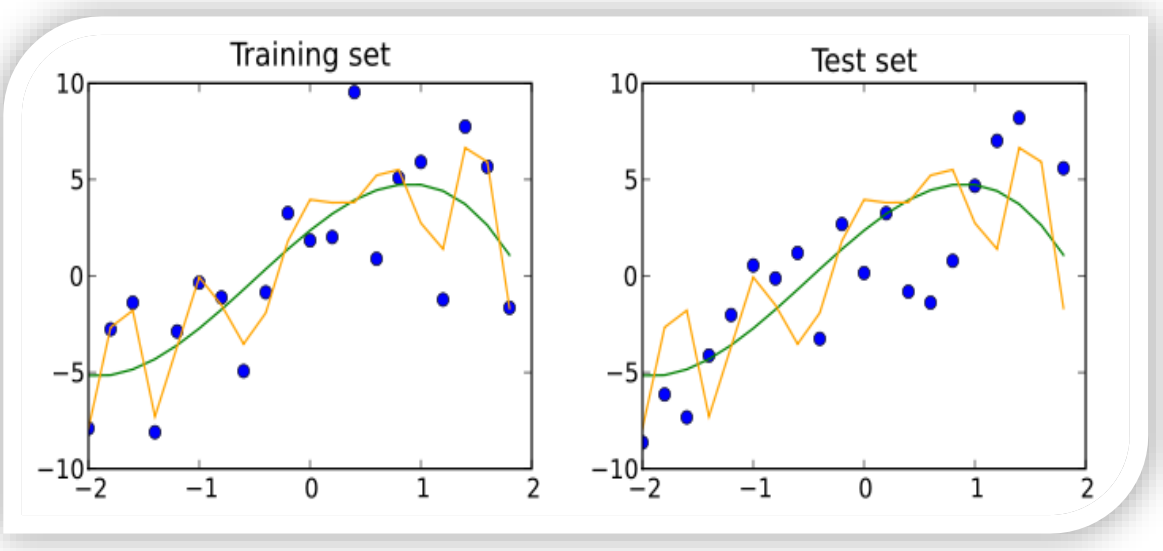
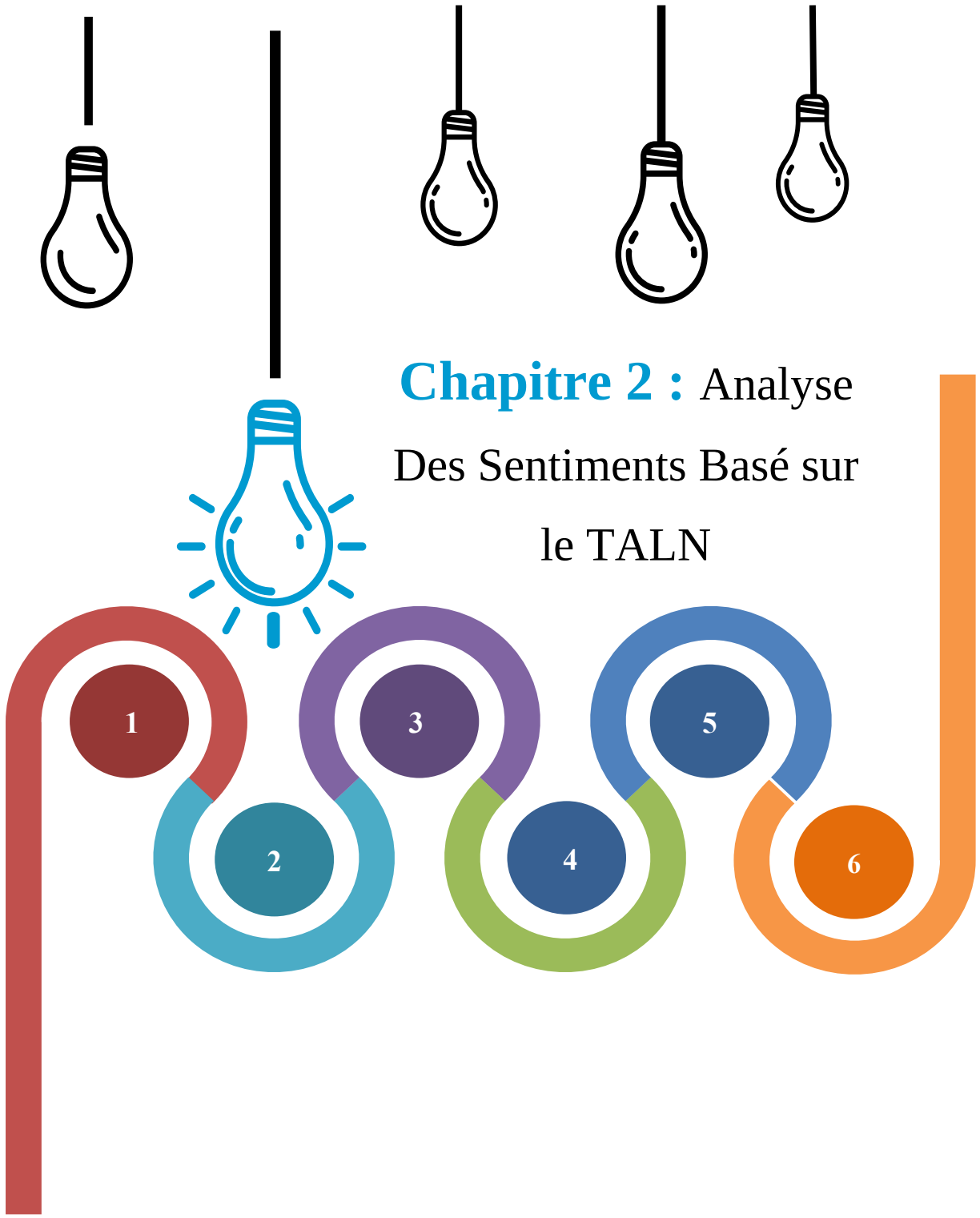


Figure 1.11 : Traininge set et Test set [20]**.**

L'ensemble d'apprentissage (à gauche) et l'ensemble de test (à droite) de la même population sont représentés par des points bleus[20].

1.5 Conclusion

Dans ce chapitre intitulé «Généralité sur l'apprentissage automatique », nous avons examiné les différentes définitions dont nous avons besoin dans notre mémoire . Où nous avons donné une définition complète de l'intelligence artificielle, son importance et son fonctionnement, puis nous avons défini l'apprentissage automatique et ses types, y compris l'apprentissage supervisé et l'apprentissage non supervisé et mentionné quelques algorithmes pour l'apprentissage automatique, enfin nous avons donné une explication de l'apprentissage, de la validation et tests de packages.



Chapitre 2 : Analyse Des Sentiments Basé sur le TALN

2. Analyse des Sentiments basé sur le TALN (la fouille de textes)

Intorduction

Le développement des reseau sociaux, des blogs pousse les utilisateurs à laisser de plus en plus d'informations en libre accès sur le web. Une partie de ces informations décrit des sentiments: elles permettent de développer des modèles d'analyse des sentiments et de faire des sondages dans divers domaines en récupérant simplement ces données textuelles.

L'analyse de sentiment demande bien plus de compréhension de la langue que l'analyse de texte et la classification par sujet. En effet, si les algorithmes les plus simples considèrent uniquement les statistiques de fréquence d'apparition des mots, cela se révèle en général insuffisant pour définir l'opinion dominante dans un document, surtout lorsque le contenu est court comme des messages ou des commentaires.

Donc ce chapitre présente une brève revue de littérature sur la technologie des réseaux sociaux numériques et leurs spécificités ,et nous expliquerons les concepts de base de l'analyse des sentiments ou se qu'on appel classification des opinions basé sur l' un des meilleurs techniques utilisés pour la catégorisation des textes :l'approche TALN et en fin les Outils d'analyse des sentiments .

2.1 les réseau sociaux

Dans cette section, nous présenter quelques concepts sur les réseau sociaux et leur impactes sur le plan social.

2.1.1 Définition

Le concept de « social » a été inventé en 1954 par un anthropologue du nom de"John A. Barnes" Le principe de réseau se définit par deux éléments : les contacts et les liaisons entre les contacts, Plus nous avons de contacts plus notre réseau est important et donc plus nous sommes utiles (la notion d'utilité ici se résume à la capacité à transmettre des informations). [21]

D'autre part, les réseaux sociaux sont considérés comme des services web qui permettent aux individus de construire un profil public ou semi -public liée avec une liste d'autres profils qui nécessite une confirmation bidirectionnelle pour l'amitié, mais parfois unidirectionnels comme fans ou abonnés. [22]

2.1.2 historique

À la fin des années 1990, au fur et à mesure que s'est accrue la popularité de l'Internet à large bande, des sites Web permettant de créer et de télécharger du contenu ont commencé à apparaître. Le premier site de réseautage social (SixDegrees.com) date de 1997. À partir de 2002, les sites de réseautage social se sont multipliés. Certains – comme Friendster – ont créé un véritable engouement initial, qui s'est cependant vite dissipé. D'autres se sont trouvé des créneaux, comme MySpace, qui a séduit les adolescents friands de musique.

À la fin des années 2000, les médias sociaux étaient largement entrés dans les mœurs, et certains services ont vu le nombre de leurs adeptes grossir de façon exponentielle. Ainsi, en novembre 2012, Facebook annonçait qu'il comptait un milliard d'usagers dans le monde. En juillet 2012, on estimait à 517 millions le nombre d'utilisateurs Twitter.

Plusieurs facteurs ont contribué à la croissance rapide de l'utilisation des médias sociaux : des facteurs technologiques, comme l'accès croissant à la large bande, l'amélioration des outils logiciels et la puissance toujours plus grande des ordinateurs et des appareils mobiles, des facteurs sociaux, comme l'engouement instantané de la part des groupes d'âge plus jeunes, des facteurs économiques, comme le prix de plus en plus abordable des ordinateurs et des logiciels et l'intérêt commercial croissant que suscitent les sites de médias sociaux [23]

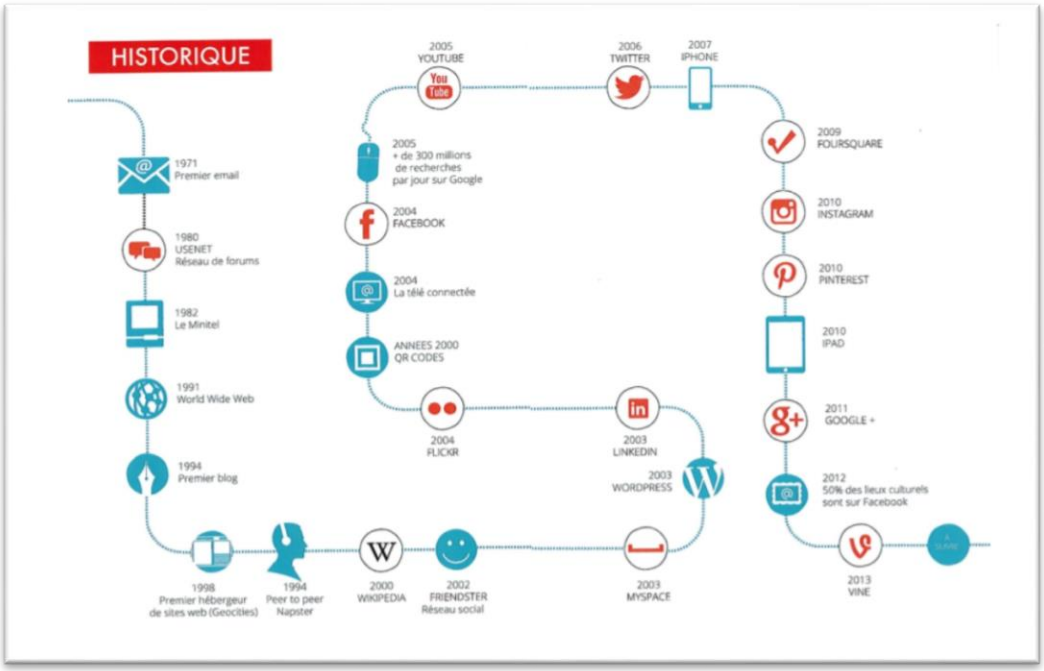


Figure 2.1 : Historique des réseaux sociaux **[24]**

2.1.3 Les principales catégories de réseaux sociaux et leurs usages

les réseaux sociaux ayant des catégories principales comme suit :

A-Les médias sociaux à usage professionnel : Vidéo, Linkedin, Xing… déstinés à la mise en relation professionnelle mais également Digg, Reddit… qui permettent le partage de liens. Sans oublier bien sûr Slideshare ou Scribd pour le partage de documents

B-Les réseaux sociaux de partage de contenus : Ils intègrent les sites de partage de liens et de documents cités ci-dessus mais également les sites de partage de vidéos comme Youtube et Dailymotion ainsi que ceux de partage de photographies comme Picasa et Flickr. Il est a noter que si les sites de partage de vidéos et de photographies sont largement utilisés par les entreprises, par les professionnels, ils ont également conquis la sphère des loisi rs et touchent par conséquent le grand public

C-les médias sociaux de « loisirs » : Ce sont les plate-formes sociales comme l'incontournable facebook, mais également Copains d'avant. L'usage concerne principalement « l'individu » mais également le « professionnel » qui peut se cacher derrière l'individu.

Ainsi, Facebook propose la création de pages destinées aux entreprises, organisations, associations… Dans cette catégorie s'intégrent également les réseaux sociaux de joueurs comme Habbo ou Zynga utilisés par des profils personnels

D-les médias sociaux destinés au partage d'expression : Ils intègrent les plate-formes sociales évoquées dans le paragraphe précédent mais également les forums, les blogs, les microblogs comme twitter, les wikis…[25]

2.1.4 Les principaux réseaux sociaux plus utilisés

On pose toujours la question : quel est le réseau social le plus utilisé à l'échelle mondial ? pour répondre à cette question on doit disposer un outil informatique qui classe ces réseaux selon des critères prédéfinies, Alexa à l'un de ces outils.

Alexa est une société Internet qui analyse le trafic des autres sites avec des outils et des algorithmes qui lui sont propres. Une fois ces analyses faites, elle établit un classement des sites les uns par rapport aux autres. Ce classement est plus ou moins contesté, cependant il est un indicateur utile. Il est possible d'améliorer le classement Alexa de son site, moyennant quelques

transformations. Il faut se polariser sur trois aspects-clés : le trafic, l'optimisation et la visibilité de votre site Selon les réseaux sociaux classifiés selon Heure quotidienne sur le site et les Pages vues quotidiennes par visiteur et d'après du trafic de la recherche et Nombre total de sites liés . [26]

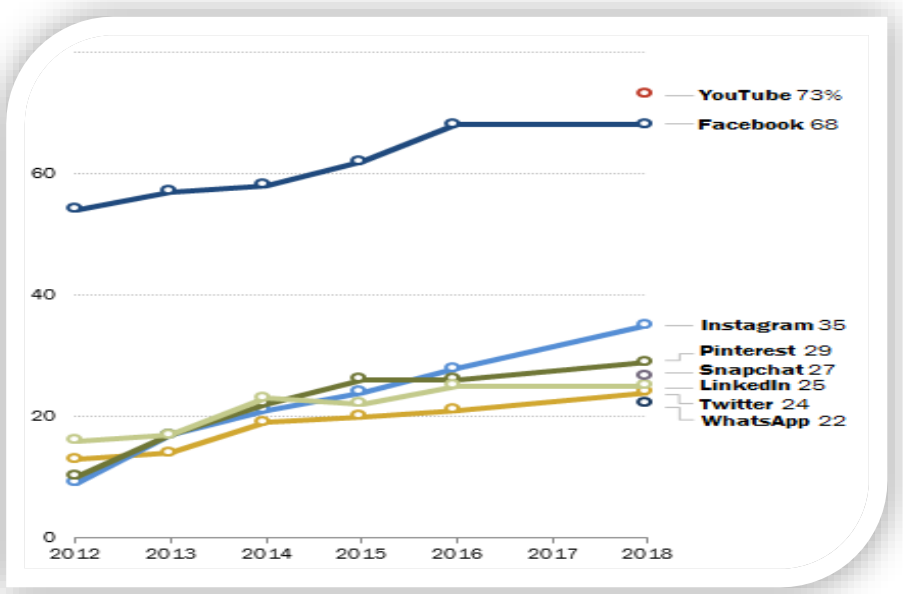


Figure 2.2 : Mise à jour des médias sociaux 2018 [26]

2.1.5 L'autre coté des réseaux sociaux

Le bilan des impacts des réseaux sociaux sur la société est plutôt partagé, Nous nous concentrerons sur les aspects négatifs qui affectent les utilisateurs :

- Les réseau sont addictif : les réseaux sociaux peuvent être très addictif car Il y a des gens qui ont besoin d'être sur ces sites toutes les heures ou toutes les minutes !
- La vie privée : Informer vos « amis » de certaines facettes de votre vie privée peut entraîner des problèmes comme la cyber intimidation , C'est à ce moment que les utilisateurs disent des choses méchantes et blessantes et peut être extrême sur d'autres utilisateurs.
- La dérive vers le terrorisme : L'inconscience mène aux une déviation religieuse et idéologique ou à une déviation psychologique là où se tuer devient un jeu divertissant.
- Les réseaux sociaux ont des inconvénients mais s'ils sont utilisés correctement et avec modération, ils sont utiles et amusants et permettent de rapprocher les gens au niveau privé ou professionnel .

2.2 Analyse des Sentiments basé sur le TALN (la fouille de textes)

Dans cette section, nous avons fait un survole sur les principaux fondements, techniques de classification des sentiments et d'opinion mining.

2.2.1 Traitement Automatique des Langues Naturelles TALN :

2.2.1.1 définition

La TALN est un moyen pour les ordinateurs d'analyser, de comprendre et de dériver le sens du langage humain d'une manière intelligente et utile. En utilisant le langage TALN, les développeurs peuvent organiser et structurer les connaissances pour exécuter des tâches telles que la synthèse automatique, la reconnaissance d'entités nommées, l'extraction de relations, l'analyse des sentiments, la reconnaissance de la parole et la segmentation des rubriques.
"En dehors des opérations courantes de traitement de texte qui traitent le texte comme une simple séquence de symboles, la TALN considère la structure hiérarchique du langage : plusieurs mots forment une phrase, plusieurs phrases forment une phrase et, finalement, des phrases qui transmettent des idées." expert de Meltwater Group, a déclaré dans Comment le traitement du langage naturel aide à découvrir le sentiment des médias sociaux. "En analysant le langage pour sa signification, les systèmes TALN ont depuis longtemps rempli des rôles utiles, tels que la correction de la grammaire, la conversion de la parole en texte et la traduction automatique entre les langues." .[27]

La TALN est utilisée pour analyser le texte, permettant aux machines de comprendre comment l'humain parle. Cette interaction homme-ordinateur permet des applications réelles telles que la synthèse automatique de texte, l'analyse de sentiment, l'extraction de sujet, la reconnaissance d'entité nommée, l'étiquetage de parties de discours, l'e§xtraction de relations, le stemming, etc. La TALN est couramment utilisée pour l'exploration de texte, la traduction automatique et la réponse automatique aux questions. .[27]

Les algorithmes TALN sont généralement basés sur des algorithmes d'apprentissage automatique. Au lieu de coder manuellement de grands ensembles de règles, la TALN peut s'appuyer sur l'apprentissage automatique pour apprendre automatiquement ces règles en analysant un ensemble d'exemples (un grand corpus, comme un livre, jusqu'à une collection de phrases) et en faisant une inférence statique. . En général, plus le nombre de données analysées est élevé, plus le modèle sera précis.[27]

2.2.1.2 Champs de recherche et applications de TALN

Le domaine du traitement des langues naturelles comprend un grand nombre de disciplines de recherche variés en termes d'objectif ou des méthodes de traitement, ainsi que la forme d'information. Ce domaine est divisé en trois axes principaux : Sémantique, Extraction d'informations et Syntaxe, dont chacun comprend nombreux domaines, nous mentionnons certains dans la figure suivante :[28]

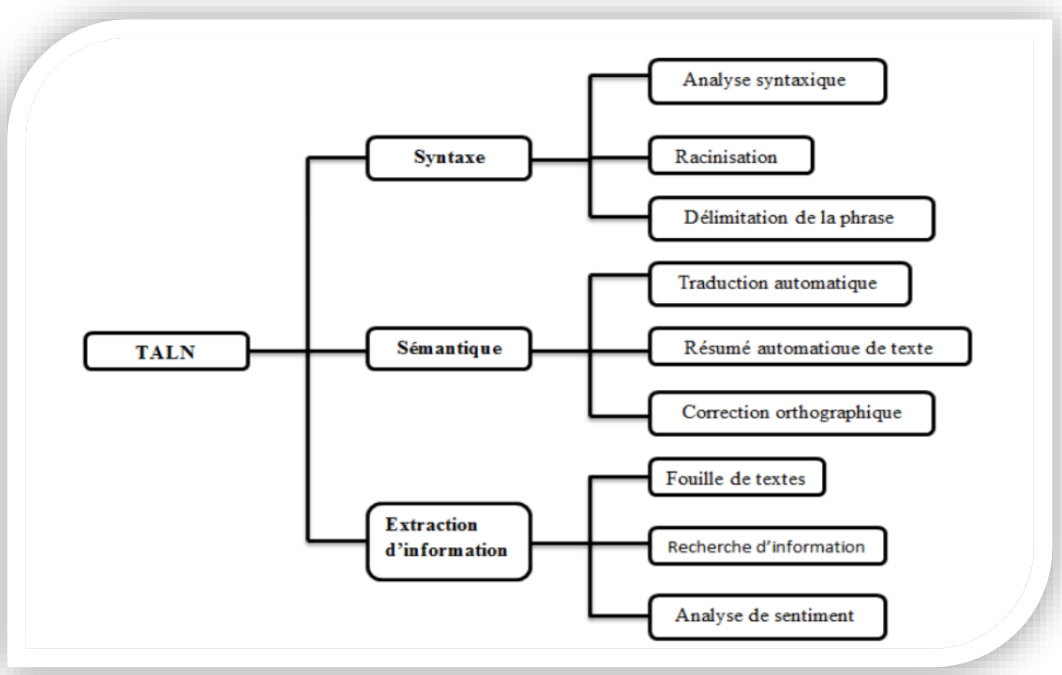


Figure 2.3 : Les champs de recherche et applications de TALN :[28]

Parmi les processus couramment utilisés dans le domaine du traitement du langues et en particulier lors du travail sur les textes, nous trouvons plongement de mots, que nous expliquerons dans la section suivante :

2.2.2 Plongement de mots (en :Word Embedding)

Plongement de mots ou Word Embeddings est l'une des méthodes qui ont une grande contribution dans le domaine du traitement automatique du langage naturel. Cette technique cherche à représenter chaque mot du dictionnaire avec un vecteur de nombres réels.

Cette représentation a la particularité que les mots qui apparaissent dans des contextes similaires ont des vecteurs convergents relativement proches.

Il existe différents types de plongement de mots, Nous mentionnons trois des tech-niques les plus utilisées :

2.2.2.1 Fréquence du terme/Fréquence inverse de document (en:TF-IDF)

Cette méthode est basée sur la représentation de chaque mot en lui donnant un poids spécifique. Son récit est lié à d'occurrences du mot dans le document ainsi qu'à son occurrence dans le corpus. Il s'appuie également sur la réduction de poids des mots courants qui se produisent dans presque tous les documents et donnent plus d'importance aux mots qui apparaissent dans un sous-ensemble de documents.

Le poids de cette méthode est calculé en deux parties : La première partie est TF, qui dépend de la fréquence du mot dans le document. Elle est calculée de plusieurs façons :

Schéma de pondération	Formule du TF
binaire	0,1
Fréquence brute	$f_{t,d}$
Normalisation logarithmique	$\log(1 + f_{t,d})$
normalisation « 0.5 » par le max	$0{,}5 + 0{,}5 \cdot \frac{f_{t,d}}{\max\{t' \in d\} f_{t',d}}$
Normalisation par le max	$K + (1 - K) \frac{F_{t,d}}{\max\{t' \in d\} f_{t',d}}$

Tableau 2.1 : Les méthodes de calculer le TF

La deuxième partie IDF, qui représente la fréquence du mot dans le corpus Calculé comme suit : **idfi=ln[(N+1)/dfi]+1**

Dj : nombre de documents où le terme Ti apparaît. Alors, le poids est :

tf-idfi=tfi×idfi

Le principal problème avec cette méthode est qu'elle traite le document comme un sac de mots, ce qui signifie que les mots sont indépendants .

2.2.2.2 Mot-à-Vecteur (en : Word2Vec)

Mot-à-Vecteur ou Word2vec fournit un vecteur pour chaque jeton / mot et ces vecteurs codent la si-gnification du mot. la signification des vecteurs est compréhensible / interprétable en les comparant à d'autres vecteurs et à diverses équations intéressantes (par exemple roi-hommes + femmes = reine, ce qui prouve à quel point ces vecteurs détiennent la séman-tique des mots).

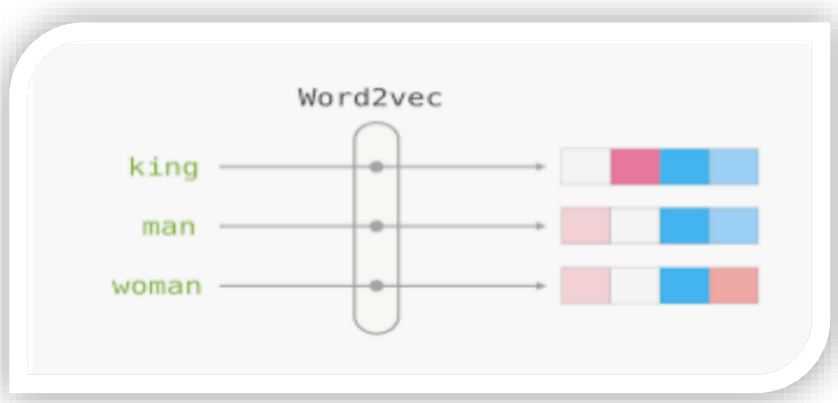


Figure 2.4 : Fonctionnement de word2vec

2.2.2.3 Le Modèle bidirectionel : « BERT »

BERT est l'acronyme de « Bidirectionnal encoder representations from transformers ». C'est un modèle de langage dit pré-entraîner qui repose sur des réseaux neuronaux. Il a été développé par les équipes Google qui planchent depuis des années sur le traitement de langage naturel (TALN).

C'est un modèle TALN dit bidirectionnel parce qu'il lit un contenu rédactionnel dans sa globalité et dans les deux sens. Et grâce à une architecture faite par Google, le Trans-formateur, il apprend les relations contextuelles entre les mots. BERT peut donc mieux interpréter les différents types de demande et apporter une réponse très pertinente aux

requêtes écrites ou orales.

BERT utilise Transformer,une mécanisme d'attention qui apprend les relations contex-tuelles entre les mots (ou sous-mots) dans un texte. Sous sa forme vanille, Transformateur comprend deux mécanismes distincts - un encodeur qui lit l'entrée de texte et un décodeur qui produit une prédiction pour la tâche. Le but du BERT étant de générer un modèle de langage,

seul le mécanisme de l'encodeur est nécessaire. Le fonctionnement détaillé de Transformer est décrit dans un document de Google.

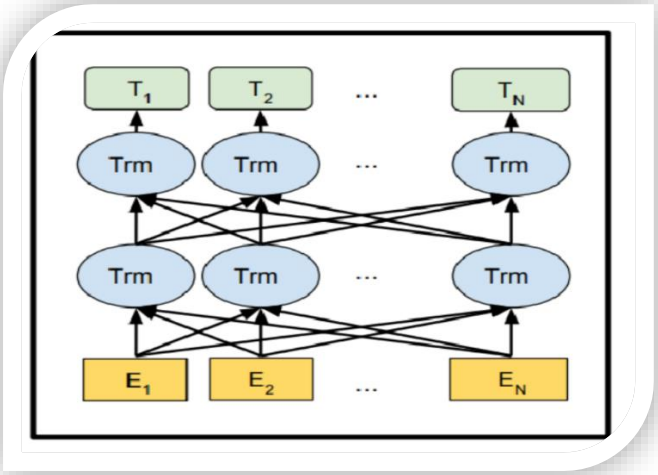


Figure 2.5 : Architecture de modèle BERT

Contrairement aux modèles directionnels, qui lisent l'entrée de texte séquentiellement (de gauche à droite ou de droite à gauche), le codeur Transformateur lit la séquence entière de mots à la fois. Par conséquent, il est considéré comme bidirectionnel, mais il serait plus exact de dire qu'il n'est pas directionnel. Cette caractéristique permet au modèle d'apprendre le contexte d'un mot en se basant sur l'ensemble de son environnement (gauche et droite du mot).

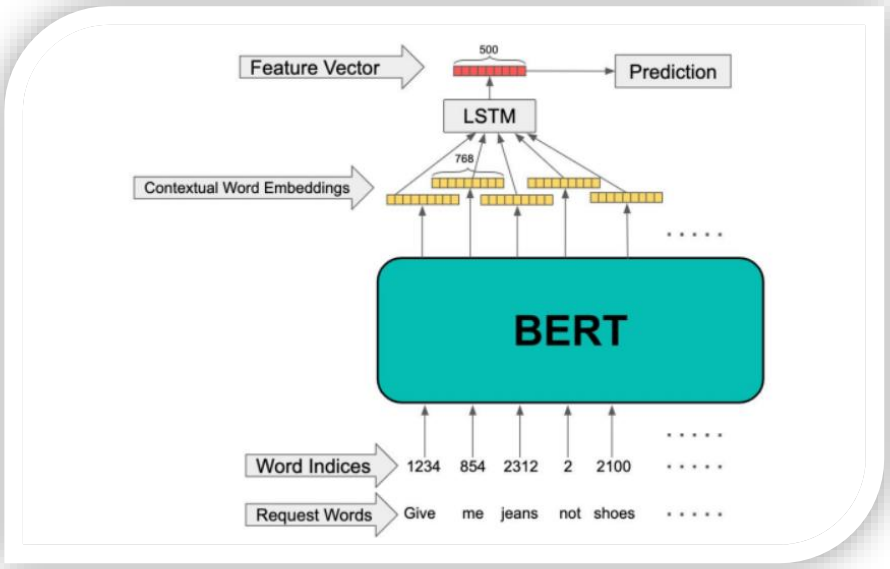


Figure 2.6 : Mécanisme de classification a l'aide de BERT

2.2.3 Analyse des sentiments

Dans cette section, nous expliquons le domaine d'analyse des sentiments qui est considéré comme un axe de TALN.

2.2.3.1 Sentiment

- **Définition :**

Il est défini comme un élément émotionnel qui implique les fonctions cognitives du corps et un moyen d'appréciation. Le sentiment est à l'origine de la connaissance directe ou d'une simple impression. Il indique la perception de l'état physiologique du moment [29].

Bien qu'il existe de nombreux types de sensations, chacune a une façon différente de l'exprimer, mais elle est essentiellement divisée en 3 catégories principales, les sentiments positifs et négatifs et neutre. C'est de cela que dépend l'analyse des sentiments, où vous classez l'opinion ou le sentiment à ces trois catégories.

2.2.3.2 Définition de l'analyse des sentiments

L'analyse des sentiments est une démarche principalement basée sur le text mining et l'analyse sémantique qui permet de déterminer la « position » des individus étudiés à l'égard d'une marque ou d'un événement. L'analyse des sentiments peut cependant également reposer sur d'autres éléments que les données textuelles. Elle peut par exemple être basée sur l'usage des émoticônes (emojis), sur les « émotions » facebook, sur l'analyse de la voix ou même sur le facial coding / decoding. Lorsqu'elle est basée sur la fouille textuelle, l'analyse des sentiments peut porter sur des verbatims issus des réseaux sociaux, des avis, forums, de données d'études qualitatives ou quantitatives, etc. Les sentiments sont généralement classés en trois types : négatifs, neutres et positifs. [30]

2.2.3.3 Taches de l'analyse des sentiments

De nos jours, l'analyse des sentiments a gagné encore plus de valeur avec l'avènement des réseaux sociaux. Leur grande diffusion et leur rôle dans la société moderne représentent l'une des nouveautés les plus intéressantes de ces dernières années, captant l'intérêt des chercheurs, des journalistes, des entreprises et des gouvernements. L'interconnexion dense qui surgit souvent parmi les utilisateurs actifs génère un espace de discussion capable de motiver et

d'impliquer les individus d'un espace plus large, reliant les personnes avec des objectifs communs et facilitant diverses formes d'action collective.

Les réseaux sociaux créent donc une révolution numérique, permettant l'expression et la diffusion des émotions et des opinions à travers le réseau, ouvrant une fenêtre sur les mondes des autres et fouillant dans leur vie. Les données d'opinion sur le net, si elles sont correctement collectées et analysées, permettent non seulement de comprendre et d'expliquer de nombreux phénomènes sociaux complexes, mais aussi de les prédire.

Les progrès technologiques actuels permettent de stocker et de récupérer efficacement une énorme quantité de données, l'accent est désormais mis sur les méthodes d'extraction de l'information et de création de connaissances à partir de sources brutes. En effet, les réseaux sociaux représentent un nouveau secteur stimulant dans le contexte de big data : les expressions en langage naturel des gens peuvent être facilement rapportées par des textes de messages courts, créant rapidement un contenu unique de dimensions énormes qui doit être analysé de manière efficace pour créer des connaissances exploitables pour les processus de prise de décision.

Cependant, l'analyse des sentiments est souvent utilisée incorrectement lorsqu'on se réfère à la classification de polarité, qui est plutôt une sous-tâche visant à détecter la polarité des textes de sentiments ; positifs, négatifs ou neutres. Bien qu'une opinion puisse aussi avoir une polarité neutre (par exemple, "Je ne sais pas si j'ai aimé le film ou pas. Je devrais le regarder tranquillement."). La plupart des travaux d'analyse des sentiments ne supposent généralement que des sentiments positifs et négatifs pour des raisons de simplicité.

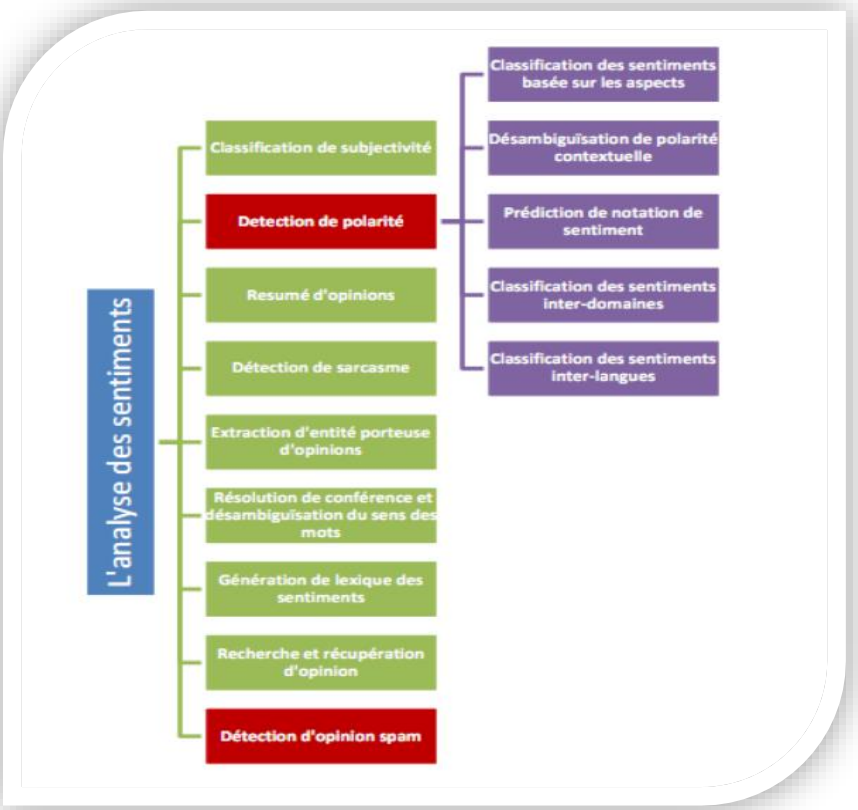


Figure 2.7 : Tâches d'analyse des sentiments [31].

Selon le domaine d'application, plusieurs noms sont utilisés pour l'analyse des sentiments (par exemple : fouille d'opinions, extraction d'opinion, fouille des sentiments, analyse de subjectivité, analyse de l'affect, analyse des émotions et analyse des avis). Une taxonomie des tâches d'analyse des sentiments les plus populaires est présentée par la Figure 2-7.

2.2.3.4 Les niveaux d'analyse des sentiments

Comme mentionné précédemment, le but de l'analyse des sentiments est de " définir des outils automatiques capables d'extraire des informations subjectives à partir des textes en langage naturel ". Le premier choix lorsqu'on applique l'analyse des sentiments est de définir ce que signifie le texte (c'est-à-dire l'objet analysé) dans le cas d'étude considéré.

En général, l'analyse des sentiments dans les réseaux sociaux peut être étudiée principalement à trois niveaux (représenté graphiquement dans la Figure 2.8) :

- **Niveau de texte** : : détermine l'opinion générale de l'ensemble du document. Cette analyse fonctionne bien pour des documents qui présentent un point de vue précis, mais moins pour des comparaisons car elle ne fera pas la différence entre les sujets abordés.

- **Niveau de phrase** : : détermine l'opinion générale d'une phrase (positive, négative ou neutre). Cette analyse peut donner une mesure de la "neutralité" d'un texte par exemple pour analyser des entrées de Wikipédia. Les méthodes utilisées sont celle de l'analyse de subjectivité.
- **Niveau d'entité et d'aspect** : (appelé en anglais Feature level) : au lieu de déterminer les entités à analyser en fonction de critère structuraux (phrase, paragraphe, document) ces méthodes se basent sur l'analyse de corrélation entre l'opinion émise et la cible de cette opinion. Par exemple, la phrase "Le sujet du cours me passionne mais le professeur est ennuyeux." présente deux sentiments sur l'entité "cours" : le sujet qui est perçu comme positif et le professeur, qui est perçu comme négatif. Ce niveau d'analyse permet de différencier les aspects qui sont aimé ou non par les auteurs des textes et ainsi permet plus facilement de déterminer des remédiations possibles. En revanche il est très difficileà mettre en place car ce type d'analyse est extrêmement complexe.[32]

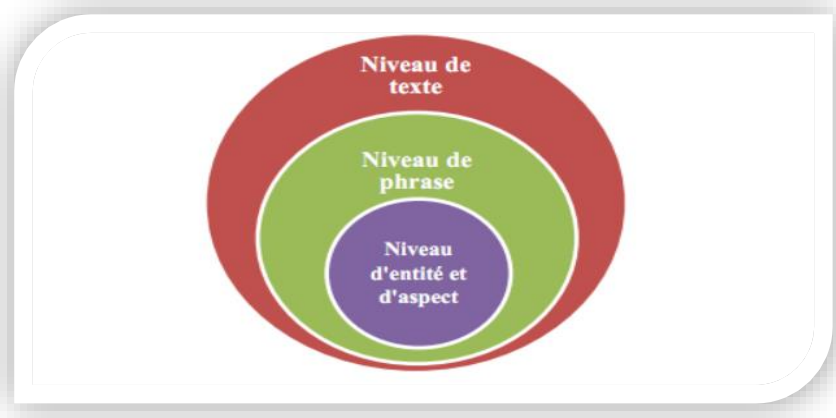


Figure 2.8 : Les différents niveaux d'analyse.[33]

2.2.3.5 Méthodes d'analyse des réseaux sociaux

On distingue alors deux grandes familles : traditionnelle et fouille de données

A- Méthodes traditionnelles

A-1 La centralité : Les acteurs importants sont ceux qui sont liés et impliqués avec les autres acteurs d'une façon extensive. Après avoir modélisé les contacts par des liens dans un réseau, nous pouvons définir un acteur central comme étant celui qui est impliqué dans plusieurs liens.

Il existe différents types de liens entre les noeuds (acteur), ce qui a permet de générer plusieurs types de centralités dont :

A-2 Degré de Centralité : Soit n le nombre total de noeud (acteurs) dans un réseau (graphe). Le degré de centralité d'un acteur i noté CD(i) est le degré du noeud acteur (le nombre d'arêtes) noté d(i) normalisé par le degré maximal n-1

$$CD(i) = \frac{d(i)}{(n-2)} \text{ ……..1}$$

A-3 La centralité de proximité(closenesscentrality) indique si le sommet est situé à proximité de l'ensemble des sommets du graphe et s'il peut rapidement interagir avec Ces sommets. Il s'écrit formellement :

$$C_c(V) = \frac{1}{\sum u \in V \setminus \{V\} dG(u,v)} \text{ ……..2}$$

avec dG (u,v) la distance entre les sommets u et v.[34]

A-4 La centralité d'intermédiarité (en :betweeness centrality) : est un des concepts les plus importants. Il mesure l'utilité du sommet dans la transmission de l'information au sein duréseau. Le sommet joue un rôle central si beaucoup de plus courts chemins entre deux sommets doivent emprunter ce sommet. Elle s'écrit :

$$C_B(v) = \sum_{i \neq j \neq v} i,j \ \frac{\delta\, i,j(v)}{\delta\, ij} \text{ ……..3}$$

Avec σij(v) le nombre de chemins entre i et j qui passent par v.

A-5 Le prestige: La notion de prestige est une autre manière de mesurer l 'importance d'un noeud.

Un noeud prestigieux est un noeud à lequel un grand nombre d'autres noeud se lient- il reçoit un grand nombre de liens entrants. Nous distinguons les mesures suivantes :

A-6 Le degré de prestige : avec dI(i) le degré entrant du noeud i, n le nombre noeuds dans le réseau, le degré de prestige est donné par la relation :

$$PD(i) = \frac{dI(i)}{(n-1)}$$

A-7 Le prestige de tri : Les mesures proposées jusqu'au là sont fondées sur les liens entrants et sortants d'un acteur donné. La mesure du prestige du tri considère la éputation et I importance des acteurs choisissant l'acteur i. Ce type d'algorithme est utilisé pour trier les résultats de recherche comme Google (algorithme Rank Page). [35]

B- **Fouille de données dans les réseaux sociaux:**

B-1 définition :

La fouille de données ou le data mininig est en fait un terme générique englobant toute une fami lle d'outils facilitant l'exploration et l'analyse des données contenues au sein d'une base décisionnelle de type Data Warehouse ou DataMart. Les techniques mises en action lors de l'utilisation de cet instrument d'analyse et de prospection sont particulièrement efficaces pour extraire des informations significatives depuis de grandes quantités de données.[36]

Donc la fouille de texte est un outil incroyablement riche et puissant de modéliser le réseau sous forme mathématique, il permet d'analyser pour tirer le maximum d'information mais aussi prédire en partie l'évolution future du réseau.

B-2. Text Mining :

Est l'exploration de texte et l'extraction de l'implicite, jusqu'alors inconnue et des informations otentiellement utiles provenant de quantités de ressources textuelles.

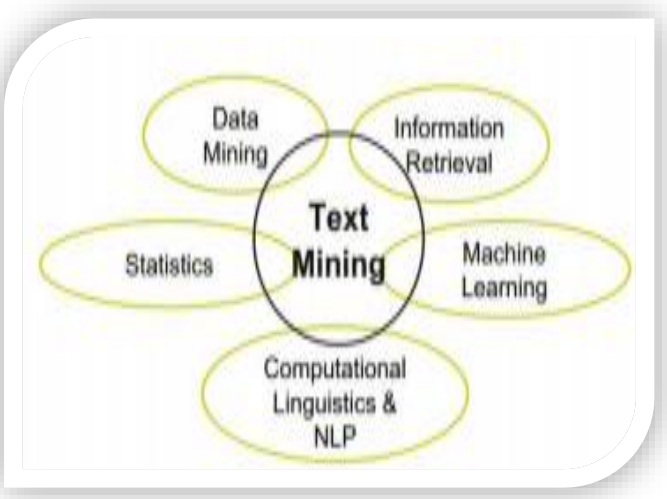


Figure 2.9 : Text Mining

B-3 Les applications d'exploration de texte : nous utilisons la fonction d'exploration de texte à plusieurs fins , notamment :

1. Classification des actualités ou des pages Web
2. Filtrage des e-mails et des actualités / détection du SPAM
3. Analyse des sentiments
4. Regroupement de documents ou de pages Web
5. Saisie automatique des termes de recherche
6. Extraction d'information

B-4 Le processus d'exploration de texte : L'exploration du texte comprend un ensemble d'opérations, comme :

1. Prétraitement du texte Syntaxique et / ou analyse sémantique .
2. Génération de fonctionnalités
    - Sac de mots, intégration de mots

B-5 .Sélection des fonctionnalités : est le processus de sélection du meilleur sous-ensemble parmi l'ensemble existant de fonctionnalités qui contribuent davantage à la prédiction de la sortie. Cela aide à améliorer les performances du modèle (c'est-à-dire supprime les fonctionnalités redondantes ou / et non pertinentes qui sont porteuses de bruit et diminuent la précision du modèle) .[37]

B-6 .Exploration de données : Il existe de nombreuses techniques d'exploration de données que nous peuvent utiliser pour transformer les données brutes en insights exploitables comme :

- Clustering
- Classification
- Association une analyse

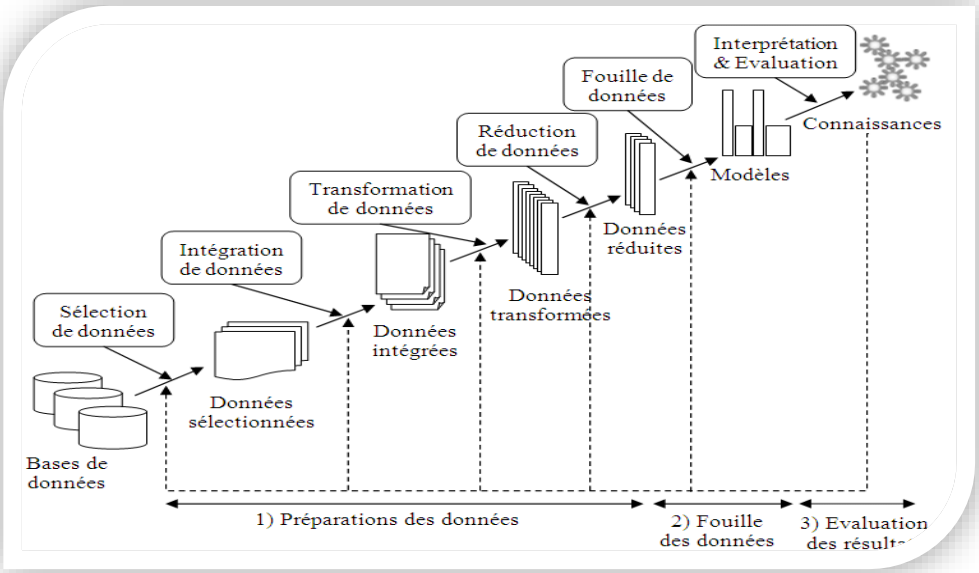


Figure 2.10 : Le processus d'exploration de texte [38]

B-7 Objectifs et applications

Parmi les principaux objectifs de la fouille de données textuelles, les suivants :

1. L'identification de thèmes, qui vise à regrouper des textes (ou parties de textes, ou ensembles de tags, etc.) en thèmes inconnus a priori. Cela implique en général l'utilisation de méthodes de classification automatique sur des données textuelles dont la représentation a été adaptée, par exemple par le passage à des représentations vectorielles.

2. Le classement de textes, qui vise à affecter des textes (ou parties de textes) à des catégories (classes) prédéfinies. Cela passe en général par l'application de modèles décisionnels, obtenus par apprentissage supervisé, à des représentations vectorielles de textes . La détection d'entités nommées, à représenter séparément, peut être utile.

3. L'extraction d'informations, qui vise à mettre en correspondance des textes avec des « schémas » d'interprétation prédéfinis. Un schéma d'interprétation regroupe plusieurs variables qui reçoivent des valeurs à partir du texte traité. Ces variables sont ensuite utilisées pour la fouille, conjointement avec d'autres variables (quantitatives, nominales) décrivant la même population.

B-8 Sac de Mots (en :Bag of words)

Définition :

Bag of words (Un sac de mots) est une technique de traitement du langage naturel de la modélisation de texte. En termes techniques, on peut dire qu'i l s'agit d'une méthode d'extraction de caractéristiques avec des données textuelles. Cette approche est un moyen simple et flexible d'extraire des fonctionnalités à partir de documents.

Pourquoi sac de mots ! L'un des plus gros problèmes avec le texte est qu'il est désordonné et non structuré, et les algorithmes d'apprentissage automatique préfèrent des entrées de longueur fixe structurées et bien définies.En utilisant la technique du sac de most, nous pouvons convertir des textes de longueur variable en une longueur fixe de plus, à un niveau beaucoup plus granulaire, les modèles d'apprentissage automatique fonctionnent avec des données numériques plutôt qu'avec des données textuelles. Donc, pour être plus précis, en utilisant la technique du sac de mots (BoW), nous convertissons un texte en son vecteur équivalent de nombres.

Comprendre le sac de mots avec un exemple :

Voyons un exemple de la façon dont la technique du sac de mots convertit le texte en vecteurs

Exemple (1) sans prétraitement :

Sentence 1: "Welcome to Great Learning, Now start learning"

Sentence 2: "Learning is a good practice"

Sentence 1	Sentence 2
Welcome	Learning
to	is
Great	a
Learning	good
,	practice
Now	
start	
learning	

Étape 1: Parcourez tous les mots du texte ci-dessus et faites une liste de tous les mots de notre vocabulaire modèle.

- Welcome
- To
- Great
- Learning
- ,
- Now
- start
- learning
- is
- a
- good
- practice

Notez que les mots ‘Learning’ et‘ learning’ ne sont pas les mêmes ici en raison de la différence entre leurs cas et sont donc répétés. Notez également qu'une virgule «,» est également prise dans la liste.

Parce que nous savons que le vocabulaire a 12 mots, nous pouvons utiliser une représentation de document de longueur fixe de 12, avec une position dans le vecteur pour noter chaque mot.

La méthode de notation que nous utilisons ici est de compter la présence de chaque mot et de marquer 0 pour l'absence. Cette méthode de notation est utilisée plus généralement.

La notation de la phrase 1 ressemblerait à ceci:

Word	Frequency
Welcome	1
to	1
Great	1
Learning	1
,	1
Now	1
start	1
learning	1
is	0
a	0
good	0
practice	0

Ecrire les fréquences ci-dessus dans le vecteur Phrase 1→ [1,1,1,1,1,1,1,1,0,0,0]

Maintenant pour la phrase 2, le score voudrait

Word	Frequency
Welcome	0
to	0
Great	0
Learning	1
,	0
Now	0
start	0
learning	0
is	1
a	1
good	1
practice	1

De même, écrire les fréquences ci-dessus sous forme vectorielle

Phrase 2→ [0,0,0,0,0,0,0,1,1,1,1,1]

Sentence	Welcome	to	Great	Learning	.	Now	start	learning	is	a	goo
Sentence1	1	1	1	1	1	1	1	1	0	0	0
Sentence2	0	0	0	0	0	0	0	1	1	1	1

Mais est-ce la meilleure façon d'exécuter un sac de mots. L'exemple ci-dessus n'était pas le meilleur exemple de la façon d'utiliser un sac de mots. Les mots Learning et learning, bien qu'ayant le même sens, sont pris deux fois. De plus, une virgule "," qui ne transmet aucune information est également incluse dans le vocabulaire.

Apportons quelques changements et voyons comment nous pouvons utiliser «sac de mots de manière plus efficace

Exemple (2) avec prétraitement :

Phrase 1 : "Welcome to Great Learning, Now start learning"

Phrase 2 : «Learning is a good practice»

Étape 1 : Convertissez les phrases ci-dessus en minuscules car la casse du mot ne contient aucune information.

Étape 2 : supprimez les caractères spéciaux et les mots vides du texte. Les mots vides sont les mots qui ne contiennent pas beaucoup d'informations sur le texte comme 'is', 'a', et bien d'autres.

Après avoir appliqué les étapes ci-dessus, les phrases sont remplacées par

Phrase 1: "welcome great learning now start learning"

Phrase 2: «learning good practice» Bien que les phrases ci-dessus n'aient pas beaucoup de sens, le maximum d'informations est contenu dans ces mots uniquement.

Étape 3: Parcourez tous les mots du texte ci-dessus et faites une liste de tous les mots de notre vocabulaire modèle.

- welcome
- great
- learning
- now
- start
- good
- practice

Maintenant que le vocabulaire ne comporte que 7 mots, nous pouvons utiliser une représentation de document de longueur fixe de 7, avec une posi tion dans le vecteur pour noter chaque mot. La méthode de notation que nous utilisons ici est la même que celle utilisée dans l'exemple précédent. Pour la phrase 1, le nombre de mots est le suivant :

Word	Frequency
welcome	1
great	1
learning	2
now	1
start	1
good	0
practice	0

Ecrire les fréquences ci-dessus dans le vecteur Phrase 1→ [1,1,2,1,1,0,0]

Maintenant, pour la phrase 2, le score serait comme

Word	Frequency
welcome	0
great	0
learning	1
now	0
start	0
good	1
practice	1

De même, écrire les fréquences ci-dessus sous forme vectorielle Phrase 2 → [0,0,1,0,0,1,1]

Sentence	welcome	great	learning	now	start	good	practice
Sentence1	1	1	2	1	1	0	0
Sentence2	0	0	1	0	0	1	1

L'approche utilisée dans l'exemple deux est celle qui est généralement utilisée dans la technique du sac de mots, la raison étant que les ensembles de données utilisés dans l'apprentissage automatique sont extrêmement volumineux et peuvent contenir un vocabulaire de quelques milliers, voire des millions de mots. Par conséquent, le prétraitement du texte avant d'utiliser un sac de mots est une meilleure façon de procéder.

Il existe différentes étapes de prétraitement qui peuvent augmenter les performances de Bag-of-Words.

- **Limitations du sac de mots**

Bien que sac de mots soit assez efficace et facile à mettre en œuvre, cette technique présente néanmoins quelques inconvénients qui sont indiqués ci-dessous :

Le modèle ignore les informations de localisation du mot. Les informations de localisation sont une information très importante dans le texte. Par exemple, «aujourd'hui est éteint» et «est aujourd'hui éteint», ont exactement la même représentation vectorielle dans le modèle BoW.

Sac de modèles de mots ne respecte pas la sémantique du mot. Par exemple, les mots «football» et «football» sont souvent utilisés dans le même contexte. Cependant, les vecteurs correspondant à ces mots sont assez différents dans le modèle du sac de mots. Le problème devient plus sérieux lors de la modélisation des phrases. Ex: «Acheter des voitures d'occasion» et «Acheter des voitures anciennes» sont représentés par des vecteurs totalement différents dans le modèle du sac de mots. [39]

2.4.6 Outils d'analyse des sentiments :

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

1. L'analyste "Werfamous"

Outil d'analyse en ligne gratuit, donnant un score de sentiment sur une échelle de -100 à 100, ainsi qu'un niveau de confiance lié à ce score. Le site sauvegarde des traceurs textes (cookies) sur votre appareil afin de vous garantir de meilleurs contenus et à des fins de collectes statistiques. Vous pouvez désactiver l'usage des cookies en changeant les paramètres de votre navigateur. En poursuivant votre navigation sur notre site sans changer vos paramètres de navigateur vous nous accordez la permission de conserver des informations sur votre appareil. Vous placez votre requête dans la boîte de dialogue et obtenez une analyse sur le web Les requêtes sont sauvegardées dans leur Jeu de données et Historique sections [40]

2. La list de mots "AFINN"

AFINN est une liste de mots valant pour valence avec un nombre entier entre moins cinq (négatif) et plus cinq (positif) . évalue la positivité/négativité d'un mot à l'aide d'un dictionnaire contenu dans une archive

Bien que le titre du document associé suggère qu'il est basé sur le corpus étiqueté ANEW, il ne l'est pas. Le titre est simplement un wordpun. Il a été développé indépendamment de la liste de mots, et ce n'est pas une révision de celui-ci. Par rapport à ANEW, la liste de mots AFINN a plus de mots et comprend des mots obscènes. ANEW d'autre part a (outre la valence) l'excitation et la dominance pour chaque mot et chaque

Mot a été étiqueté par plusieurs personnes et la moyenne et la déviation standard sont données. L'AFINN n'a été étiqueté que par Finn Årup Nielsen. Finn Årup Nielsen n'était aucunement impliqué dans le éveloppement d'ANEW. ANEW a été développé par Margaret M. Bradley et Peter J. Lang. [41]

3. L’ enquêteur général : (en :General Inquirer)

Lemmatise les mots, effectue une analyse graphique et statistique et produit un rapport contenant des phrases avec les mots les plus significatifs .

Ce site est divisé en plusieurs sections, donnant à la fois des informations sur l'Inquirer et des points vers d'autres systèmes. Les pages de notre site Web contiennent des liens vers les 100 premiers mots de chaque catégorie dans les dictionnaires Harvard et Lasswell. [42]

4. l’analyste des sentiments : “SenticNet”

Parler de SenticNet, c'est parler d'analyse des sentiments au niveau conceptuel, c'est-à-dire effectuer des tâches telles que la détection de polarité et la reconnaissance des émotions en s'appuyant uniquement sur la sémantique et la linguistique.

Dans ce contexte, SenticNet peut être l'une des choses suivantes:

- une base de connaissances conceptuelle;
- un cadre multidisciplinaire;
- une entreprise privée.

En tant que base de connaissances, SenticNet fournit un ensemble de sémantiques, de sentiques et de polarité associées à 100 000 concepts de langage naturel. En particulier, la sémantique est la plus sémantiquement liée au concept d'entrée (ie les cinq concepts qui partagent plus de caractéristiques sémantiques avec le concept d'entrée), les sentics sont des valeurs de catégorisation des émotions exprimées en termes de quatre dimensions affectives (Agréable, Attention, Sensibilité, et Aptitude) et la polarité est un nombre

Flottant entre -1 et +1 (où -1 est la négativité extrême et +1 est la positivité extrême). La base de connaissances est téléchargeable gratuitement en tant que fichier XML autonome et sa dernière version (publiée tous les deux ans) est également accessible en tant qu'API.

En tant que cadre, SenticNet se compose d'un ensemble d'outils et de techniques pour l'analyse des sentiments combinant le raisonnement de bon sens, la psychologie, la linguistique et l'apprentissage automatique. Dans ce contexte, SenticNet est plus communément appelé «informatique sismique», un paradigme multidisciplinaire qui va au-delà des simples approches statistiques pour se concentrer sur une représentation préservant la sémantique des concepts de langage naturel et sur la structure des phrases [43]

5 . L'extrateur : "SentiWordNet"

SentiWordNet est une ressource lexicale pour l'extraction d'opinion. SentiWordNet attribue à chaque synset de WordNet trois scores de sentiment: positivité, négativité, objectivité.

Dans le domaine de l'analyse de sentiment, une étude comparative a été effectuée afin de déterminer quels étaient les avantages et inconvénients de chaque source de données.

Dans le cadre d'analyse de tweets relatifs à des évènements majeurs, l'étude met en avant le fait que plusieurs de ces tweets n'ont pas pu être reconnus par les sources de données.

On peut y voir que SentiWordNet, SenticNet et SentiStrength semblent couvrir un plus grand nombre de tweets. Cependant l'article met également en évidence que le taux de couverture n'est pas synonyme de reconnaissance efficace et que la polarité d'un mot donné n'est pas fiable. C'est pourquoi l'article se propose de combiner plusieurs de ces méthodes afin d'exploiter les avantages de chacun et d'obtenir le résultat le plus proche possible de la réalité [44]

6. Sentiment140

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

Sentiment140 permet de découvrir des sentiments des tweets d'une marque, un produit ou un sujet sur Twitter. Site officiel : http://www.sentiment140.com . [45]

7. le service "Tweetfeel"

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

L'évaluation de TweetFeel se fait sur la base de présence de mots clés précis dans les tweets tels que Good, Bad, etc... (Uniquement anglais pour l'instant), a quand un service similaire en français.

Ensuite un pourcentage est calculé selon le nombre de tws ou négatifs un sentiment global de Twitter sur la marque. [46]

8. le classificateur Twitrratr

Twitrratr est un outil en ligne gratuit, qui a émergé à partir d'un projet Startup Weekend. Twitrratr fonctionne à partir d'une liste de mots positifs et d'une liste de mots négatifs [47]

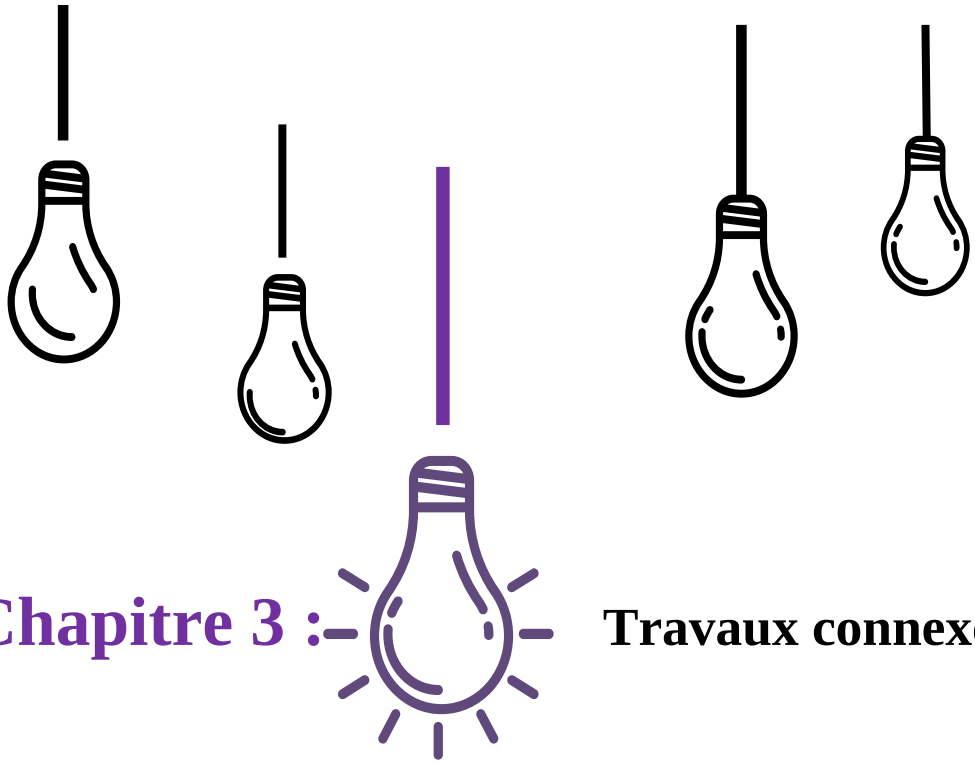
Cet outil classe une opinion sur le mot clé de la requête s'il est capable de le croiser avec un mot d'une des deux listes. Les mots positifs et négatifs qui servent à classer les tweets sont surlignés dans l'interface . [48]

9. Tweet Sentiments Analyses

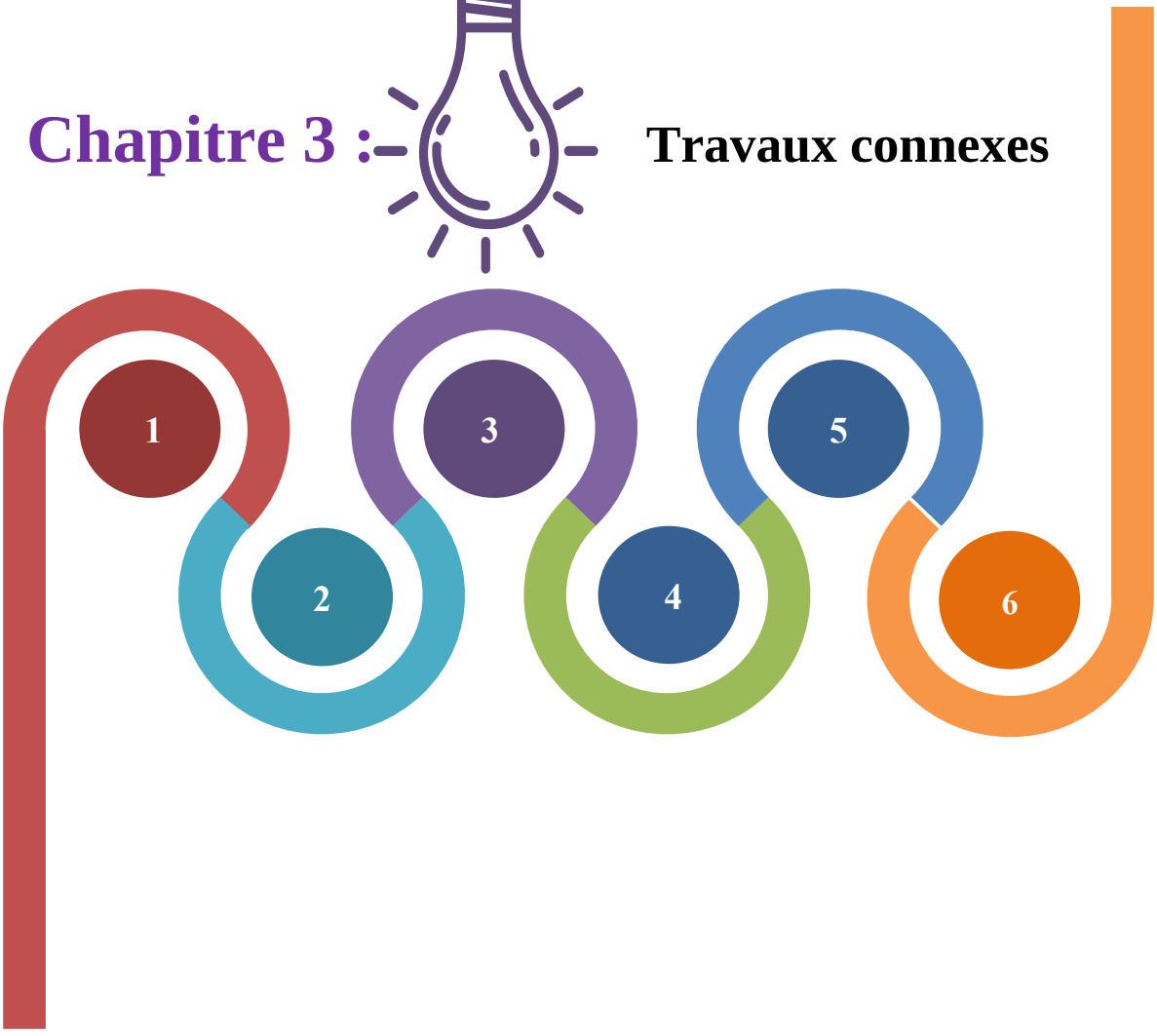
Tweet Sentiments Analyses est un outil en ligne gratuit et open source d'analyse du sentiment sur Twitter. Il peut donner des sentiments positifs, négatifs et neutres des tweets sur le mot clé lancé dans la requête. Il peut travailler sur 12 langues. Il donne les résultats sous forme graphique . [49]

2.5 Conclusion

Dans ce chapitre nous avons fait un survole sur les réseau sociaux et les principaux fondements, méthodes et techniques de classification des sentiments et d'opinion mining. dans le chapitre suivant dans le chapitre suivant nous allons détailler plus les travaux faites dans ce domaine en relation avec le traitement automatique du langage naturel et présenter leur résultats.



Chapitre 3 : Travaux connexes



3.Travaux connexes

3.1 Introduction:

Dans ce chapitre, nous présentons les travaux liés à nos recherches. Nous avons trouvé plusieurs études de notation, chacune selon sa méthode. Certains d'entre eux sont présentés dans ce chapitre. Il y a des travaux qui dépendent de la façon dont l'apprentissage automatique, y compris le premier travail «Analyse du sentiment de grand data à l'aide d'algorithmes d'apprentissage machine » et le second travail «Détecter les affiliations extrémistes basées sur les médias sociaux Classification sociale et classification à l'aide de l'analyse des sentiments», mais le troisième ouvrage compare plus d'un algorithme de classification «Une application conjointe des techniques d'exploration de données pour l'analyse de la prévention et de la prévision des attaques terroristes mondiales pour lutter contre le terrorisme» et le dernier ouvrage «Extrême Propagande classe les tweets avec apprentissage en profondeur dans des scénarios réalistes » Utilisez d'autres méthodes de classification. Nous en apprendrons davantage sur elles à travers ce chapitre.

3.2 Analyse à l'aide d'algorithmes d'apprentissage machine

Dans le travail .[50] ,Kamal Al-Barzanji et Atanas Atanasov ont analysé le sentiment des méga données à l'aide d'algorithmes d'apprentissage automatique. car les gens vivent à l'ère des méga données. réseaux sociaux, etc... Ce sont les principales sources de génération de données textuelles énormes par les utilisateurs. Ces « données textuelles » peuvent fournir aux entreprises un aperçu de la façon dont les consommateurs perçoivent leur marque et leur permettre de prendre activement des décisions commerciales pour maintenir leur activité. Par conséquent . .[50]

3.2.1 EXPÉRIENCES

Il règle ce qui suit

3.2.1.1 Ensemble de données

Ensemble de données prédéfini et prêt à être utilisé par des algorithmes d'apprentissage automatisés dans une zone d'analyse .Il comprend 500 commentaires positifs et 500 suspension négative ont été choisis au hasard dans chaque site. L'examen est évalué avec des émotions négatives de 0 et l'examen positif est évalué. [50]

3.2.1.2. Ensembles de données de formation et de test

Les ensembles de données ont été divisés en un ensemble de données d'exercice de 80% et à 20% de données de test aléatoire. [50]

Les ensembles de données créés dans le groupe de données de pré-entraînement ou de nettoyage ont été traités via des concepts NLP, afin de convertir la chaîne de saisie en petits caractères et de les diviser en mots (codes) à l'aide d'espaces vides tels que les séparateurs et supprimez des codes inutiles. TF-IDF a été mis en œuvre pour un sac de mots. Les algorithmes d'analyse émotionnelle ont ensuite été appliqués à l'aide d'un sac de mots. [50]

Le dessin suivant montre toutes les étapes

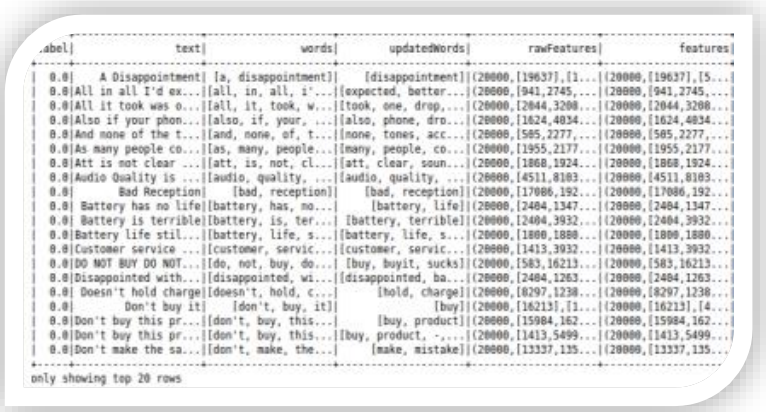


Figure 3.1:Étapes prétraitées Résultats pour Amazon Base de données [50]

3.2.1.3. conséquences

Les résultats des algorithmes d'apprentissage automatique semblent analyser le texte dans les données Cellule Amazone définies aux figures 3.2 et 3.3, respectivement. Ces résultats résultent des résultats du modèle de conduite final pour chaque ensemble de données de test, après avoir exécuté des modules de formation sur le jeu de données de formation. Le modèle naïf Bayes produit des probabilités et des prévisions de fonctionnalités .

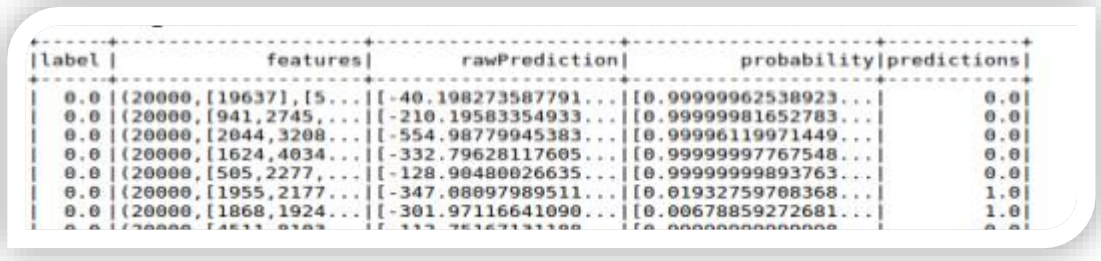
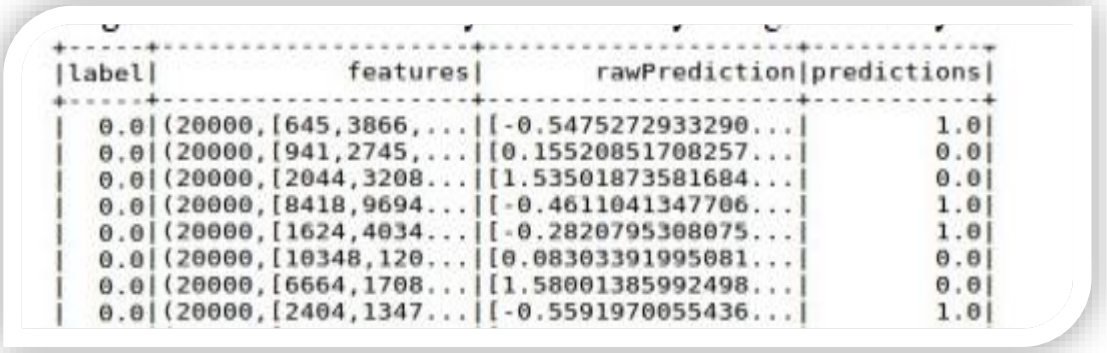


Figure 3.2:Résultats de l'analyse des sentiments en utilisant Naïve Bayes[50]



label	features	rawPrediction	predictions
0.0	(20000,[645,3866,...	[-0.5475272933290...	1.0
0.0	(20000,[941,2745,...	[0.15520851708257...	0.0
0.0	(20000,[2044,3208...	[1.53501873581684...	0.0
0.0	(20000,[8418,9694...	[-0.4611041347706...	1.0
0.0	(20000,[1624,4034...	[-0.2820795308075...	1.0
0.0	(20000,[10348,120...	[0.08303391995081...	0.0
0.0	(20000,[6664,1708...	[1.58001385992498...	0.0
0.0	(20000,[2404,1347...	[-0.5591970055436...	1.0

Figure 3.3: Résultats de l'analyse des sentiments à l'aide de SVM linéaire[50]

3.2.1.4 Évaluations du modèle

Les résultats sont présentés dans le tableau 1 et la figure 3.4 respectivement, et les mesures de résolution intermédiaire révèlent que le modèle Linéaire SVM est meilleur que le modèle Naïf Bayes en utilisant des ensembles de données de test pour évaluer cet ensemble de données. Tous les travaux ci-dessus ont été effectués à l'aide d'algorithmes d'apprentissage automatique supervisé et de bibliothèques d'apprentissage automatique grand Data . [50]

Techniques with Evaluation Metrics	**Naïve Bayes**	**Linear SVM**
Accuracy Evaluation Metrics1	0.8	0.79803
Accuracy Evaluation Metrics2	0.78421	0.7973
Accuracy Evaluation Metrics3	0.77941	0.83333
Accuracy Evaluation Metrics4	0.77251	0.80729
Accuracy Evaluation Metrics5	0.80303	0.8
Average Accuracy Evaluation Metrics	**0.78783**	**0.80719**
Test Error	0.21217	0.19281

Tableau 3.1: Résultats des mesures d'évaluation des techniques[50]

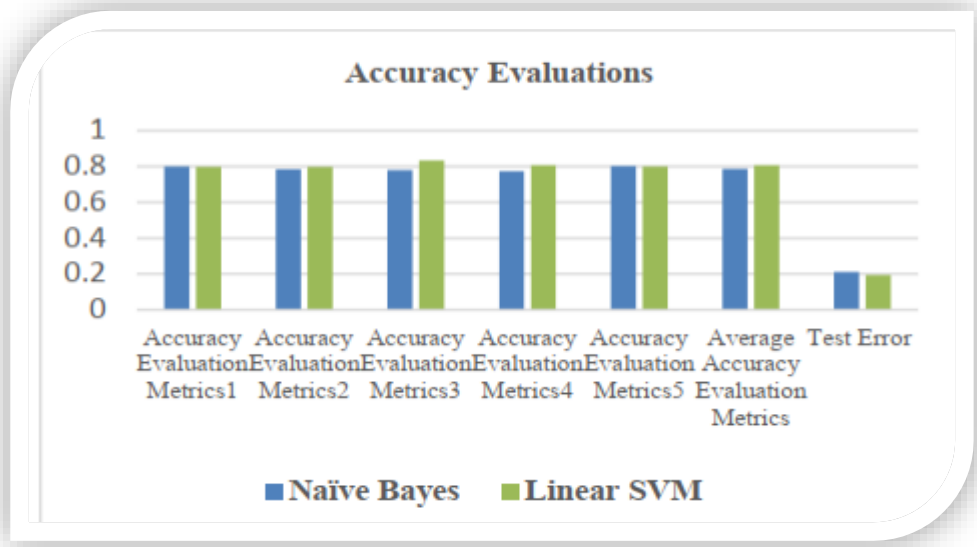


Figure 3.4: Résultats des mesures d'évaluation des techniques[50]

3.3 Détection et classification à l'aide de techniques d'analyse des sentiments

Identifier et catégoriser les tweets liés aux extrémistes est une question urgente. Les gangs extrémistes tentent d'utiliser des sites de médias sociaux tels que Facebook et Twitter pour diffuser leur idéologie et recruter des individus. Par conséquent, Shakil Ahmed et ses collègues ont tenté de présenter le travail .[51] ,cette étude sur la détection et la classification des affiliations extrémistes collectées sur les réseaux sociaux à l'aide de techniques d'analyse des sentiments dans le but de fournir un cadre d'analyse du contenu lié au terrorisme en mettant l'accent sur la classification des tweets en extrémistes et catégories non extrémistes.

3.3.1 Collecte de données:

Utilisez l'API de streaming Twitter pour supprimer les Tweets contenant un ou plusieurs mots clés liés à l'extrémisme. De plus, ils ont également enquêté sur divers forums du Sombre Web, tels que Al-Firdaws. Ils ont collecté plus de 25 000 articles en traduisant des articles non anglais en anglais à l'aide de l'API de traduction basée sur Google Python. Chaque avis est mis en correspondance avec des mots-clés dans notre dictionnaire de vocabulaire extrême fabriqué à la main, provenant de BiSAL, un dictionnaire bilingue pour l'analyse des forums sur le Web sombre. De cette manière, tous les articles contenant un ou plusieurs mots-clés du glossaire obtenu sont collectés manuellement.. [51]

3.3.2 Prétraitement

Ils ont appliqué différentes techniques de prétraitement, telles que la création de jetons, la suppression de mots vides, la conversion de casse et la suppression de symboles spéciaux. La tokenisation produit un ensemble de jetons uniques (356, 242), qui aident à construire un vocabulaire à partir de la formation. [51]

3.3.3 Apprentissage, validation et tests:

Ils ont divisé l'ensemble de données en trois parties: la formation, la validation et les tests. Le modèle DL a été formé avec la bibliothèque Keras basée sur Tensor Flow. La configuration matérielle requise comprend 4 GPU son Titan X et 128 Go de mémoire avec le nœud Intel Corei7. Les données de formation sont utilisées pour former le modèle, 80% des données sont utilisées pour la formation et peuvent varier en fonction des exigences en matière d'expérience. Les données de formation comprennent à la fois les intrants et les extrants attendus. Un jeu de validation de 10% est utilisé pour éviter les erreurs de performances lors de l'application du réglage des paramètres. À cette fin, ils ont mis en œuvre une vérification automatique des ensembles de données, qui fournit une évaluation impartiale du modèle et minimise la désaffectation . [51]

3.3.4 Modèle de réseau:

La méthode proposée est mise en œuvre et évaluant les performances de performance à long terme avec un modèle CNN (réseau de neurones concurrentiels) pour déterminer les tweets / critiques contenant du contenu avec des indicateurs extrêmes et la formation du classeur nerveux, pour classer le contenu. Extrêmement affiliation.

Le flux de travail de réseau comprend les étapes suivantes:

1. insérer des mots, dans lequel un mot est défini à partir de la vente en gros existe dans l'index
2. la couche de don est utilisée pour éviter une modification supplémentaire, (3)
3. La couche LSTM a été fusionnée de dépendance à longue distance dans les propriétés d'extrait Tweets / III (3) à l'aide du processus d'exécution,
4. Objectifs de couche de piscine pour réduire la taille des caractéristiques en collectant des informations
5. aplatir la couche convertit la collection de l'entité collectée sur le vecteur de colonne,
6. à la couche de sortie, la fonction Soft max est utilisée pour la classification. [51]

3.3.5Analyse des sentiments utilisateur par rapport à leur développement émotionnel avec des extrémistes:

Ce module répond à la notation des émotions des utilisateurs. Chaque texte est classé avec une classe émotionnelle à l'aide de l'API Python Analyzer basée sur Python. Restaure une gamme d'émotions: {en colère, chagrin, peur, joie, confiance, analyse et vitesse}. Dans ce module, le texte d'entrée est analysé et la classe de passion correspondante est déterminée. [51]

3.3.5.1. résultats et discussion

Comment identifier et classer les tweets comme extrêmes ou no extrêmes et appliquer des sentiments basés sur un apprentissage profond?

Pour répondre à cette question de recherche, ils ont appliqué une technique d'émotions d'apprentissage en profondeur et sont LSTM + CNN.

Ils ont effectué des expériences sur la couche CNN 1, LSTM 1 couche. Les modèles sont évidemment un modèle proposé LSTM + CNN obtient les meilleurs résultats. Le niveau suivant, la couche CNN récupère des fonctionnalités supplémentaires (N-gram) du groupe le plus riche, offrant ainsi de meilleurs résultats de performance améliorés.

Le tableau 3.2 montre les résultats obtenus grâce à l'application de différents modèles DL, évidemment le modèle proposé LSTM + CNN obtient les meilleurs résultats.

Models	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
1-layer LSTM	85.07	87	85	85
1-layer CNN	83.09	84.4	82	82
LSTM + CNN	92.66	88.32	89.47	90.71

Tableau 3.2: Résultats expérimentaux de différents modèles SA basés sur DL[51].

Le tableau 3.3 montre la configuration du modèle proposé (LSTM + CNN). L'expérience est réalisée en utilisant divers paramètres,

LSTM + CNN model layers	Parameters with values
Convolutional layer	Number of filters = 2, 4, 6, 8, 10, 12, 14, 16 Kernel size = 2 × 2 Padding = same
Pooling layer	Pooling size = 2 × 2
LSTM layer	Units = 100
Additional	Vocabulary size = 2000 Input vector size = 50 Embedding dimension = 128 Batch size = 8 Number of epochs = 7

Tableau 3.3: Paramétrage du modèle proposé LSTM + CNN. [51]

3.3.5.2 Réglages des paramètres:

Le tableau 3.3 présente le paramétrage du modèle proposé (LSTM + CNN) ,L'expérience est menée en utilisant différents paramètres, [51]

3.3.5.3Méthodes d'apprentissage en profondeur avec variantes

Différentes variantes de techniques DL (CNN + Intégration aléatoire, LSTM + Intégration aléatoire , texte rapide + incorporation aléatoire et GRU + Intégration aléatoire) sont utilisées pour la classification des tweets comme extrémistes et non extrémistes. Les résultats sont rapportés dans le tableau 3.4 . Le LSTM + CNN proposé fonctionne mieux que les autres méthodes DL. [51]

Study	Technique	Results		
		P	R	F
Part-A Baselines	N-gram + SVM [12]	73	78	79
	Bag of words + SVM [13]	73	74	73
	TF-IDF + RF	78	84	90
	BOW + KNN [8]	76	75	76
Part-B Deep learning methods with variants	CNN + word embedding	84	84	84
	LSTM + word embedding	86	86	86
	FastText + word embedding	87	87	87
	GRU + word embedding	85	85	85
Part-C Proposed	LSTM + CNN	0.90	0.88	0.88

Tableau 3.4 : Comparaison du mot proposé avec les méthodes de base[51].

3.3.5.4 La méthode

Pour la tâche de classification extrémiste, ils ont expérimenté différents modèles SA basés sur DL, à savoir CNN, LSTM, Texte rapide et GRU, initialisés avec des plongements de mots aléatoires. Le modèle proposé LSTM + CNN pour la classification extrémiste surpasse les méthodes de base et il fonctionne également mieux que les autres modèles DL .Ils ont mené des expériences en utilisant une technique de référence supervisée, basée sur le lexique, proposée pour la classification extrémiste sur Twitter. [51]

3.3.5.5 Techniques supervisées:

Dans le tableau 3.5, les résultats de la méthode proposée (LSTM + CNN) sont comparés à diverses techniques avancées basées sur des classificateurs de machines, à savoir KNN huit et Naïve Bayes Classifier. D'autres classificateurs appliqués à l'apprentissage automatique et en profondeur, tels que la forêt aléatoire, KNN a fourni le score de performance le plus bas (précision de 72%). [51]

Study	Techniques	Features	P (%)	R (%)	F (%)	A (%)
Wei et al. [8]	Machine learning (KNN)	Classical features (tf, idf, tfxidf) BOW	73	71	71	74
Azizan and Aziz [7]	Machine learning (NB)	Classical features (tf, idf, tfxidf)	70	68	69	72
Proposed	Deep learning (LSTM + CNN)	Word embedding	90	86	84	92
Chalothorn and Ellman [17]	Lexicon-based (Senti WordNet)	Sentiment scores and polarity classes	69	68	69	73
Other Models	Deep learning (LSTM)	Word embedding	85	84	83	85
	Deep learning (CNN)	Word embedding	88	84	83	88
	Machine learning (SVM)	Tf xidf	79	78	79	79
	Machine learning (RF)	Tf xidf	83	81	82	84

Tableau 3.5: Comparaison avec la technique de pointe[51]

La technique proposée (LSTM + CNN) a donné les meilleurs résultats par rapport aux autres méthodes de comparaison. Le travail proposé fonctionne en trois modules: (1) la collecte des tweets des utilisateurs, (2) le prétraitement et (3) le profilage contre les extrémistes. Et des catégories non extrêmes utilisant le modèle LSTM + CNN et les classificateurs ML et DL.

Les résultats expérimentaux ont montré la supériorité du système proposé sur les méthodes de comparaison en termes d'exactitude, de rappel, de mesure et d'exactitude. [51]

3.4 Une application conjointe des techniques d'exploration de données

Dans le travail [52**]** nous visons à mettre que Le terrorisme est devenu l'un des problèmes les plus difficiles à traiter et un grand danger pour l'humanité. Pour renforcer la lutte contre le terrorisme, il y a beaucoup de recherche pour développer des systèmes efficaces et précis, et l'exploration de données ne fait pas exception. Les méga données flottent dans nos vies, bien que les données réelles sur les attaques terroristes soient rarement disponibles dans le domaine public, ce qui rend la lutte contre le terrorisme difficile. .

3.4.1 Description de l'ensemble de données

Les données sur les attentats terroristes sont obtenues à partir de sources de données sur Internet, en les collectant individuellement. Cet ensemble de données comprend 156 772 attaques signalées dans le monde, qui se sont produites entre 1970 et 2015. L'analyse des données se termine après avoir affiné les données primaires avec neuf caractéristiques pour catégoriser les données de cette étude: mois d'attaque, année d'attaque, zone, type d'arme, cible , type d'attaque et source Perte de données et de biens. [52**]**

Pour ce travail, la «responsabilité de l'attaque» était une catégorie, divisée en trois catégories qui sont revendiquées et anonymes. Une classification supplémentaire des données est également effectuée pour faciliter la classification. Le total des attaques est divisé en fonction des responsabilités d'attaque de classe présumées et anonymes. [52**]**

3.4.1.1 Résultats et discussion

. Ils ont obtenu des résultats avec une grande précision allant de 90 à 95%.Cette étude a tenté de classer les attaques terroristes mondiales à l'aide de différents classificateurs d'exploration de texte tels que l'arbre de décision aléatoire, le classificateur Paresseux IBK Linéaire NN, le classificateur Paresseux IBK Voisin filtré et le classificateur Fainéant K.-star. Cette analyse peut être utilisée pour dessiner des modèles pour extraire des informations sur les attentats terroristes. Voici quelques résultats des travaux:

- Sur le total des attaques signalées de 1970 à 2015, 156 772 attaques ont été signalées, 14 664 attaques, 75 966 n'ont pas été approuvées et 66 142 étaient inconnues.
- Le pourcentage de plaintes liées à des attaques d'organisations terroristes varie de 10 à 15%. Dans cette étude, nous sommes arrivés à un taux de 11%.

Le pari principal est l'Asie pour soutenir les attaques qui représentent environ 45% du total des attaques.

• Le gouvernement et les forces de l'ordre (armée et police) sont les principales cibles à 40%, suivis des citoyens ordinaires et des biens à 22%.

• Les pertes dues aux attentats terroristes sont les plus élevées pour les pertes moyennes, qui ne s'élèvent qu'à 64%.

Les types d'attaques tels que les bombardements et les agressions armées étaient les plus courants, avec respectivement 48,45% et 24,50%. [52]

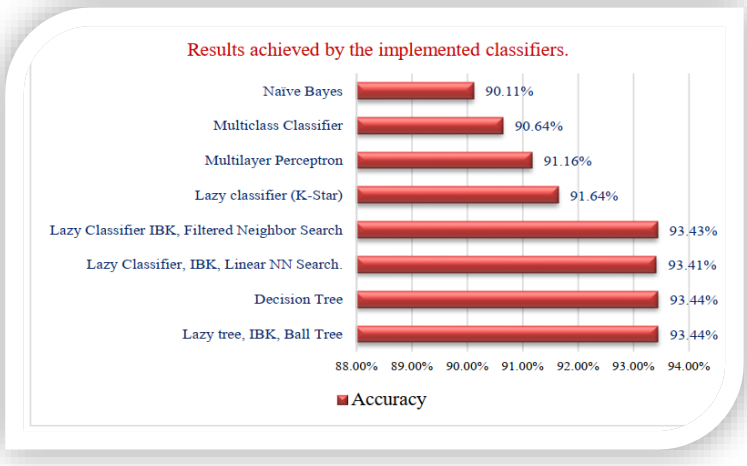


Figure 3.5:Résultats obtenus par les classificateurs implémentés. [52]

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.987	0.987	0.987	0.987	0.987	0.987	0.987	0.987	Claimed
0.114	0.114	0.114	0.114	0.114	0.114	0.114	0.114	No-Claim
1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	Anonymous

Tableau 3.6Résultat du classificateur paresseux IBK. [52]

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.999	0.161	0.853	0.999	0.921	0.845	0.972	0.962	Claimed
0.104	0.00	0.965	0.104	0.187	0.303	0.923	0.572	No-Claim
1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	Anonymous

Tableau 3.7 . Résultat de K-star du classificateur paresseux[52]**.**

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.987	0.114	0.891	0.987	0.936	0.875	0.983	0.979	Claimed
0.332	0.005	0.851	0.332	0.477	0.509	0.948	0.669	No-Claim
1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	Anonymous

Tableau 3.8 . Le résultat de l'arbre de décision . [52]

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.984	0.156	0.856	0.984	0.916	0.834	0.954	0.929	Claimed
0.141	0.008	0.611	0.141	0.230	0.268	0.864	0.389	No-Claim
1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	Anonymous

Tableau 3.9: Le résultat de Multicouche Perceptron. [52]

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.999	0.180	0.839	0.999	0.912	0.828	0.942	0.910	Claimed
0.009	0.001	0.511	0.009	0.018	0.059	0.826	0.269	No-Claim
1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	Anonymous

Tableau 3.10: Le résultat du classificateur multiclasse. [52]

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.986	0.176	0.841	0.986	0.908	0.817	0.939	0.906	Claimed
0.010	0.002	0.358	0.010	0.019	0.048	0.818	0.246	No-Claim
1.00	1.004	1.994	1.00	1.997	1.995	1.00	1.00	Anonymous

Tableau 3.11 Le résultat de Naïve Bayes. [52**]**

3.5 le classification à l'aide apprentissage en profondeur

Le travail [53**]** confirme que Le volume et la variété du flux Twitter est une cachette confortable aux activités de propagande malveillantes. Ainsi, la communauté de recherche Social Media Intelligence (SOCMINT) a consacré des efforts croissants au développement de méthodes automatiques de filtrage des tweets malveillants. Si la littérature apporte plusieurs contributions sur ce sujet , les techniques proposées sont généralement testées et évaluées en laboratoire. Par conséquent, les parties prenantes ne peuvent pas estimer comment ces modèles fonctionneraient dans la nature. [53**]**

3.5.1 Matériels et méthodes

Ils ont construit plusieurs ensembles de données en combinant des tweets pro-ISIS et des tweets aléatoires dans des proportions variables. Des échantillons de 15684 tweets pro-ISIS de langue anglaise ont été extraits d'un vaste ensemble de données fiable et accessible au public publié par Cinquième tribu. Ils ont collecté des tweets aléatoires via l'API Twitter Streaming, via la méthode cas / échantillon GET, qui renvoie un petit échantillon aléatoire. de tous les cas publics. L'ensemble de données le plus grand et le plus confus (1%) a été inclus, moins d'ensembles de données déséquilibrés (2% à 10% équilibrés) ont été dérivés d'ensembles de données déséquilibrés, supprimant les cas négatifs. Chaque ensemble de données a été divisé en groupes pour établir les ensembles de formation (64%), la validation (16%) et les tests (20%)..

3.5.2 Résultats et discussion

Figure 3.6: Tendances du score F1, pour différentes stratégies d'entraînement, à mesure que le déséquilibre du groupe de test change. Dans 1% de la formation, ChCNN s'est comporté comme un classificateur majoritaire (aucun score F1 n'a été spécifié)[53]

Ils ont formé les modèles dans des groupes d'entraînement équilibrés, 10%, 5%, 1%, et ont évalué leurs performances lorsque le déséquilibre du groupe d'essai a changé dans la plage [1%, 10%]. La figure 3.5 montre les tendances du score F1 obtenu. Les modèles qui ont été formés sur le groupe d'entraînement équilibré ont souffert d'une grave dégradation des performances lors de la réduction du pourcentage de cas positifs dans le groupe de test. Au lieu de cela, la performance a montré une tendance plus stable à mesure que nous augmentions le nombre d'exemples négatifs dans l'ensemble d'entraînement. Baseline Linéaire ChSVM a obtenu le score F1 le plus élevé lors de l'utilisation d'un ensemble d'entraînement à 10%, mais a surpassé 1%. ChCNN était compétitif avec le groupe de formation à 5%, mais il en a résulté que la majorité a été notée, avec un score F1 indéterminé, de 1%. Les RCNN basés sur Word ont clairement surpassé de 1% les autres candidats de l'ensemble de données plus équilibré, tout en maintenant la cohérence dans l'ensemble de la plage examinée. En particulier, WeRCNN avait le score F1 le plus élevé (0,895) dans le groupe de test 1%.

Les RCNN décrivent, sur la base de faux mots pré-testés, des technologies appropriées. En outre, ils soutiennent le choix d'un ensemble d'entraînement très déséquilibré si un défaut significatif inconnu dans le flux réel de tweets est à prévoir. Enfin, il permet d'apprécier les performances dans un large éventail de conditions . [53]

3.6 Etude comparative

Dans ce qui suit, nous tenterons d'analyser les résultats de chacune des études précédentes.

3.6.1 Comparer les résultats de Analyse à l'aide d'algorithmes d'apprentissage machine

Ils ont introduit un système d'analyse du Gros Data à l'aide d'algorithmes d'apprentissage automatique. Ils ont conçu et développé des algorithmes Naïf Bayes et un support vectoriel pour le système d'analyse de texte pour les revues avec Apache Étincelle, les performances des algorithmes ont été mesurées sous forme de précision. L'ensemble de données est pré-marqué et prêt à être utilisé par des algorithmes d'apprentissage automatique supervisés dans la zone d'analyse des sentiments.

Les ensembles de données ont été divisés en deux parties: les données d'entraînement et les données de test. Les ensembles de données ont été divisés en un ensemble de données de train de 80% et un ensemble de données de test aléatoire de 20%, et l'ensemble de données a été chargé dans le système à partir du système de fichiers local. Les ensembles de données créés dans l'ensemble de données du train ont ensuite été nettoyés via les concepts PNL, pour convertir la chaîne d'entrée en lettres minuscules, la diviser en mots pour utiliser des espaces vides comme virgules et les supprimer. Le vecteur d'attribut a ensuite été généré pour chaque phrase de l'ensemble de données, de sorte que TF-IDF a été exécuté pour obtenir un sac de mots. Ensuite, elle a appliqué des algorithmes d'analyse émotionnelle à l'aide d'un sac de mots.

Les résultats des algorithmes d'apprentissage automatique pour l'analyse de texte apparaissent dans l'ensemble de données balisé cellule amazonienne et constituent le résultat du test final des résultats du modèle de pilotage pour chaque ensemble de données de test, après l'exécution des échantillons. Formation sur les ensembles de données. Le modèle Naïf Bayes produit des probabilités et des prédictions à partir d'entités. Les résultats sont présentés dans le tableau 3.12 et la figure 3.7, respectivement, et les mesures moyennes de précision montrent que le modèle Linéaire SVM est meilleur que le modèle Naïf Bayes en utilisant des ensembles de données de test pour évaluer cet ensemble de données. Étant donné que la précision sera comprise entre 0 et 1 et que plus la précision est élevée, cela signifie qu'il y a moins d'erreur. Tous les travaux ci-dessus ont été effectués à l'aide d'algorithmes d'apprentissage automatique supervisé et de bibliothèques d'apprentissage automatique grand Data.

Techniques with Evaluation Metrics	Naïve Bayes	Linear SVM
Accuracy Evaluation Metrics1	0.8	0.79803
Accuracy Evaluation Metrics2	0.78421	0.7973
Accuracy Evaluation Metrics3	0.77941	0.83333
Accuracy Evaluation Metrics4	0.77251	0.80729
Accuracy Evaluation Metrics5	0.80303	0.8
Average Accuracy Evaluation Metrics	**0.78783**	**0.80719**
Test Error	0.21217	0.19281

Tableau 3.12 : Résultats des mesures d'évaluation technique [50]

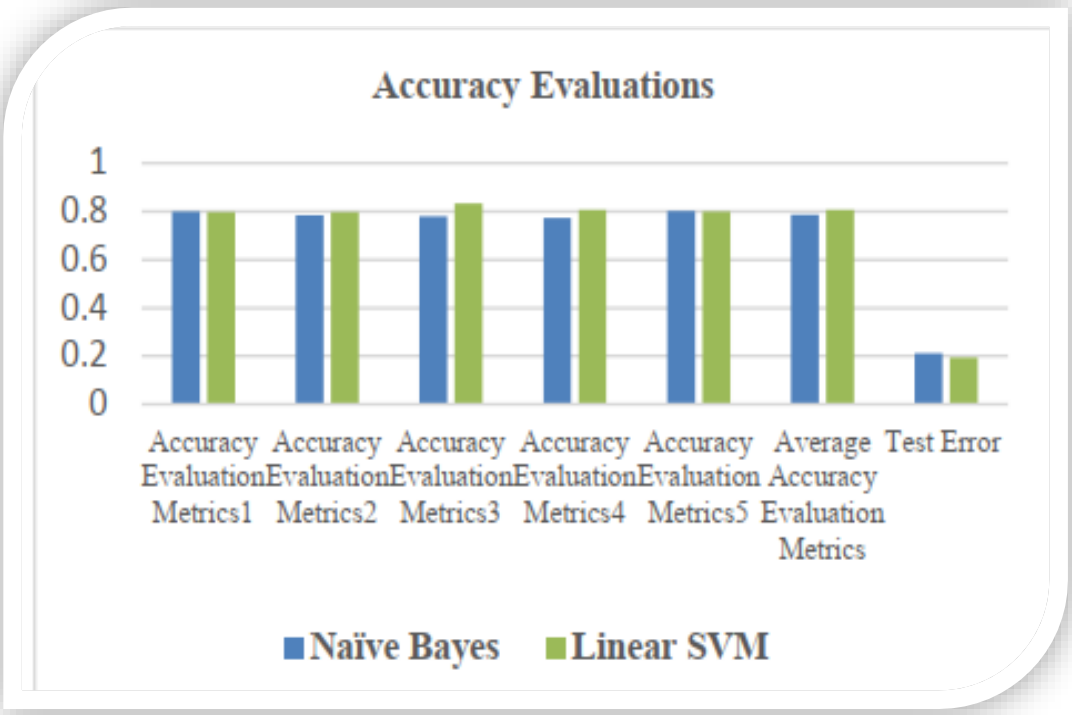


Figure 3.7: Résultats des mesures d'évaluation technique [50]**]**

3.6.2 Comparer les résultats de Détection et classification à l'aide de techniques d'analyse des sentiments :

Ce travail "Détection et classification des affiliations extrémistes basées sur les réseaux sociaux à l'aide de techniques d'analyse des sentiments" vise à proposer un cadre d'analyse de contenu lié au terrorisme en mettant l'accent sur la classification des tweets en classes extrémistes et non extrémistes. La détection d'un tel contenu extrémiste est importante pour analyser le sentiment des utilisateurs à l'égard d'un groupe extrémiste et pour décourager de tels actes illégaux associés.

L'ensemble de données de formation comprend 12 754 tweets étiquetés «extrémistes» et 8432 «non extrémistes». Le tableau 3.13 montre le détail de l'ensemble de données utilisé.

Class label	# of Tweets	Avg. length of a Tweet	# of tokens with stop words	# of tokens without stop words
Extremist	12,754	10.47	86,971	74,354
Non-extremist	8432	8.92	64,183	52,485

Tableau 3.13 : Statistiques de l'ensemble de données [51]

Les données d'entraînement sont utilisées pour entraîner le modèle. Dans ce poste, 80% des données sont utilisées pour la formation et cela peut varier en fonction des exigences en matière d'expérience. Les données de formation comprennent à la fois les intrants et les extrants attendus. Il comprend à la fois l'entrée et la sortie attendues correspondantes.

Cette unité traite de la catégorisation des émotions des utilisateurs et de l'identification de l'affiliation des utilisateurs aux messages extrémistes. Chaque texte d'entrée est étiqueté avec une catégorie d'émotion à l'aide de l'API de l'analyseur de ton basé sur Python.

Reproduit un ensemble d'émotions: {colère, tristesse ...}. Dans ce module, le texte d'entrée est analysé et la catégorie d'émotion correspondante est identifiée.

Le tableau 3.14 montre les résultats obtenus en appliquant différents modèles DL et il est clair que le modèle LSTM + CNN proposé donne les meilleurs résultats. Le modèle LSTM + CNN a obtenu les meilleurs résultats car la couche LSTM génère une nouvelle représentation

de l'entrée tweet reçue de la couche de couverture en capturant les informations Tableau 3.18: Résultats expérimentaux pour différents modèles SA basés sur DL.

Models	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
1-layer LSTM	85.07	87	85	85
1-layer CNN	83.09	84.4	82	82
LSTM + CNN	92.66	88.32	89.47	90.71

Tableau 3.14: Résultats expérimentaux de différents modèles SA basés sur DL[51]

LSTM + CNN model layers	Parameters with values
Convolutional layer	Number of filters = 2, 4, 6, 8, 10, 12, 14, 16 Kernel size = 2 × 2 Padding = same
Pooling layer	Pooling size = 2 × 2
LSTM layer	Units = 100
Additional	Vocabulary size = 2000 Input vector size = 50 Embedding dimension = 128 Batch size = 8 Number of epochs = 7

Tableau 3.15 : Paramétrage du modèle proposé LSTM + CNN. **[51]**

Le tableau 3.15 présente le paramétrage du modèle proposé (LSTM + CNN) L'expérience est menée en utilisant différents paramètres, énumérés comme suit:

Le paramètre, à savoir «nombre de filtres», reçoit une valeur variant de 2 à 16, tandis que les valeurs d'autres paramètres, tels que la taille du noyau, le remplissage, la taille du pool, l'optimiseur, la taille du lot, les époques et les unités, sont fixes.

Model	Training time (s)	Loss score	Accuracy (%)
LSTM + CNN1	23	1.25	62
LSTM + CNN2	28	1.17	74
LSTM + CNN3	44	1.11	80
LSTM + CNN4	24	0.98	88
LSTM + CNN5	19	0.98	88.12
LSTM + CNN6	48	0.97	90.01
LSTM + CNN7	44	0.99	90.4
LSTM + CNN8	29	0.96	92.66

Tableau 3.16: Temps de formation des modèles LSTM + CNN, score de perte et précision **[51]**.

Ils incluaient la précision, le degré de perte et le temps de formation dans le tableau 3.16. Après avoir effectué des expériences avec différents paramètres pour les modèles LSTM + CNN, la précision obtenue est de 92,66%. Il est à noter que la précision du modèle augmente avec le nombre de filtres.

Les paramètres suivants sont utilisés pour le modèle LSTM + CNN, à savoir: (i) Caractéristiques max. (10000), (ii) Dimensions d'inclusion (128), (iii) Taille de l'unité LSTM (200), taille du filtre convolutif (200), et (4) la taille du lot 32 avec 3 périodes a donné les meilleurs résultats de performance .

Study	Technique	Results		
		P	R	F
Part-A Baselines	N-gram + SVM [12]	73	78	79
	Bag of words +SVM [13]	73	74	73
	TF-IDF +RF	78	84	90
	BOW + KNN [8]	76	75	76
Part-B Deep learning methods with variants	CNN + word embedding	84	84	84
	LSTM + word embedding	86	86	86
	FastText + word embedding	87	87	87
	GRU + word embedding	85	85	85
Part-C Proposed	LSTM + CNN	0.90	0.88	0.88

Tableau 3.17 : Comparaison des travaux proposés avec les méthodes de référence. **[51]**

La technique proposée (LSTM + CNN) a produit les meilleurs résultats comme le montre le tableau 3.18, par rapport aux autres méthodes de comparaison.

Study	Techniques	Features	P (%)	R (%)	F (%)	A (%)
Wei et al. [8]	Machine learning (KNN)	Classical features (tf, idf, tfxidf) BOW	73	71	71	74
Azizan and Aziz [7]	Machine learning (NB)	Classical features (tf, idf, tfxidf)	70	68	69	72
Proposed	Deep learning (LSTM + CNN)	Word embedding	90	86	84	92
Chalothorn and Ellman [17]	Lexicon-based (Senti WordNet)	Sentiment scores and polarity classes	69	68	69	73
Other Models	Deep learning (LSTM)	Word embedding	85	84	83	85
	Deep learning (CNN)	Word embedding	88	84	83	88
	Machine learning (SVM)	Tf xidf	79	78	79	79
	Machine learning (RF)	Tf xidf	83	81	82	84

Tableau 3.18 : Comparaison avec les techniques de pointe. **[51]**

Les résultats rapportés dans le tableau 3.19 montrent que le module proposé pour la classification des émotions surpasse les classificateurs d'apprentissage automatique supervisé en termes de précision, de précision, de rappel et de mesure F.

Technique/classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 measure (%)
Random forest (RF)	0.82	0.84	0.83	0.83
Support vector machine (SVM)	0.79	0.79	0.78	0.79
K-nearest neighbour (KNN)	0.72	0.72	0.72	0.71
Naïve Bayesian (NB)	0.71	0.70	0.70	0.69
Classification of users sentiments w.r.t emotional affiliation with extremists (proposed)	0.90	0.86	0.84	0.90

Tableau 3.19:Résultats comparatifs des sentiments des utilisateurs par rapport à leur affiliation émotionnelle avec des extrémistes **[51]**.

3.6.3 Comparer les résultats Une application conjointe des techniques d'exploration de données

Cet article se concentre sur la responsabilité d'attaque par catégorie pour un ensemble de données composé d'attaques terroristes signalées de 1970 à 2015. Les algorithmes de classification appliqués sont Arbre de décision aléatoire Forêt , fainéant IBK Linéaire Classifier NN, fainéant IBK Recherche de voisins filtrée Classifier, Arbre paresseux , IBK, Ball Arbre, classificateur K-star paresseux, classificateur en couches, classificateur en couches, classificateur d'achats simples.

Les données sur les attentats terroristes proviennent de sources de données en ligne. Cet ensemble de données comprend 156 772 attaques et conclut l'analyse des données avec des données primaires améliorées avec neuf caractéristiques de classification des données pour cette étude.

Pour ce travail, la «responsabilité de l'attaque» était une catégorie, divisée en trois catégories adoptées et inconnues. Une classification supplémentaire des données est également effectuée pour faciliter la classification.

Les résultats étaient d'une grande précision, variant entre 90 et 95%.

La clarification n'est obtenue que dans le contexte des performances du classificateur. Cette analyse peut être utilisée pour dessiner des modèles pour extraire des informations sur les attentats terroristes. Voici quelques résultats des travaux:

• Sur le total des attaques signalées de 1970 à 2015, 156 772 attaques, 14 664 attaques, 75 966 non approuvées et 66 142 inconnues.

• Le pourcentage de plaintes liées à des attaques d'organisations terroristes varie de 10 à 15%. Dans cette étude, un taux de 11% a été atteint.

Le pari principal est l'Asie pour soutenir les attaques, qui représentent environ 45% du total des attaques.

• Le gouvernement et les forces de l'ordre (militaires et policiers) sont les principales cibles à 40%, suivis des citoyens ordinaires et des biens à 22%.

• Les pertes résultant d'attaques terroristes sont les plus élevées de la moyenne, qui n'est que de 64%. Les types d'attaques tels que les bombardements et les agressions armées étaient les plus courants, avec respectivement 48,45% et 24,50%.

3.6.4 Comparer les résultats le classification à l'aide apprentissage en profondeur

Dans ce travail, ils ont abordé le problème de la classification automatique de la propagande extrémiste sur Twitter, en mettant l'accent sur l'État islamique en Irak et au Levant (ISIS). Ils ont construit et publié plusieurs ensembles de données. Ils ont étudié trois techniques avancées d'apprentissage en profondeur, les principales méthodes actuelles de classification de texte et deux bases de référence solides pour l'apprentissage automatique linéaire. Ils ont comparé leurs performances lors de la modification de la composition des groupes de formation et de test, afin d'explorer différentes stratégies de formation et d'évaluer les résultats à mesure qu'ils se rapprochent des conditions réelles. Ils ont démontré, sur la base de combinaisons de mots préalablement entraînées, le réseau neuronal convolutif répétitif

Ils ont construit plusieurs ensembles de données en combinant des tweets pro-ISIS et des tweets aléatoires dans des proportions variables. Des échantillons de 15 684 tweets pro-ISIS ont été extraits d'un vaste ensemble de données fiable et accessible au public. Moins d'ensembles de données déséquilibrés (équilibrés de 2% à 10%) ont été dérivés d'ensembles de données déséquilibrés, supprimant les cas négatifs. Chaque ensemble de données a été divisé

en ensembles de formation (64%), validation (16%) et tests (20%). Ils ont publié 3 ensembles de données pour la reproduction et des recherches supplémentaires.

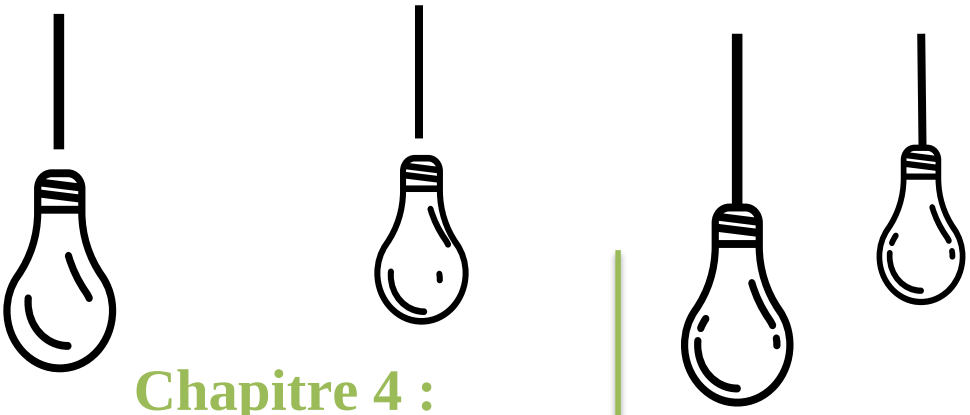
. Les modèles qui ont été formés sur le groupe d'entraînement équilibré ont souffert d'une grave dégradation des performances lors de la réduction du pourcentage de cas positifs dans le groupe de test. Au lieu de cela, la performance a montré une tendance plus stable à mesure que le nombre de cas négatifs augmentait dans l'ensemble d'entraînement. Baseline Linéaire ChSVM a obtenu le score F1 le plus élevé lors de l'utilisation d'un groupe d'entraînement à 10%, mais a surpassé 1%. ChCNN était compétitif avec le groupe d'entraînement à 5%, mais le résultat était que la majorité avec un score F1 non spécifié, ils ont obtenu 1%. Les RCNN basés sur Word ont clairement surpassé de 1% les autres candidats de l'ensemble de données plus équilibré, tout en maintenant la cohérence dans l'ensemble de la plage examinée. En particulier, WeRCNN avait le score F1 le plus élevé (0,895) dans le groupe de test 1%.

3.7 Conclusion:

Dans ce chapitre, nous avons étudié 4 travaux liés à notre thèse. Leurs titres étaient les suivants:

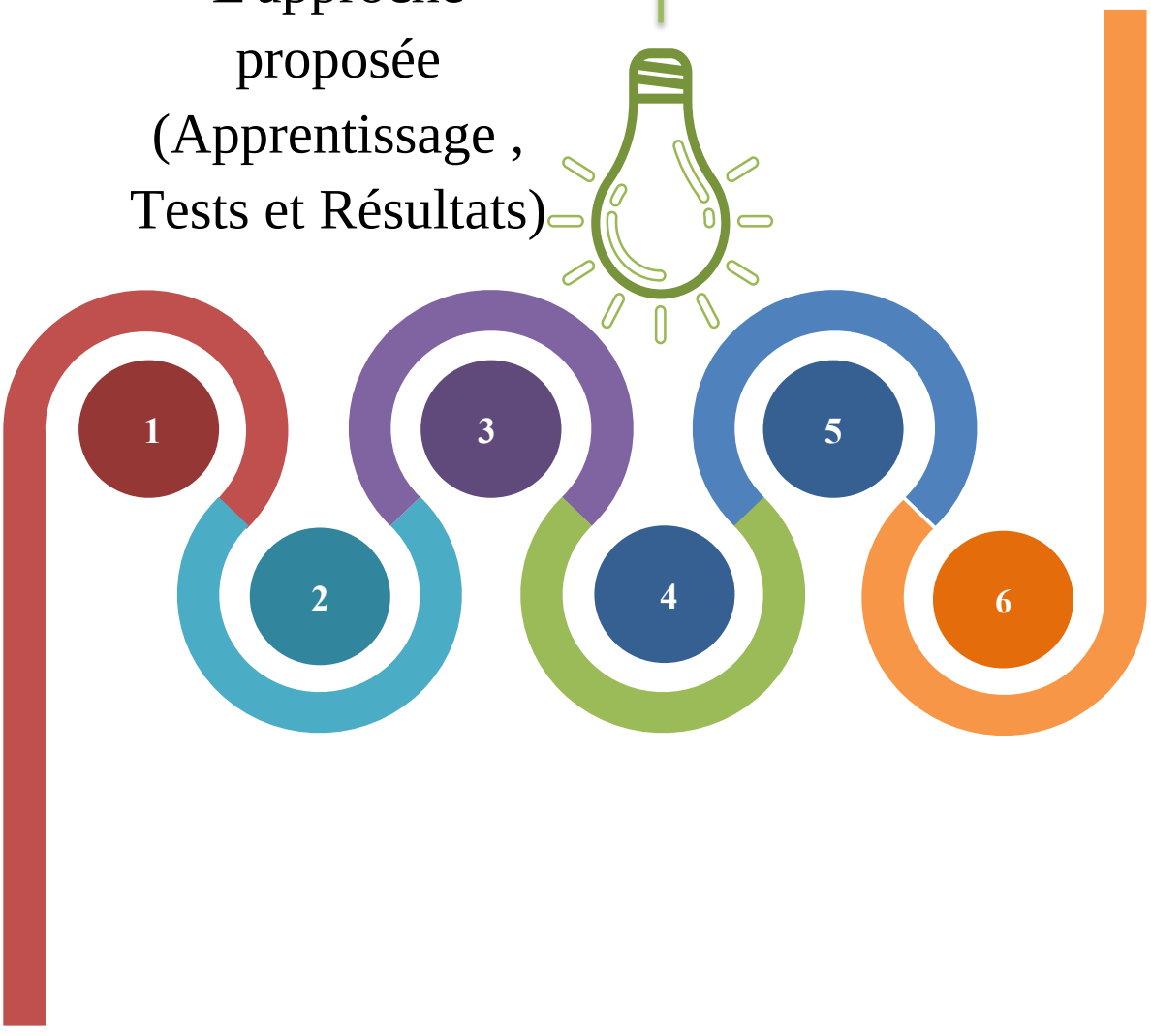
1) Analyse du sentiment de grand data à l'aide d'algorithmes d'apprentissage machine
2) Détection et classification des affiliations extrémistes basées sur les réseaux sociaux à l'aide techniques d'analyse des sentiments
3) Une application conjointe des techniques d'exploration de données pour l'analyse de la prévention et de la prévision des attaques terroristes mondiales pour lutter contre le terrorisme
4) Extrême Propagande classe les tweets avec apprentissage en profondeur dans des scénarios réalistes

Où nous avons mentionné la méthode et les algorithmes les plus importants utilisés dans leurs recherches. Nous avons présenté les résultats obtenus. Ensuite, nous avons comparé les résultats de chacune des actions que nous avons mentionnées



Chapitre 4 :

L'approche proposée (Apprentissage , Tests et Résultats)



4.1 Introduction :

Dans ce chapitre, nous commençons par donner une description de notre modèle d'analyse d'opinion. Ensuite, nous décrivons la source de données sur laquelle le modèle est appliqué. Finalement, nous avons fait l'annotation de corpus , qu'ils consomment beaucoup de temps et nécessitent un grand effort.

4.2 Le Modéle d'analyse les sentiments :

Comme de coutumes des travaux d'apprentissage, notre expérimentation passe par des étapes de prétraitement et d'apprentissage et test, tel que illustre la Figure suit :

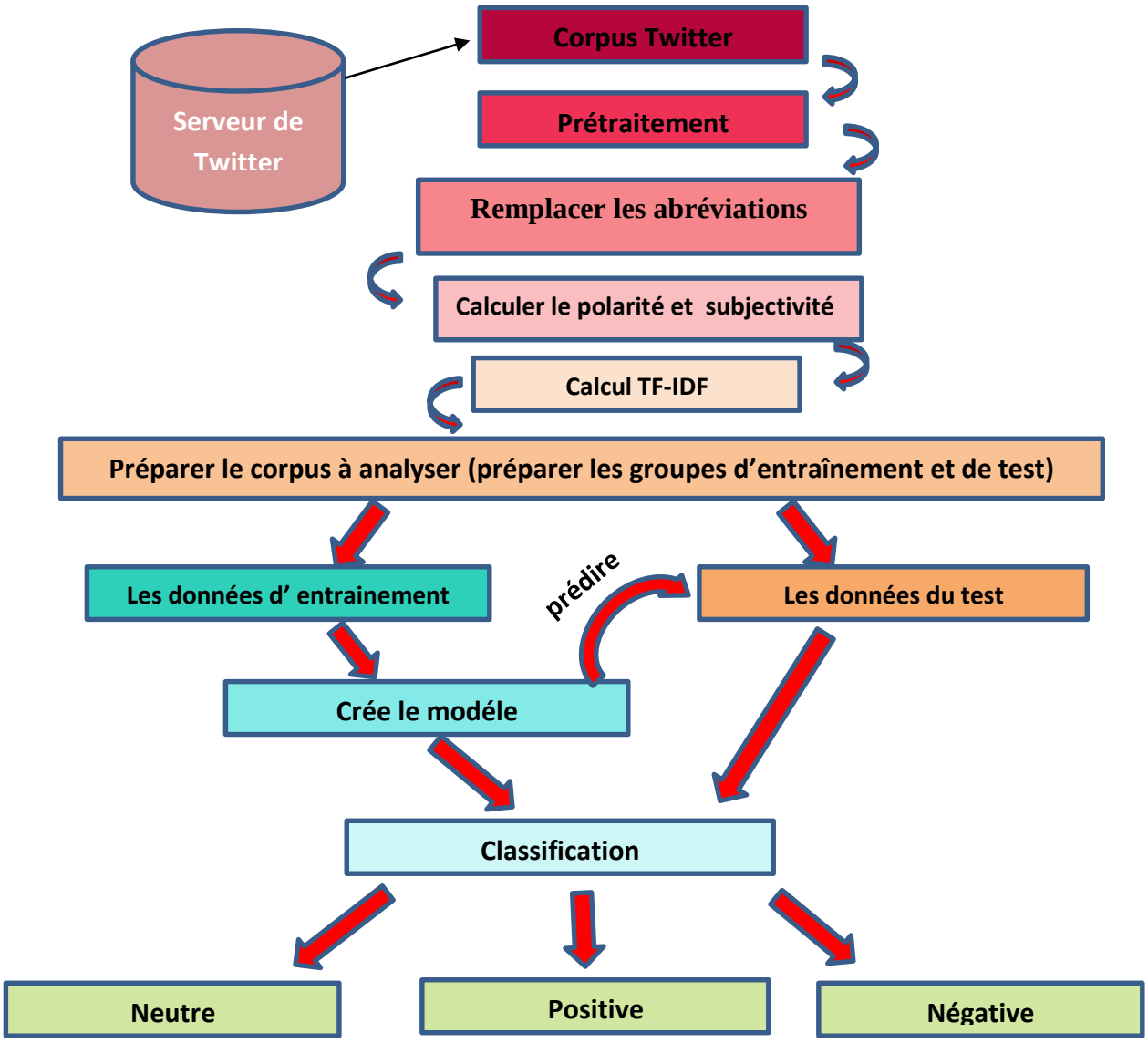


Figure 4.1 : Représentation du modèle d'analyse.

4.3 Contribution :

Mes contributions sont les suivantes :

a) La Collection des données "dataset" qui sont des 20550 commentaires obtenu de site kaggle .
b) L'application des algorithmes d'apprentissage automatique supervisé "SVM, NB, RF,DT,KNN".
c) Liste du mots similaires.
d) Des modèles de tests crée pour la classification des commentaires a partir de l'apprentissage.

4.4 Phase d'Apprentissage

La phase d'apprentissage comporte le prétraitement des données d'apprentissage ainsi que l'extraction et la présentation de descripteurs :

4.4.1 Source des données (Data set)

A- Définition : Est un ensemble de valeurs (ou données) où chaque valeur est associée à une variable (ou attribut) et à une observation. Une variable décrit l'ensemble des valeurs décrivant le même attribut et une observation contient l'ensemble des valeurs décrivant les attributs d'une unité (ou individu statistique) et c'est la première étape d'un algorithme supervisé. consiste à importer un Dataset qui contient les exemples que la machine doit étudier. Ce Dataset inclut 2 types de variables : [54]

- Une variable objectif (target) .
- Une ou plusieurs variables caractéristiques (features) .

B- Source des commentaires :

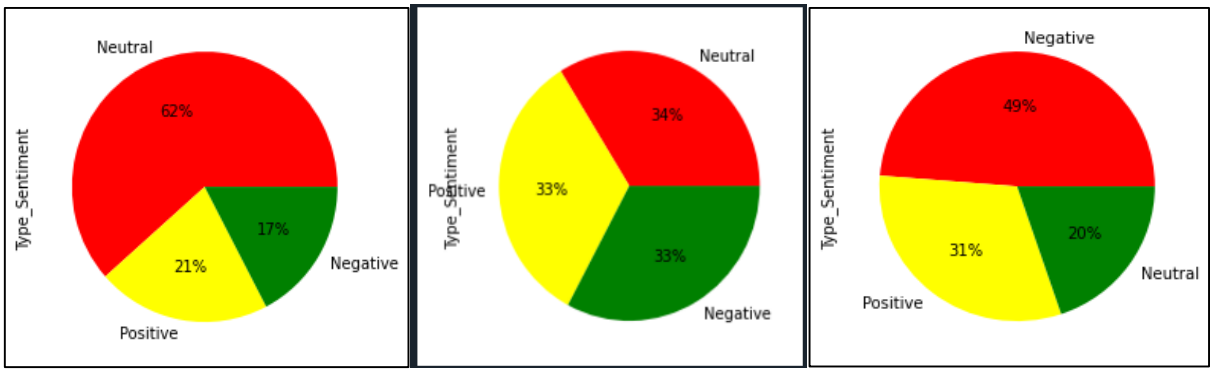
Nous avons utilisé l'ensemble des tweets télécharger par le site ' Kaggle ' . Il se présente sous forme d'un fichier d'extension (.csv) contenant 20550 tweets . Cet ensemble du données comporte trois classes des sentiments, `a savoir positive, négative, neutre. Chaque entrée de notre ensemble de données est structuré comme suit :

- Tweet id : un identifiant du tweet.
- Tweet texte : il contient le texte du tweet publié par l'utilisateur.
- TweetDate : date de publication du tweet.
- Tweet username : contient le nom de personne qui écrit le commentaire .

Nous avons organiser deux types de dataset . Le tableau 4.1 montre la répartition des échantillons de dataset qui nous utilisons dans les Tests :

Polarité / échantillons	Positive	Négative	Neutre	Total	Type de dataset
01	4344	3631	12575	20550	Non équilibré
02	1009	991	1022	3022	équilibré
03	950	1471	598	3019	Non équilibré

Tableau 4.1 : Ensemble de donnée collectées .



Echantillons 01 **Echantillons 02** **Echantillons 03**

Figure 4.2 : Représentation de la polarisation de l'ensemble d'échantillons

4.4.2 Prétraitement :

Généralement l'utilisateur de Twitter écrit des abréviations, des émoticons, et des argots pour exprimer ses opinions et ses sentiments ,donc Par conséquence une étape de prétraitement est indispensable. dont le but de cette étape est du nettoyer les tweets et leur rendre le plus proche possible à un langage formel et prêt à utiliser.

Dans ce qui suit nous allons présenter la procédure de prétraitement :

1- Au début nous avons commencé par le filtrage de tweets, on ne considerant que ceux écrit en langue anglaise. Car un corpus de différents langages est un corpus qui contient du bruit.

2- Ensuit les mots sont séparés avec la fonction tokenization .
3- Supprimer les identifiants des utilisateurs (USER) : nous avons utilisé l'expression régulière '@[\w]+'pour détecter les mots qui représentent les identifiants des utilisateurs Twitter dont le signe il le faut détecter.
4- Nous avons supprimé les liens web (URL) : et utilisé l'expression régulière '((www \ .[b\ s]+)|(https? : //[b\ s]+))' pour d´détecter les liens des sites cité dans le tweet.
5- Supprimer les Hashtags (TAG) : nous avons utilisé l'expression régulière r '#([b\ s]+)' pour d´détecter les mots clé (sur des sujet précis) dans le tweet.
6- Supprimer les chiffres : il faut supprimer les chiffres qui n'ont aucun impact sur la classification.
7- Eliminer les commandes VIA, RT : Twitter possède son propre vocabulaire et fonctions, il y'a les commande VIA et RT indique que le tweet a été rediffusé par un autre utilisateur, nous les avons éliminé à cause de son influence négligeable sur la classification .
8- Eliminer les ponctuations : les utilisateurs utilisent dans leurs tweets beaucoup de ponctuations qui n'ont pas une importance dans notre classification, donc il a été mieux de les éliminer dans cette phase.
9- Supprimer les mots vides (Stop-words), nous avons utilisé les mots vides en langue Anglaise prédéfinies dans le package nltk.corpus.

4.4.3 Remplacer les mots :

Pour des résultats plus précis, nous recourons dans notre travail à remplacer les abréviations et les mots que l'utilisateur écrit parfois de manière particulière afin de calculer leur vraie valeur dans la phase d'apprentissage. Pour cela, nous avons crée un dictionnaire qui contient les plus fréquemment mots utilisés dans les commentaires et leurs écrits attendus. Voici quelques exemples :

Mots	Remplacement
Allahu , alah, ilah, rab	Allah
American ,amirika, amirica	Usa
Idlib ,deir ,damascu, deirezzor ,alleppo	Syria

Tableau 4.2 : Exemple Des Abréviations.

séparer les mots

19	of the week! https://twitter.com/ uspended	['the', 'word', 'of', 'the', 'week', 'https…
20	ıjir2 @jundhullah14 nothing like Ji…	['', 'ironmuhajir2', 'jundhullah14', 'nothi…
21	:tate now just 8 kilometres from Gr…	['', 'islamicstate', 'now', 'just', '8', 'k…

Supprimer des utilisateurs

@__Kebab_ayran__ @Raheelk what is this	['…	
@__Kebab_ayran__ your chatsecure may be not working i tried many time	['…	chatsecure workingtried time
@_croxetti_ @11_HaqqPrevails @ProJ32 @Somaliyyah then who is that one person?	['…	person
@_croxetti_ @11_HaqqPrevails @ProJ32 @Somaliy…	['…	blame akhi human inteligence maybe know better

Supprimer les liens web

@1436RpG: Coming soon... Wilayat Kirkuk https:// twitter.com/account/suspended		Coming soon Wilayat Kirkuk
@AK47_PK: Coming soon... Wilayat Kirkuk https:// twitter.com/account/suspended		Coming soon Wilayat Kirkuk
@Mustafaklashin7: Correction: while ghouta is besi…		Correction ghouta besieged allloush eating inrestaurant Amman Jordan
(2) https://twitter.com/account/suspended	['…	

Supprimer les Hashtags

RT @Paradoxy13: Explosions heard in the Baramkeh area in #Damascus	['r…	Explosions heard Baramkeh area
RT @osamahfakih: Explosions of weapon depots in …	['r…	Explosions weapon depots Alhafa rocking city led coalition airstrikes

Supprimer les chiffres

I have 80 new followers from India. Pak…	['i…	I new followers India Pakistan week See
I have 43 new followers from Indonesia.…	['i…	I new followers Indonesia week See
I have 35 new followers from Maldives. …	['i…	I new followers Maldives week See

Eliminer les commandes VIA, RT

RT @MuslimPrisoners: Bothers & Sisters plz …	['rt', 'muslimprisoners', 'bothers', 'siste…	Bothers Sisters plz determination If haters trier banning Abu Haleemah stop sho…
RT @3bdUlkaed6r: Russian troops 'fighting a…	['rt', '3bdulkaed6r', 'russian', 'troops', …	Russian troops fighting alongside Assads army Syrian rebels
RT @syrializer: #Syria: ISIS Destroys More …	['rt', 'syrializer', 'syria', 'isis', 'dest…	ISIS Destroys More Ancient Artifacts In

Eliminer les ponctuations

La hawla wal? quwwata ill? billah. AQAP betrayed from their own al	La hawla wal quwwata ill billah AQAP betrayed al
ly the tagh?t H?d?. abt 100 killed. tens captured & injured by Hout	ly taght Hd abt killed tens captured injured Hout

Supprimer les mots vides (Stop-words)

RT @Paradoxy13: Explosions heard in the Baramkeh area in #Damascus	['r…	Explosions heard Baramkeh area
RT @osamahfakih: Explosions of weapon depots in …	['r…	Explosions weapon depots Alhafa rocking city led coalition airstrikes

Figure 4.3 : Tweets avant et après le prétraitement .

4.4.4 Calcul TF- IDF :

Dans la recherche d'information, TF-IDF , abréviation de fréquence de document à fréquence , est une statistique numérique destinée à refléter l'importance d'un mot pour un document dans une collection ou un corpus. Il est souvent utilisé comme facteur de pondération dans les recherches d'extraction d'information, l'exploration de texte et la modélisation d'utilisateur. La valeur de Tf-idf augmente proportionnellement au nombre de fois qu'un mot apparaît dans le document et est compensée par la fréquence du mot dans le corpus, ce qui aide à ajuster le fait que certains mots apparaissent plus fréquemment en général. Tf-idf est l'un des systèmes de pondération de terme les plus populaires aujourd'hui; 83% des systèmes de recommandation textuels dans les bibliothèques numériques utilisent tf-idf. Les variations du schéma de pondération de Tf-idf sont souvent utilisées par les moteurs de recherche comme un outil central de notation et de classement de la pertinence d'un document en fonction d'une requête de l'utilisateur. Tf-idf peut être utilisé avec succès pour le filtrage des mots d'arrêt dans divers domaines, y compris la synthèse de texte et la classification. [55]

4.4.5 Expériences et résultats :

Notre démarche d'analyse de sentiments s'inscrit dans l'approche d'apprentissage automatique supervisé. Nous avons utilisé les cinq algorithmes d'apprentissage (SVM , NB, RF,DT,KNN) qui sera utilisé dans l'étape de prédiction. concernant le coté implémentation du package Sklearn mentionné dans la figure 5.1 .

A - Préparation de corpus :

Tant que nous utilisons la méthode d'apprentissage supervisé, nous divisons le corpus en deux parties, 80% pour la formation et 20% pour les tests.

Pour chaque test nous utilisé la fonction Train_test_split() pour crée les groupes d'entrainement et de test comme ressortir dans le tableau 4.3 :

Nombre de Tweets / Echantillons	Data pour l'entraînement (80%)	Data pour le test (20%)
01	16440	4110
02	2417	605
03	2415	604

Tableau 4.3 : les échantillons traitées par la fonction Train_test_split() .

B - Constat 01 :

Nous avons réalisé la premier test sur l'échantillons 01 . Le tableau 4.4 présente les résultats de la classification et une discussion de test.

Test	échantillon	Classificateur	Résultats (Accuray)	Discussion
01	- Le dataset n'est pas équilibrée elle contient 18 % des tweets négatif et 21 % positive le et le reste contient les neutres(61 %) - l'ensemble de test est composé de 4110 tweets répartir comme suit : - Négative 695 - Positive 908 - Neutre 2507	RF	0.926	Le meilleur résultat de ce test est de 92% donné par le RF et le pire résultat de ce test est de 64% donné par le KN
		DT	0.887	
		SV	0.877	
		NB	0.810	
		KN	0.641	

Tableau 4.4 : Résultats de Premier Test .

La graphe suivant montre la différence des ratios obtenus pour chaque classificateur.

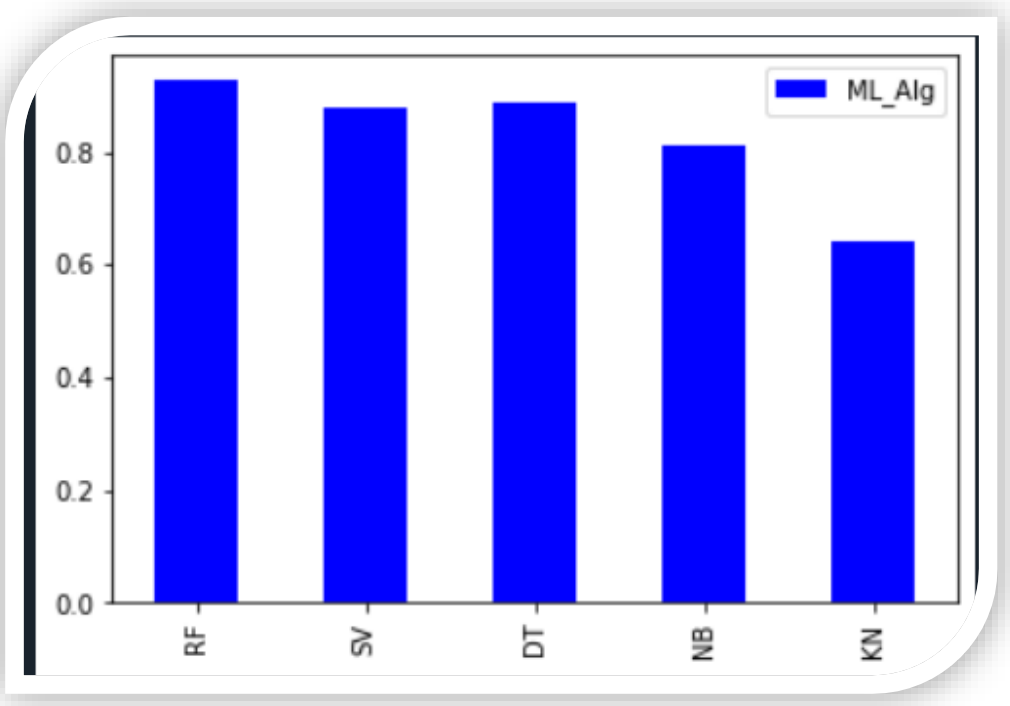


Figure 4.4 : Description d'accuray obtenu pour chaque classificateur .

Selon les resultats du tableau il est clair que la meilleur classificateur est le RF sur le DATASET Non équilibré.

C- Constat 02 :

Dans la 2éme constat Nous avons réalisé plusieurs tests sur les échantillons 02 et 03. Le tableau 4.5 illustre les résultats de la classification pour chaque test.

Test	échantillon	Classificateur	Résultats (Accuray)	Discussion	
01	Le dataset équilibrée elle contient 33 % des tweets négatif et 33 % positive le et le reste contient les neutres (34 %) l'ensemble de test est composé de 605 tweets répartir comme suit : - Négative 217 - Positive 191 - Neutre 197	RF	0.795	Le meilleur résultat de ce test est de 80% donné par le DT et le pire résultat de ce test est de 33% donné par le KN	Sans utilisé la fonction du remplacement
		DT	0.801		
		SV	0.773		
		NB	0.737		
		KN	0.337		
02	Le dataset non équilibrée elle contient 49 % des tweets négatif et 31 % positive le et le reste contient les neutres(20 %) - l'ensemble de test est composé de 604 tweets répartir comme suit : - Négative 318 - Positive 178 - Neutre 108	RF	0.781	Le meilleur résultat de ce test est de 79% donné par le SV et le pire résultat de ce test est de 22% donné par le KN	
		DT	0.754		
		SV	0.794		
		NB	0.746		
		KN	0.223		
03	Le dataset équilibrée elle contient 33 % des tweets négatif et 33 % positive le et le reste contient les neutres(34 %) l'ensemble de test est composé de 605 tweets répartir comme suit : - Négative 217 - Positive 191 - Neutre 197	RF	0.778	Le meilleur résultat de ce test est de 81% donné par le DT et le pire résultat de ce test est de 33% donné par le KN	en utilisé la fonction du remplacement
		DT	0.818		
		SV	0.765		
		NB	0.732		
		KN	0.338		

04	Le dataset non équilibrée elle contient 49 % des tweets négatif et 31 % positive le et le reste contient les neutres(20 %) l'ensemble de test est composé de 604 tweets répartir comme suit : - Négative 318 - Positive 178 - Neutre 108	RF	0.786	Le meilleur résultat de ce test est de 80% donné par le SV et le pire résultat de ce test est de 23% donné par le KN	
		DT	0.763		
		SV	0.799		
		NB	0.743		
		KN	0.235		

Tableau 4.5 : : Analyse des resultats.

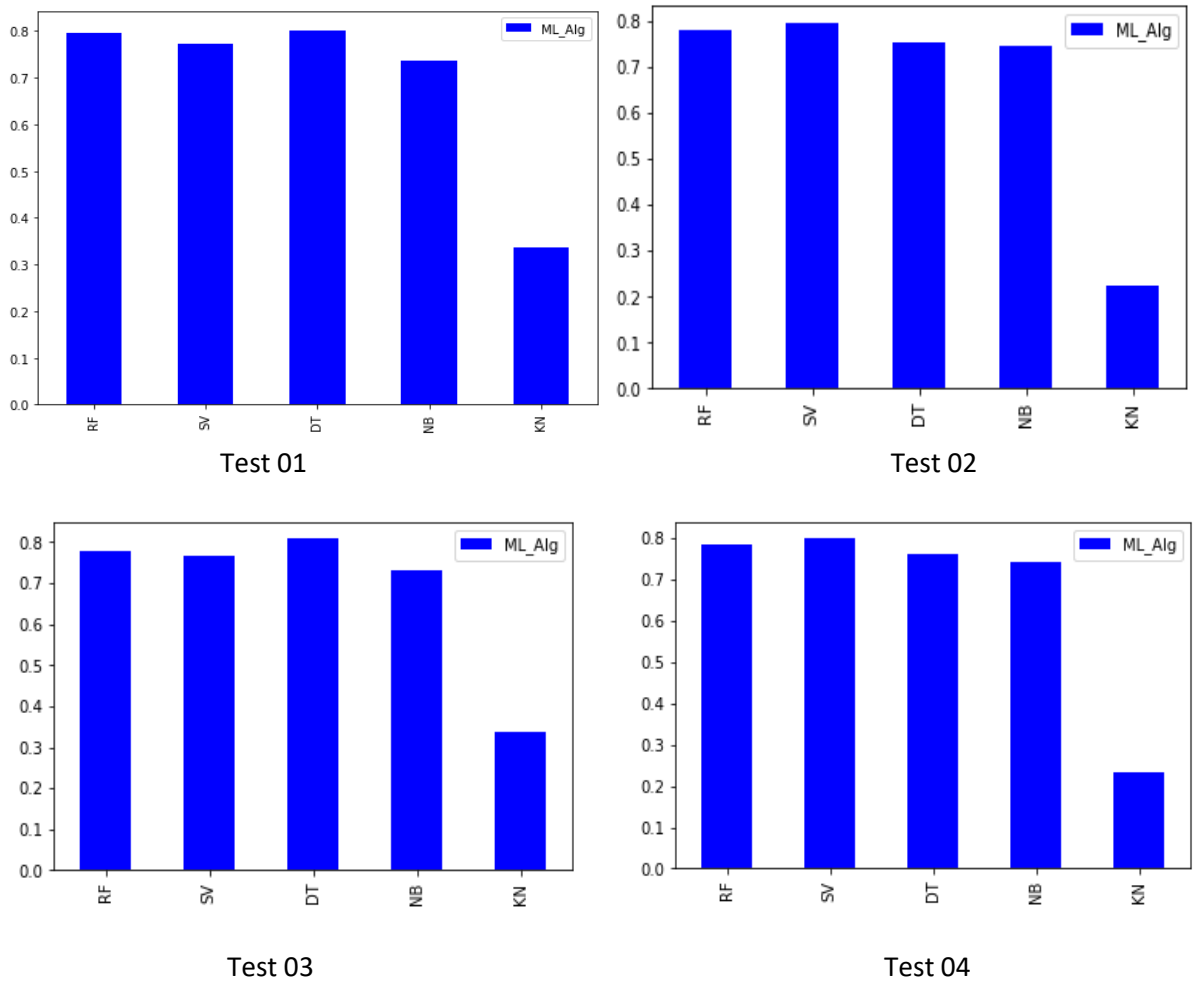


Figure 4.5 : Description du pourcentage obtenu pour chaque classificateur

4.4.6 Analyse et discussion :

Selon les résultats du constats Nous concluons que :

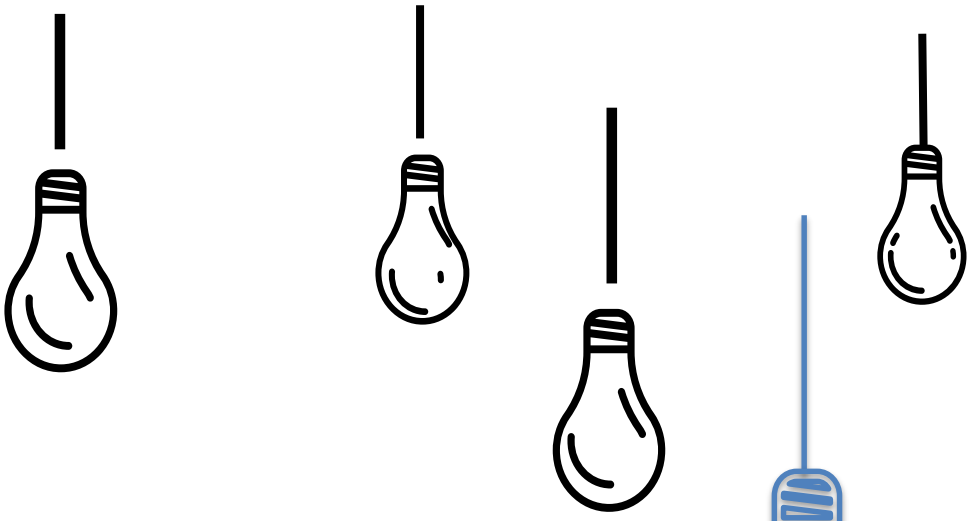
- Le meilleur résultat dans tous les tests était le 92% que RF a donné avec le premier constat s'applique au DATASET non équilibré (20000 Tweets) , Contrairement au reste des tests, il a donné moins de résultats avec moins de nombre des tweets .
- Le meilleur résultat encore a été donné par DT dans le test 1 et 3 sur DATASET équilibré de 3022 tweets. Et le SVM aussi dans le test 2 et 4 sur DATASET non équilibré de 3019 Tweets .
- Les algorithmes de classification conservent leurs valeurs les plus élevées, que la substitution de vocabulaire soit appliquée ou non.
- Le pire résultat de tous les tests a été donné par le KNN au pourcentage entre [22% - 64 %].
- Nous concluons que chaque classificateur augmenter son pourcentage avec l'augmentation de nombre de dataset c'est ta dire la quantité de **Tweets diversifiées** d'entrainement joue un grand rôle dans l'amélioration du résultat.

4.5 Conclusion :

Dans ce chapitre, nous avons analysé un ensemble de données une fois équilibrées et une autre fois déséquilibrées en termes de nombre de commentaires diverse tel que L'équilibré contenait près de 3022 textes et le déséquilibré contenait parfois 20550 et en autre 3019 Tweets

Pour chaque test , nous utilisons les outils d'apprentissage supervisé sur les datasets pour obtenu un meilleur modèle d'apprentissage qui permet du classifier chaque tweet selon leur polarité , positive , négative ou neutre .

Les expérimentations effectuées montre que les algorithmes d'apprentissage ayant des meilleur résultats à chaque fois qui'il entrainer a nouvelles dataset ou Nous fournissons des données améliorées et diversifiées pour l'entainement sans **les exagérer** pour éviter le sur-apprentissage (Overffiting).



Chapitre 5 :

Implimentation

5. L’Implimentation

5.1 Introduction

Avant d’entrer dans les détails, nous d´écrivons d’abord les aspects pratiques relatifs à la mise en œuvre d’application de Classification des opinions. Alors, nous parlerons de Python et des libraires, en particulier les libraires d'apprentissage automatique, puis les étapes du programme les plus importants .

5.1.1 Environnement de Travail

D’abord, nous donnons une description de l’environnement de travail :

A- Environnement matériel :

Afin de mener notre expérimentation et évaluation, nous avons utilisé un PC marque LENOVO, équipé d’un processeur Intel® Core5(TM) ™ i7-8550U cadenc´e par une horloge d’une fréquence de 1,80 GHz, avec 8 GO Octets de RAM, un système d'exploitation Windows 10 .

B- Environnement logiciel

B.1 python :

Nous avons utilisé le langage de programmation Python. C’est parce qu'elle un langage de programmation portable, dynamique, extensible, gratuit, qui permet (sans l’imposer) une approche modulaire et orientée objet de la programmation. Python est d´eveloppé depuis 1989 par Guido van Rossum et de nombreux contributeurs b´enévoles. Il prend en charge les modules et les packages, ce qui encourage la modularité du programme et la réutilisation du code. L'interpréteur Python et la bibliothèque standard étendue sont disponibles sous forme source ou binaire sans frais pour toutes les principales plates-formes, et peuvent être librement distribués.

B.2 Editeur Spyder :

Nous avons utilisé l’environnement de développement Spyder (Scientific PYthon Development EnviRonment)qui est un IDE MATLAB-like pour le calcul scientifique avec Python. Il permet d'éditer, d'exécuter et de déboguer du code dans un seul environement. Spyder fournit entre autres :

- Un éditeur de code efficace avec coloration syntaxique, permettant de l'introspection

dynamique de code, intégré au débogueur de Python.

- Un exploreur de variables, une invite de commande IPython
- Variable exploré, IPython command prompt.
- Une documentation intégrée.

B.3 Les bibliothèques de Python :

La bibliothèque est une ensemble de fichiers de code et de fichiers d'aide que vous pouvez inclure dans votre code qu'un groupe de développeurs a écrit et publié pour le rendre disponible à tout le monde et le but des bibliothèques de programmation est de faciliter la mise en œuvre de certaines choses dans n'importe quelle application en premier lieu en plus de donner à l'application des capacités supplémentaires. Il ne peut le faire qu'après avoir utilisé ces bibliothèques .

Et dans notre travail nous avons utilisé les packages suivants:

- Package CSV : CSV (Comma Separated Values) La bibliothèque csv fournit des fonctionnalités de lecture et d'écriture dans des fichiers CSV. Conçu pour fonctionner directement avec les fichiers CSV générés par Excel, il est facilement adapté pour fonctionner avec une variété de formats CSV. La bibliothèque csv contient des objets et d'autres codes pour lire, écrire et traiter des données depuis et vers des fichiers CSV.
- Package re : (Regular expressions) Ce module fournit des op´erations correspondant aux expressions r´eguli`eres. Package numpy : numpy (NUMeric Python) est une bibliothèque numérique apportant le support efficace de larges tableaux multidimensionnels, et de routines mathématiques de haut niveau (algèbre linéaire, statistiques, .. etc.).
- Package pandas : Pandas est une bibliothèque Python open source permettant de manipuler des structures de données hautes performances et faciles à utiliser ainsi que des outils d'analyse de données pour le langage de programmation Python.
- Package Nltk : Nltk (Natural Language Toolkit) est une plate-forme pour la cr´eation de programmes Python pour travailler avec des donn´ees de langage humain.
- Package Sklearn : est un module en Python pour l'apprentissage automatique.
- Gensim : Est une bibliothèque gratuite Python pour l'extraction automatique des sujets sémantiques des documents.

5.1.2 Les étapes d'exécution (exemples de codes sources) :

Nous allons présenter quelques exemples de codes sources.

1) La déclaration et l'importation des bibliothèques nécessaires :

```
In[Importation]
# Importation general bib
from nltk.corpus import stopwords
import pandas as pd
import numpy as np
from textblob import TextBlob
import nltk
import string
from gensim.parsing.preprocessing import remove_stopwords
from nltk.tokenize import word_tokenize
from nltk.stem.porter import PorterStemmer
import re
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
import matplotlib.pyplot as plt

# Importation des algorithmes ML de sklearn

from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import  MultinomialNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn import svm, tree, model_selection
```

Figure 5.1 : La déclaration et l'importation des bibliothèques nécessaires .

2) Importation des données (DATASET) :

```
# In[open_file]

df = pd.read_csv('C:\\Users\\Aures\\Desktop\\Application_ML\\1.csv')
df.head()
```

Figure 5.2 : Importation des données .

3) Prétraitement le dataset :

Dans ce qui suit nous allons présenter la procédure de prétraitement suivie dans notre travail, dont le but de nettoyer les tweets et leur rendre le plus proche possible à un langage formel.

A- D'abord nous avons commencé par séparer les mots avec la fonction tokenization et calculer leur nombres .

```
def tokenization(text):
    text = re.split('\W+', text)
    return text


df['Tweet_tokenized'] = df['Tweets'].apply(lambda x: tokenization(x.lower()))
df.head()

for index in df.index:
    a = df.loc[index, 'Tweet_tokenized']
    df.loc[index, 'Tweets_size_Before'] = len(a)
```

Figure 5.3: Séparer et compter le nombre de mots .

- Le résultat :

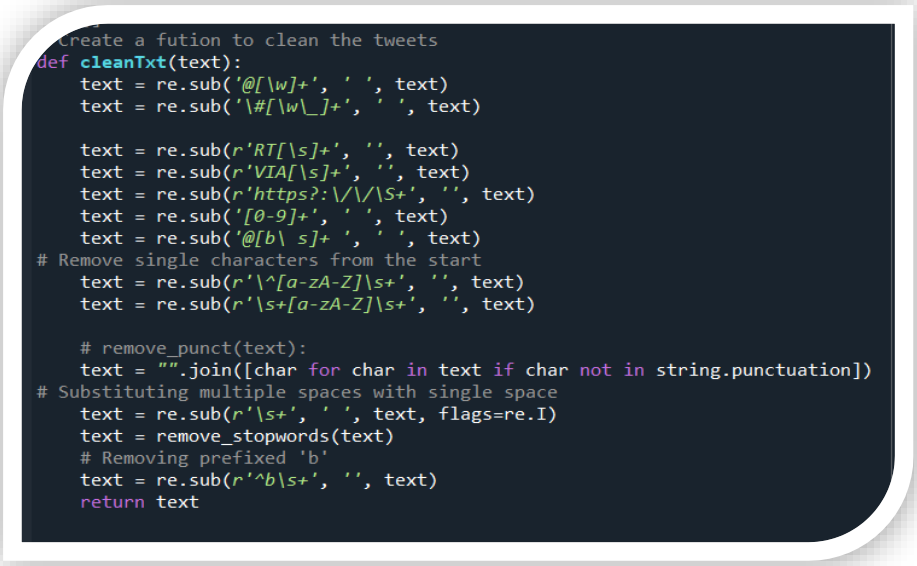
	Tweets	Tweet_tokenized	Tweets_size_Before
0	english translation: 'a message to the truthful in syria - sheikh ab…	['english', 'translation', 'a', 'message', 'to', 'the', 'truthful', 'in', 'syria'…	23
1	english translation: sheikh fatih al jawlani 'for the people of inte…	['english', 'translation', 'sheikh', 'fatih', 'al', 'jawlani', 'for', 'the', 'peo…	24
2	english translation: first audio meeting with sheikh fatih al jawlan…	['english', 'translation', 'first', 'audio', 'meeting', 'with', 'sheikh', 'fatih'…	20
3	english translation: sheikh nasir al wuhayshi (ha) leader of aqap: '…	['english', 'translation', 'sheikh', 'nasir', 'al', 'wuhayshi', 'ha', 'leader', '…	24
4	english translation: aqap: 'response to sheikh baghdadis statement 'although the disbelievers dislike it.' https://justpaste.it/i4xj	['english', 'translation', 'aqap', 'response', 'to', 'sheikh', 'baghdadis', 'stat…	17
5	the second clip in a da'wah series by a soldier of jn: video link :h…	['the', 'second', 'clip', 'in', 'a', 'da', 'wah', 'series', 'by', 'a', 'soldier',…	30
6	english transcript : oh murabit! : https://justpaste.it/ohmurabit https://twitter.com/account/suspended	['english', 'transcript', 'oh', 'murabit', 'https', 'justpaste', 'it', 'ohmurabit', 'https', 'twitter', 'com', 'account', 'suspended']	13
7	english translation: 'a collection of the words of the u'lama regard…	['english', 'translation', 'a', 'collection', 'of', 'the', 'words', 'of', 'the', …	22
8	aslm please share our new account after the previous one was suspended. @khalidmaghrebi @seifulmaslul123 @cheerleadunited	['aslm', 'please', 'share', 'our', 'new', 'account', 'after', 'the', 'previous', 'one', 'was', 'suspended', 'khalidmaghrebi', 'seifulmaslul123', 'cheerleadunited']	15
9	english translation: aqap statement regarding the blessed raid in fr…	['english', 'translation', 'aqap', 'statement', 'regarding', 'the', 'blessed', 'r…	19
10	@khalidmaghrebi @seifulmaslul123 @cheerleadunited @ibnnabih1	['', 'khalidmaghrebi', 'seifulmaslul123', 'cheerleadunited', 'ibnnabih1']	5
11	new link after previous one taken down:aqap-'the faces have been brightened' -regarding the blessed attack in france	['new', 'link', 'after', 'previous', 'one', 'taken', 'down', 'aqap', 'the', 'face…	20
12	english translation- sheikh abu hasan al kuwaiti (ha) advice of nabi…	['english', 'translation', 'sheikh', 'abu', 'hasan', 'al', 'kuwaiti', 'ha', 'advi…	27
13	sheikh abu hassan al kuwaiti (ha): advice of nabi saw to the mujahid…	['sheikh', 'abu', 'hassan', 'al', 'kuwaiti', 'ha', 'advice', 'of', 'nabi', 'saw',…	22
14	@IbnNabih1 @MuwMedia @Dawlat_islam7 Not translating it either.	['', 'ibnnabih1', 'muwmedia', 'dawlat_islam7', 'not', 'translating', 'it', 'either', '']	9
15	@Skxx51 If all of you remember his tweets from his first acc he used…	['aslm', 'anybody', 'translating', 'the', 'new', 'jn', 'video', 'will', 'translat…	887
16	@3diyah1000: Most popular attacks in Somalia last years : #dayniile attack. #Lego attack. #Jannale attack. #Yurkud attack. #Somalia	['', '3diyah1000', 'most', 'popular', 'attacks', 'in', 'somalia', 'last', 'years'…	18
17	US-trained Division 30 has entered Marea to fight the Islamic State. JN has reportedly allowed this (?!) http://t.co/xC7g/s/6KhJ.	['us', 'trained', 'division', '30', 'has', 'entered', 'marea', 'to', 'fight', 'th…	24
18	@Tawheed1_IS but one is important to live when the other one is not!	['', 'tawheed1_is', 'but', 'one', 'is', 'important', 'to', 'live', 'when', 'the', 'other', 'one', 'is', 'not', '']	15
19	the word of the week! https://twitter.com/account/suspended	['the', 'word', 'of', 'the', 'week', 'https', 'twitter', 'com', 'account', 'suspended', '']	11
20	@IronMuhajir2 @jundhullah14 nothing like Jihad. God give us the help and power to do our FARRIDA	['', 'ironmuhajir2', 'jundhullah14', 'nothing', 'like', 'jihad', 'god', 'give', 'us', 'the', 'help', 'and', 'power', 'to', 'do', 'our', 'farrida', '']	18
21	#IslamicState now just 8 kilometres from Great Umayyad Mosque. March…	['', 'islamicstate', 'now', 'just', '8', 'kilometres', 'from', 'great', 'umayyad'…	20
22	BREAKING Nigeria Islamic State advance on Lagos the business capital…	['breaking', 'nigeria', 'islamic', 'state', 'advance', 'on', 'lagos', 'the', 'bus…	26
23	#WilayatAlFallujah New Video : The Knights of Victory (Part 3) Coming Soon.....InshaAllah https://twitter.com/account/suspended	['', 'wilayatalfallujah', 'new', 'video', 'the', 'knights', 'of', 'victory', 'par…	19
24	@Jazrawi_Missk: Apostate FSA group declares hiwar this area circled as a military zone. https://twitter.com/account/suspended	['', 'jazrawi_missk', 'apostate', 'fsa', 'group', 'declares', 'hiwar', 'this', 'a…	20

Format Resize Background color Column min/max

Figure 5.4 : listes des tweets et leur nombre de mots .

B- Nettoyé les Tweets comme suit :

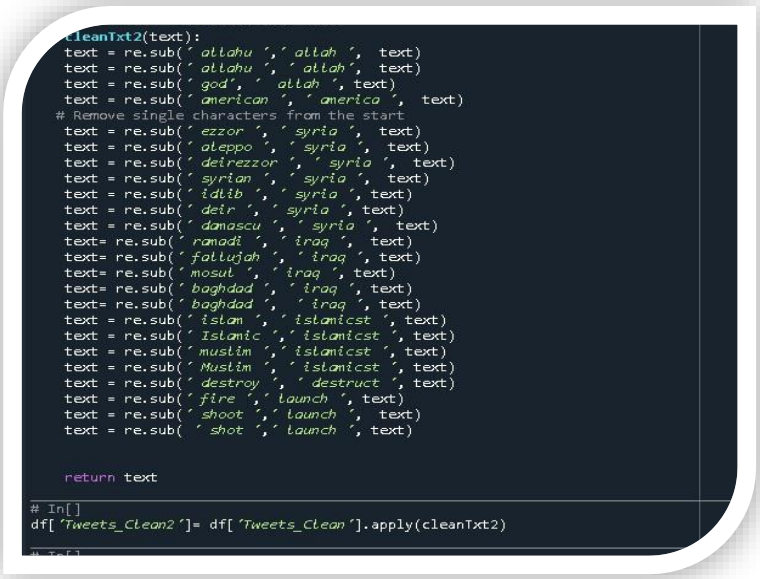
- le fonction re.sub() : pour supprimé les chiffres et les expression régulières (@ # 'http'…ect) et les commandes VIA RT et ponctuation et hashtag …ect qui n'ont aucun impact sur la classification avec
- la fonction remove_stopword() pour Supprimer les mots vides (Stop-words), nous avons utilisé les mots prédéfinies dans le package nltk.corpus.



```
Create a futions to clean the tweets
def cleanTxt(text):
    text = re.sub('@[\w]+', ' ', text)
    text = re.sub('\#[\w\_]+', ' ', text)

    text = re.sub(r'RT[\s]+', '', text)
    text = re.sub(r'VIA[\s]+', '', text)
    text = re.sub(r'https?:\/\/\S+', '', text)
    text = re.sub('[0-9]+', ' ', text)
    text = re.sub('@[b\ s]+ ', ' ', text)
# Remove single characters from the start
    text = re.sub(r'\^[a-zA-Z]\s+', '', text)
    text = re.sub(r'\s+[a-zA-Z]\s+', '', text)

    # remove_punct(text):
    text = "".join([char for char in text if char not in string.punctuation])
# Substituting multiple spaces with single space
    text = re.sub(r'\s+', ' ', text, flags=re.I)
    text = remove_stopwords(text)
    # Removing prefixed 'b'
    text = re.sub(r'^b\s+', '', text)
    return text
```

Figure 5.5 : fonction qui Nettoyer les Tweets.

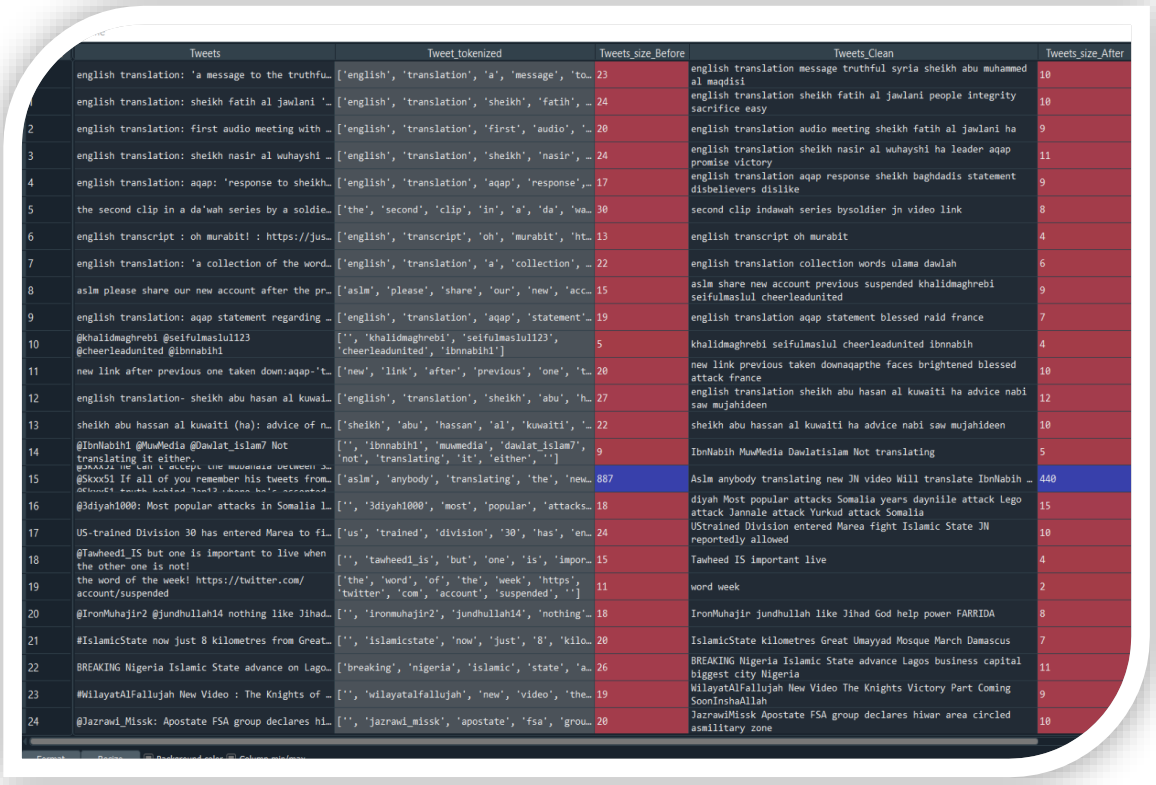
C- .Remplacer les abréviation :



```
CleanTxt2(text):
    text = re.sub(' allahu ',' allah ', text)
    text = re.sub(' allahu ', ' allah', text)
    text = re.sub(' god', ' allah ', text)
    text = re.sub(' american ', ' america ', text)
# Remove single characters from the start
    text = re.sub(' ezzor ', ' syria ', text)
    text = re.sub(' aleppo ', ' syria ', text)
    text = re.sub(' deirezzor ', ' syria ', text)
    text = re.sub(' syrian ', ' syria ', text)
    text = re.sub(' idlib ', ' syria ', text)
    text = re.sub(' deir ', ' syria ', text)
    text = re.sub(' damascu ', ' syria ', text)
    text= re.sub(' ramadi ', ' iraq ', text)
    text= re.sub(' fallujah ', ' iraq ', text)
    text= re.sub(' mosul ', ' iraq ', text)
    text= re.sub(' baghdad ', ' iraq ', text)
    text= re.sub(' baghdad ', ' iraq ', text)
    text = re.sub(' islam ', ' islamicst ', text)
    text = re.sub(' Islamic ',' islamicst ', text)
    text = re.sub(' muslim ',' islamicst ', text)
    text = re.sub(' Muslim ', ' islamicst ', text)
    text = re.sub(' destroy ', ' destruct ', text)
    text = re.sub(' fire ',' launch ', text)
    text = re.sub(' shoot ',' launch ', text)
    text = re.sub( ' shot ',' launch ', text)


    return text

# In[]
df['Tweets_Clean2']= df['Tweets_Clean'].apply(cleanTxt2)
```

Figure 5.6 : fonction qui remplacer les synonyms.

- Le résultats :



	Tweets	Tweet_tokenized	Tweets_size_Before	Tweets_Clean	Tweets_size_After
0	english translation: 'a message to the truthfu…	['english', 'translation', 'a', 'message', 'to…	23	english translation message truthful syria sheikh abu muhammed al maqdisi	10
1	english translation: sheikh fatih al jawlani '…	['english', 'translation', 'sheikh', 'fatih', …	24	english translation sheikh fatih al jawlani people integrity sacrifice easy	10
2	english translation: first audio meeting with …	['english', 'translation', 'first', 'audio', '…	20	english translation audio meeting sheikh fatih al jawlani ha	9
3	english translation: sheikh nasir al wuhayshi …	['english', 'translation', 'sheikh', 'nasir', …	24	english translation sheikh nasir al wuhayshi ha leader aqap promise victory	11
4	english translation: aqap: 'response to sheikh…	['english', 'translation', 'aqap', 'response',…	17	english translation aqap response sheikh baghdadis statement disbelievers dislike	9
5	the second clip in a da'wah series by a soldie…	['the', 'second', 'clip', 'in', 'a', 'da', 'wa…	30	second clip indawah series bysoldier jn video link	8
6	english transcript : oh murabit! : https://jus…	['english', 'transcript', 'oh', 'murabit', 'ht…	13	english transcript oh murabit	4
7	english translation: 'a collection of the word…	['english', 'translation', 'a', 'collection', …	22	english translation collection words ulama dawlah	6
8	aslm please share our new account after the pr…	['aslm', 'please', 'share', 'our', 'new', 'acc…	15	aslm share new account previous suspended khalidmaghrebi seifulmaslul cheerleadunited	9
9	english translation: aqap statement regarding …	['english', 'translation', 'aqap', 'statement'…	19	english translation aqap statement blessed raid france	7
10	@khalidmaghrebi @seifulmaslul123 @cheerleadunited @ibnnabih1	['', 'khalidmaghrebi', 'seifulmaslul123', 'cheerleadunited', 'ibnnabih1']	5	khalidmaghrebi seifulmaslul cheerleadunited ibnnabih	4
11	new link after previous one taken down:aqap-'t…	['new', 'link', 'after', 'previous', 'one', 't…	20	new link previous taken downaqapthe faces brightened blessed attack france	10
12	english translation- sheikh abu hasan al kuwai…	['english', 'translation', 'sheikh', 'abu', 'h…	27	english translation sheikh abu hasan al kuwaiti ha advice nabi saw mujahideen	12
13	sheikh abu hassan al kuwaiti (ha): advice of n…	['sheikh', 'abu', 'hassan', 'al', 'kuwaiti', '…	22	sheikh abu hassan al kuwaiti ha advice nabi saw mujahideen	10
14	@IbnNabih1 @MuwMedia @Dawlat_islam7 Not translating it either.	['', 'ibnnabih1', 'muwmedia', 'dawlat_islam7', 'not', 'translating', 'it', 'either', '']	9	IbnNabih MuwMedia Dawlatislam Not translating	5
15	@Skxx51 If all of you remember his tweets from…	['aslm', 'anybody', 'translating', 'the', 'new…	887	Aslm anybody translating new JN video Will translate IbnNabih …	440
16	@3diyah1000: Most popular attacks in Somalia 1…	['', '3diyah1000', 'most', 'popular', 'attacks…	18	diyah Most popular attacks Somalia years dayniile attack Lego attack Jannale attack Yurkud attack Somalia	15
17	US-trained Division 30 has entered Marea to fi…	['us', 'trained', 'division', '30', 'has', 'en…	24	UStrained Division entered Marea fight Islamic State JN reportedly allowed	10
18	@Tawheed1_IS but one is important to live when the other one is not!	['', 'tawheed1_is', 'but', 'one', 'is', 'impor…	15	Tawheed IS important live	4
19	the word of the week! https://twitter.com/account/suspended	['the', 'word', 'of', 'the', 'week', 'https', 'twitter', 'com', 'account', 'suspended', '']	11	word week	2
20	@IronMuhajir2 @jundhullah14 nothing like Jihad…	['', 'ironmuhajir2', 'jundhullah14', 'nothing'…	18	IronMuhajir jundhullah like Jihad God help power FARRIDA	8
21	#IslamicState now just 8 kilometres from Great…	['', 'islamicstate', 'now', 'just', '8', 'kilo…	20	IslamicState kilometres Great Umayyad Mosque March Damascus	7
22	BREAKING Nigeria Islamic State advance on Lago…	['breaking', 'nigeria', 'islamic', 'state', 'a…	26	BREAKING Nigeria Islamic State advance Lagos business capital biggest city Nigeria	11
23	#WilayatAlFallujah New Video : The Knights of …	['', 'wilayatalfallujah', 'new', 'video', 'the…	19	WilayatAlFallujah New Video The Knights Victory Part Coming SoonInshaAllah	9
24	@Jazrawi_Missk: Apostate FSA group declares hi…	['', 'jazrawi_missk', 'apostate', 'fsa', 'grou…	20	JazrawiMissk Apostate FSA group declares hiwar area circled asmilitary zone	10

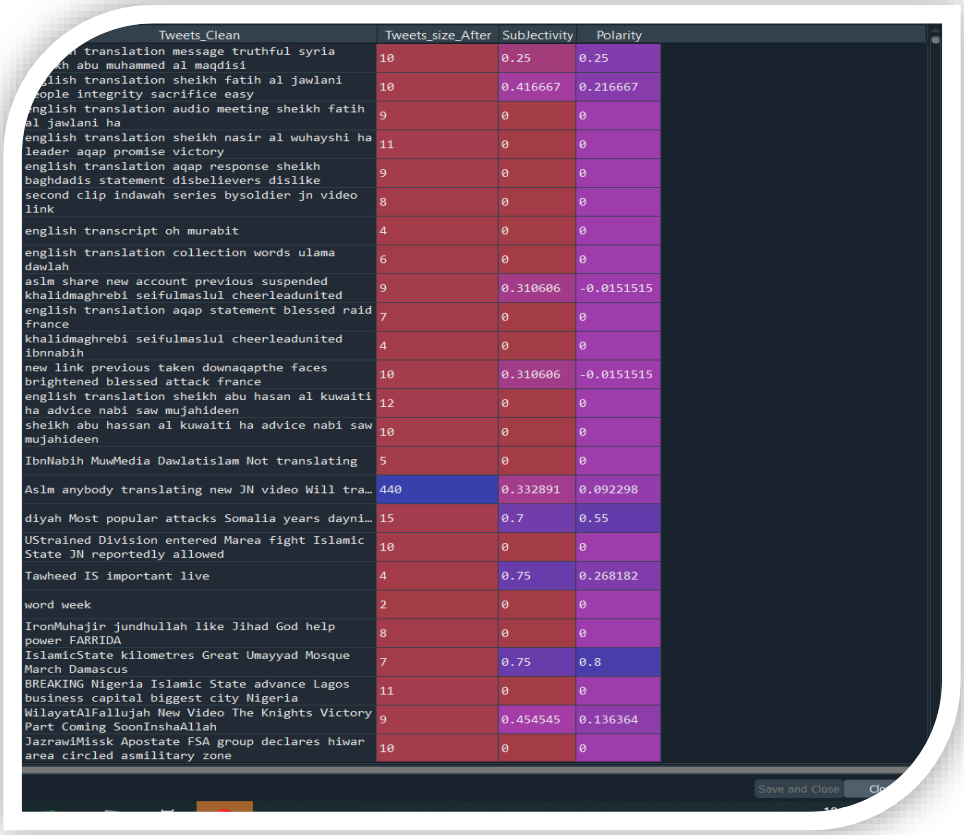
Figure 5.7 : les Tweets nettoyé (Tweet_Clean) et leur nombre de mots

4) On calculer la subjectivity et polarity de nouveau Tweets avec la fonction :

```
In[calcul sub/pol]

# Create a function to get the subjectivity
def getSubjectivity(text):
    return TextBlob(text).sentiment.subjectivity

# Create a function to get the polarity


def getPolarity(text):
    return TextBlob(text).sentiment.polarity


# Creat tow a new columns
df['SubJectivity'] = df['Tweets_Clean'].apply(getSubjectivity)
df['Polarity'] = df['Tweets_Clean'].apply(getPolarity)
```

Figure 5.8 : fonction de calcul la subjectivity et la polarity[1] .

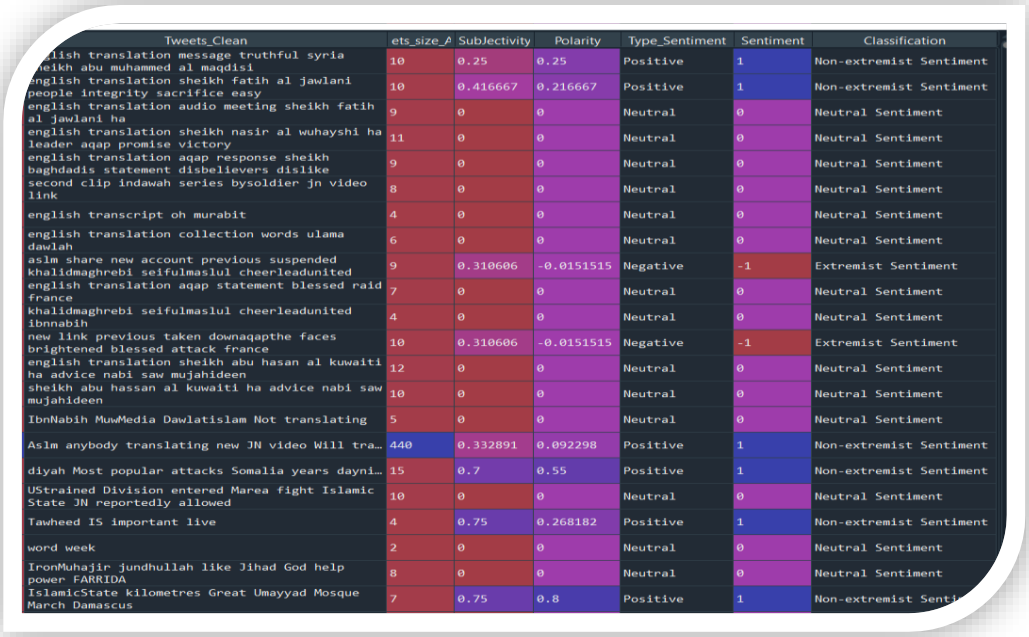
- Et les résultats :



Tweets_Clean	Tweets_size_After	SubJectivity	Polarity
translation message truthful syria kh abu muhammed al maqdisi	10	0.25	0.25
lish translation sheikh fatih al jawlani ople integrity sacrifice easy	10	0.416667	0.216667
nglish translation audio meeting sheikh fatih al jawlani ha	9	0	0
english translation sheikh nasir al wuhayshi ha leader aqap promise victory	11	0	0
english translation aqap response sheikh baghdadis statement disbelievers dislike	9	0	0
second clip indawah series bysoldier jn video link	8	0	0
english transcript oh murabit	4	0	0
english translation collection words ulama dawlah	6	0	0
aslm share new account previous suspended khalidmaghrebi seifulmaslul cheerleadunited	9	0.310606	-0.0151515
english translation aqap statement blessed raid france	7	0	0
khalidmaghrebi seifulmaslul cheerleadunited ibnnabih	4	0	0
new link previous taken downaqapthe faces brightened blessed attack france	10	0.310606	-0.0151515
english translation sheikh abu hasan al kuwaiti ha advice nabi saw mujahideen	12	0	0
sheikh abu hassan al kuwaiti ha advice nabi saw mujahideen	10	0	0
IbnNabih MuwMedia Dawlatislam Not translating	5	0	0
Aslm anybody translating new JN video Will tra…	440	0.332891	0.092298
diyah Most popular attacks Somalia years dayni…	15	0.7	0.55
UStrained Division entered Marea fight Islamic State JN reportedly allowed	10	0	0
Tawheed IS important live	4	0.75	0.268182
word week	2	0	0
IronMuhajir jundhullah like Jihad God help power FARRIDA	8	0	0
IslamicState kilometres Great Umayyad Mosque March Damascus	7	0.75	0.8
BREAKING Nigeria Islamic State advance Lagos business capital biggest city Nigeria	11	0	0
WilayatAlFallujah New Video The Knights Victory Part Coming SoonInshaAllah	9	0.454545	0.136364
JazrawiMissk Apostate FSA group declares hiwar area circled asmilitary zone	10	0	0

Figure 5.9 : Résultats en nombre entre [-1..1]

A partir de cette résultats en utilise la fonction dans la figure 5.9 pour classifier le sentiment



Tweets_Clean	ets_size_A	SubJectivity	Polarity	Type_Sentiment	Sentiment	Classification
glish translation message truthful syria heikh abu muhammed al maqdisi	10	0.25	0.25	Positive	1	Non-extremist Sentiment
nglish translation sheikh fatih al jawlani people integrity sacrifice easy	10	0.416667	0.216667	Positive	1	Non-extremist Sentiment
english translation audio meeting sheikh fatih al jawlani ha	9	0	0	Neutral	0	Neutral Sentiment
english translation sheikh nasir al wuhayshi ha leader aqap promise victory	11	0	0	Neutral	0	Neutral Sentiment
english translation aqap response sheikh baghdadis statement disbelievers dislike	9	0	0	Neutral	0	Neutral Sentiment
second clip indawah series bysoldier jn video link	8	0	0	Neutral	0	Neutral Sentiment
english transcript oh murabit	4	0	0	Neutral	0	Neutral Sentiment
english translation collection words ulama dawlah	6	0	0	Neutral	0	Neutral Sentiment
aslm share new account previous suspended khalidmaghrebi seifulmaslul cheerleadunited	9	0.310606	-0.0151515	Negative	-1	Extremist Sentiment
english translation aqap statement blessed raid france	7	0	0	Neutral	0	Neutral Sentiment
khalidmaghrebi seifulmaslul cheerleadunited ibnnabih	4	0	0	Neutral	0	Neutral Sentiment
new link previous taken downaqapthe faces brightened blessed attack france	10	0.310606	-0.0151515	Negative	-1	Extremist Sentiment
english translation sheikh abu hasan al kuwaiti ha advice nabi saw mujahideen	12	0	0	Neutral	0	Neutral Sentiment
sheikh abu hassan al kuwaiti ha advice nabi saw mujahideen	10	0	0	Neutral	0	Neutral Sentiment
IbnNabih MuwMedia Dawlatislam Not translating	5	0	0	Neutral	0	Neutral Sentiment
Aslm anybody translating new JN video Will tra…	440	0.332891	0.092298	Positive	1	Non-extremist Sentiment
diyah Most popular attacks Somalia years dayni…	15	0.7	0.55	Positive	1	Non-extremist Sentiment
UStrained Division entered Marea fight Islamic State JN reportedly allowed	10	0	0	Neutral	0	Neutral Sentiment
Tawheed IS important live	4	0.75	0.268182	Positive	1	Non-extremist Sentiment
word week	2	0	0	Neutral	0	Neutral Sentiment
IronMuhajir jundhullah like Jihad God help power FARRIDA	8	0	0	Neutral	0	Neutral Sentiment
IslamicState kilometres Great Umayyad Mosque March Damascus	7	0.75	0.8	Positive	1	Non-extremist Senti

Figure 5.10 : classification des Tweets a partir de fonction [1] .

5. appliquer le modèle de pondération TF-IDF :

```
# In[18]
vectorizer = TfidfVectorizer(max_features=3000, min_df=1, max_df=1.0,
stop_words=stopwords.words('english'), sublinear_tf=True)
Features = vectorizer.fit_transform(Tweets_Cl).toarray()
# In[19]
print(vectorizer.get_feature_names())
```

Figure 5.11 : Calculer le modèle TF-IDF

6. Utilisé la fonction train_test_split() pour crée les valeurs d'entraîner et passer le test

```
# In[20]

X_train, X_test, y_train, y_test = train_test_split(
    Features, Sentiment, test_size=0.2, random_state=0)
```

Figure 5.12 : Préparation le corpus pour entrainer et testé.

7.Appliqué les cinq algorithmes sur datset :

```
# In[21]
text_classifier = RandomForestClassifier(n_estimators=200, random_state=0)
text_classifier.fit(X_train, y_train)
predictions = text_classifier.predict(X_test)
print(confusion_matrix(y_test, predictions))
print(classification_report(y_test, predictions))
print(accuracy_score(y_test, predictions))
a_RF = accuracy_score(y_test, predictions)
# In[22]
svm = SVC()
svm.fit(X_train, y_train)
predictions_svm = svm.predict(X_test)
print(confusion_matrix(y_test, predictions_svm))
print(classification_report(y_test, predictions_svm))
print(accuracy_score(y_test, predictions_svm))
a_SV = accuracy_score(y_test, predictions_svm)
# In[23]
det = DecisionTreeClassifier()
det.fit(X_train, y_train)
predictions_det = det.predict(X_test)
print(confusion_matrix(y_test, predictions_det))
print(classification_report(y_test, predictions_det))
print(accuracy_score(y_test, predictions_det))
a_DT = accuracy_score(y_test, predictions_det)
# In[24]
NB = MultinomialNB()
NB.fit(X_train, y_train)
predictions_NB = NB.predict(X_test)
print(confusion_matrix(y_test, predictions_NB))
print(classification_report(y_test, predictions_NB))
print(accuracy_score(y_test, predictions_NB))
a_NB = accuracy_score(y_test, predictions_NB)
# In[25]
KN = KNeighborsClassifier()
KN.fit(X_train, y_train)
predictions_KN = KN.predict(X_test)
print(confusion_matrix(y_test, predictions_KN))
print(classification_report(y_test, predictions_KN))
print(accuracy_score(y_test, predictions_KN))
a_KN = accuracy_score(y_test, predictions_KN)
```

Figure 5.13: Pratiquez x et y et calculez la précision.

a_DT	float64	1	0.8878345498783455
a_KN	float64	1	0.6411192214111923
a_NB	float64	1	0.8102189781021898
a_RF	float64	1	0.9262773722627737
a_SV	float64	1	0.8776155717761557

Figure 5.14: la précision du chaque algorithme d'apprentissage .

8.Utilisé la fonction wordcloud pour calculer les mots :

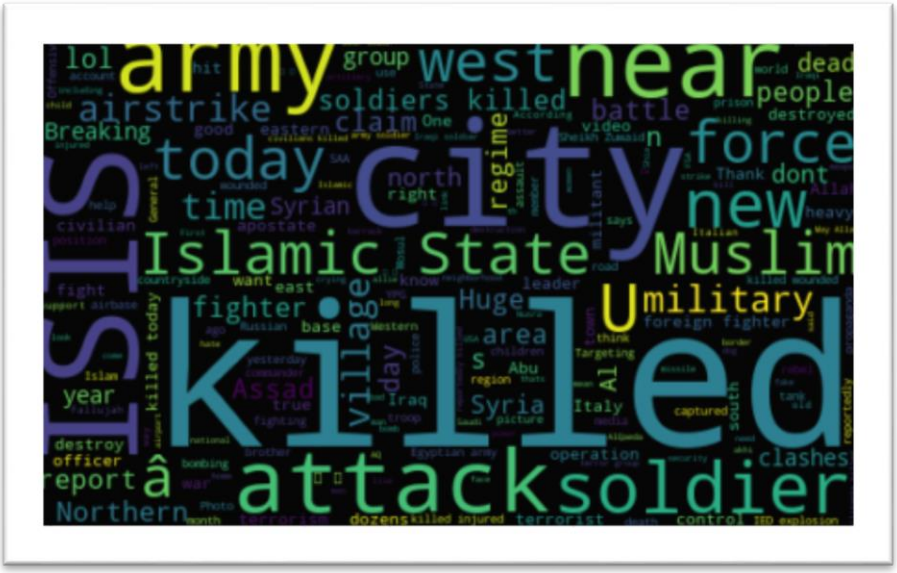


Figure 5.15 : synthétiser les mots par le wordcloud .

9 . Sauvgarder le modèle d'apprentissage

```
# In[]:
    #output a pickle file for the model
fl='Model_of_Classification.pk1'
joblib.dump(text_classifier,fl)
# out[]:
    ['Model_of_Classification.pk1']
```

Figure 5.16 : garder le modèle .

5.2 Conclusion :

Dans ce chapitre, nous avons découvert Python et ses bibliothèques les plus importantes, et nous avons appliqué cinq algorithmes d'apprentissage supervisé Ces algorithmes sont : Support Vector Machines (SVM), Tree Decision (DT), Random Forest (RF) et Naïve Bayes (NB).. sur les jeux de données et affichent leurs résultats pour obtenir un modèle.

Conclusion générale

L'objective de ce mémoire est du réaliser d'une application sous Python qui utilise une source de données (Dataset, format .csv) pour classifier des Tweets en cas l'extrémisme sur les réseaux sociaux a partir du calculer la polarités selon trois voies : positive, négative et neutre. Nous avons commencé par la définition de quelques concepts. Ensuite, nous avons met l'accent sur quatre travaux connexes, puis on a étudier l'efficacité des algorithmes d'apprentissage supervisé 'SVM, DT, RF et NB et KNN ' en créant un modèle de prédiction pour classifier les types d'extrémisme a l'aide de fonctionnalité du pondération TF-IDF qui il a fourni des résultats très encourageants comme un meilleur Accuracy obtenu par le classificateur RF(92%) quelle que soit la fonction du Remplacer les abréviations active ou pas ,

L'augmentation de la quantité de données améliore le modèle de prédiction , mais aussi il pourra donner de très bonnes prédictions sur les dataset qu'il a déjà "vues" et auxquelles il s'y est adapté , mais il prédira mal sur des données qu'il n'a pas encore vues lors de sa phase d'apprentissage donc On dit que la fonction prédictive se généralise mal. Et que le modèle souffre de "Sur-Apprentissage" ou "L'Overfitting" .

Trouver le juste modéle est le challenge de chaque data scientist lors d'un projet de Machine Learning et pour éviter l'overfitting , il existe des méthodes plus élaborée . parmi ces méthodes , il y a la validation croisée .

pas dans l'ensemble. L'augmentation des données peut amener l'algorithme à mémoriser et à ne pas apprendre.

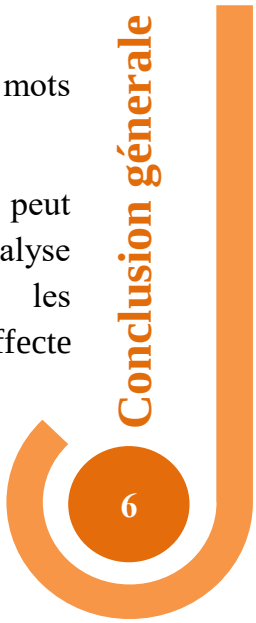
Perspectives Pour les travaux futurs, nous pouvons citer :

- L'enrichissement de liste de synonyme et abréviation par d'autres mots en langue différente fin d'obtenir des résultats bien précis.

- L'utilisation de l'analyse lexique laisse entendre que les Tweets peut prendre la direction de l'extrémisme ainsi que l'application de l'analyse sémantique latent "LSA" sur les Tweets permet d'extraire les caractéristiques de base, stylistiques et sémantiques de texte et cela affecte bien les résultats d'apprentissage .

- L'analyse par l'utilisation de la classe Mixte autre que les classes positive, négative et neutre.

-L'enrichissement d'un dictionnaire par des mots extremes de langue différente pour détecter différents types de tweets extrémistes .



Références

1) Livre .

[3] Introduction to Machine Learning The Wikipedia Guide p 10

[21] Romain Rissoan. Les réseaux sociaux Facebook. Twitter. Linkedin. Viadeo, Google+ Comprendre et maitriser ces nouveaux outils de communication, ENI. 2011

[31]Pozzi, F. A., Fersini, E., Messina, E. & Liu, B. (2016). Challenges of Sentiment Analysis in Social Networks: Livre .

[29] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. page 94.

[24] B.BENITA.et al. Lumière sur les réseaux sociaux (p14/15).

2) Mémoire

[1] Emilio FERRARA , Contagion dynamics of extremist propaganda in social networks , University of Southern California, Information Sciences Institute, USA May 2017. [

8] Rachid MIFDAL Application des techniques d'apprentissage automatique pour la prédiction de la tendance des titres financiers ,MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE ,MONTRÉAL, LE 12 NOVEMBRE 2019 ÉCOLE DE TECHNOLOGIE SUPÉRIEURE UNIVERSITÉ DU QUÉBEC

[12] Parejo pionnier, Contribution à la classification de textes et à l'extraction d'informations, une thèse pour obtenir un doctorat en sciences avec spécialisation en informatique 2012/2013 p 46.

[28] Chenni Rania Wissem , analyse des sentiments arabes en utilisant l'apprentissage en profondeur , Mémoire Présenté pour obtenir le diplôme de master académique en Informatique Parcours : Système d'information optimisation décision (SIOD), page 13

[33] Ziani Amel , la recommandation via l'analyse d'opinions Université Badji Mokhtar–Annaba 2017/2018 .

[35] M.DJELLOULI,A.TABTI, Mémoire Analyse de Contexte des Personnes Sur Les Réseaux Sociaux : Analyse des sentiments sur Twitter 2017/2018.

[50]Kamal Al-Barznji , Atanas Atanassov BIG DATA SENTIMENT ANALYSIS USING MACHINE LEARNING ALGORITHMS KAMAL AL-BARZNJI, ATANAS ATANASSOV

[51] Shakeel Ahmad, Muhammad Zubair Asghar,Fahad M. Alotaibi And Irfanullah Awan, Detection And Classification Of Social Media-Based Extremist Affiliations Using Sentiment Analysis Techniques, Research,9102 .

[52] Vivek Kumar, Manuel Mazzara , A Conjoint Application of Data Mining Techniques for Analysis of Global Terrorist Attacks Prevention and Prediction for Combating Terrorism.

[53] Leonardo Nizzoli , Marco Avvenuti , Stefano Cresci , Maurizio Tesconi, Extremist Propaganda Tweet Classification with Deep Learning in Realistic Scenarios .

3) Journal

[9] @article author = "M. Han and H. Zhang", title = "Business intelligence architecture based on internet of things", journal = "Journal of Theoretical and Applied InformationTechnology", volume = "50", number = "01", pages = "90 , 95", year = "2013".

[22] D. Boyd and E. Nicole. “Social Network Sites Definition, History,” .Journal Of Computer—Mediated communication. 2008.

[23] M. DEWING , Les médias sociaux – Introduction Publication no 2010-03-F , Le 3 février 2013

[25]Christophe VILAIN, Article de Blog agence-web-cvmh /reseaux-sociaux/

[32] Bing Liu , Draft: Due to copyediting, the published version is slightly different

Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

[54] Hadley Wickham, « Tidy Data », Journal of Statistical Software
vol. 59, no 10, 2014, p. 1-23 .

[27]Introduction to Natural Language Processing (NLP), Publié le 11/08/2016 ,

4) Sit web

[2] Le professeur Philippe Besse , Professeur de Mathématiques à l'Université de Toulouse - INSA Institut de Mathématiques UMR CNRS 5219 Apprentissage & Statistique , Institut National des Sciences Appliquées , https://openclassrooms.com/fr/courses/5919236-decouvrez-la-science-des-donnees-pour-les-objets-connectes/6305091-abordez-les-fondements-de-lintelligence-artificielle.

[4] https://experiences.microsoft.fr/business/intelligence-artificielle-ia-business/comprendre-utiliser-intelligence-artificielle/#1 Publié le 9 février 2018 30/03/2021 16:38

[5] Jaume Duch Guillot , Direction générale de la communication , 30/03/2021 https://www.europarl.europa.eu/news/fr/headlines/society/20200827STO85804/intelligence-artificielle-definition-et-utilisation .

[6] Qu'est-ce que l'Intelligence artificielle (IA)Posted on 4 février 2019 in Blog, Lab-Sense https://www.lab-sense.com/2019/02/04/quest-ce-que-lia/ 30/03/2021

[7] Data Science Team, https://datascience.eu/fr/apprentissage-automatique/quest-ce-que-lapprentissage-automatique-une-definition/

[10] https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d Devin Soni , Supervised vs. Unsupervised Learning Mar 22, 2018·4 min read

[11] https://mrmint.fr/9-algorithmes-de-machine-learning-que-chaque-data-scientist-doit-connaitre02/04/2021.

[13]Subrat,Tutorial , Simply learn ,https://www.simplilearn.com/classificationmachine-learning-tutorial .

[14]fatimaElharithi ,https://io.hsoub.com/tech/93148 02/04/2021 .

[15] Cathy O'Neil, http://cedric.cnam.fr/vertigo/cours/ml2/coursArbresDecision.html.

[16]https://www.hebergementwebs.com/apprentissage-automatique-avec-ython/algorithmes-de-classification-foret-aleatoire02/04/2021.

[17] https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

[18] https://www.kdnuggets.com/2018/06/5-clustering-algorithms-data-scientists-need-know.html

[19] https://coursee.org/blog/artificial-intelligence/a-tour-of-machine-learning-algorithms.

[20]https://mzuer.github.io/machine_learning/data_split.

[26] www.alexa.com . Consulté le 25/05/2018 .

[30]B.Bathelot, Publié le 2 février 2018 . www.definitions-marketing.com Consulté le 25/05/2018 .

[34] https://www.agence-web-cvmh.fr/reseaux-sociaux/

[55] https://www.techno-science.net consulté le 10/06/2018.

[36] https://www.piloter.org/business-intelligence/datamining.htm .

[37] https://ichi.pro/fr/methodes-de-selection-des-fonctionnalites-dans-l-apprentissage-automatique-114173558902305

[38] www.researchgate.net consulté le 04/07/2021

[39] https://www.mygreatlearning.com/blog/bag-of-words/ .

[40] https://werfamous.com/fr/ Consulté le 05/06/2018 .

[41] https://finnaarupnielsen.wordpress.com/ Consulté le 05/06/2018 .

[42] http://www.wjh.harvard.edu Consulté le 05/06/2018 .

[43] http://www.sentic.net/ Consulté le 05/06/2018 .

[44] http://www.sentiwordnet.isti.cnr.it/ Consulté le 05/06/2018 .

[45] sentiment140. URL : http://www.sentiment140.com . Consulté le 05/06/2018 .

[46] http://www.tweetfeel.com . Consulté le 05/06/2018 .

[47] http://twitrratr.com . Consulté le 05/06/2018 .

[48] https://books.openedition.org/oep/214. Consulté le 05/07/2021 .

[49]http://smm.streamcrab.com . Consulté le 05/06/2018 .