Université de Blida 1–Saad Dahlab

Faculté des sciences

Département d'Informatique

Mémoire présenté par :

Mrs. ZERROUKI Hammeme et HAMDAD Adel

Pour l'obtention du diplôme de Master

Domaine : Mathématique et Informatique

Filière : Informatique

Spécialité : Traitement Automatique de la Langue

Sujet :

Détection du Langage Agressif dans les Posts des Utilisateurs sur les Réseaux Sociaux

Soutenu le :24/09/2020, devant le jury composé de :

Mr.S.Ferfera	Université de Blida 1	Président
Mme. D. Berramdane	Université de Blida 1	Examinatrice
Mme. H. Aliane	CERIST	Encadreur
Mme. M. MEZZI	Université de Blida 1	Promotrice

Remerciement

Nous tenons à remercier avant tout,«ALLAH»
Le tout puissant de nous avoirdonné la force, lavolonté et le courage de réaliser ce travail.

La présente étude n'aurait pas été possible sans le bienveillant soutien de certaines personnes. Et nous ne sommes pas non plus capables de dire dans les mots qui conviennent, le rôle qu'elles ont pu jouer à nos côtés pour en arriver là. Cependant, nousvoudrions lesprier d'accueillir ici tous nos sentiments de gratitude qui viennent du fond de nos cœurs, en acceptant nos remerciements.

Nous tenons à remercier respectivementnotreencadreur

« Mme. ALIANE »

D'avoirproposé et encadré ce sujet pendant 6 mois. Nous luiexprimons notre profonde gratitude pour

Nous avoir fait profiter de ses connaissances, mais aussi de ses méthodesdetravail, et surtout de sa rigueur scientifique .Veuilleztrouver ici l'expression de notre estime et considération.

Nos remerciements s'étendent chaleureusement notre enseignante et promotrice

« Mme. MEZZI »

Qui a endossée son rôle de la meilleure façon qui soit, qui a bien voulu noussuperviser

Nous a confié ce riche travail d'expérience et nous a guidés à chaque étape de sa consécration.

Nous retiendrons sa patience, sa disponibilité et sa compréhensibilité ainsi

Que son aide précieuse, ses encouragements inlassables, son amabilité, sa gentillesse, ces conseils avisés et ces idées riches qui ont contribué à alimenter nos réflexions méritent toute notre attention, Nous avons eu l'honneur d'être parmi ses étudiants et de bénéficier de sonriche enseignement. Veuilleztrouver ici, l'expression de notre gratitude et de notregrande estime.

Nous remercions

« Les membres de jury »

Pour avoir accepté deJuger notre travail.

Notre reconnaissance va aussi à tous ceux qui ont collaboréà notre Formation en particulier les enseignants dudépartement d'Informatique, Universitaire Saad Dahleb Blida Aussi à nos camarades de lapromotion 2019-2020. On remercie également tous ceux quiont

Participé de près ou de loin à élaborer ce travail.

Dédicaces

C'est avec un immense plaisir

Que je dédie ce travail

A mes très chers parents qui sont toute ma vie et tout ce que j'ai de plus cher au monde, en témoignage de ma reconnaissance infinie pour leurs nombreux Sacrifices.

Qu'ils trouvent en ce travail la preuve de mon éternel amour et ma reconnaissance envers eux.

Que dieu les gardes et leur procure la santé et le bonheur.

Aussi

À mes frères, ma famille sans oublier tous mes amis.

M.Hammeme Zerrouki

Dédicaces

Avec l'aide de Dieu le tout puissant que j'ai pu arriver au terme de ce travail que Je tiens à dédier à :

A mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études,

A mes chers frères,karim, Sidahmed et Tarek pour leur appui et leur encouragement,
A toute ma famille pour leur soutien tout au long de mon parcours universitaire,
Que ce travail soit l'accomplissement de vos vœux tant allégués, et le fruit de votre soutien infaillible,
Merci d'être toujours là pour moi.

M. Hamdad Adel

Résumé

Les réseaux sociaux représentent aujourd'hui un moyen de communication incontournable. Cependant, ils peuvent également être une source fiable pour diffuser les nouvelles et de la propagande dans le monde dans différentes langues, notamment la langue Anglaise.

En effet, certains utilisateurs créent actuellement des comptes pour diffuser des contenus abusifs. En effet, ce langage abusif est interdit et la détection automatique d'un tel langage est un défi majeur à surmonter.

Notre recherche suggère l'un des programmes de base du traitement automatique de la langue. Dans ce travail nous avons utilisé l'apprentissage automatique et profond afin de proposer un programme de détection automatique de langage abusif dans les postes en langue Anglaise sur Twitter.

La solution proposée permet de détecter et de classifier un poste abusif selon cinq différents algorithmes et l'analyse des résultats obtenus sur un Dataset constitué à partir d'un scrapping sur Twitternous a permisde déterminer l'algorithme le plus précis selon f-score et la complexité temporelle.

Mots Clés :Analyse des Réseaux Sociaux, Détection de Langage Abusif, Classification Automatique, ApprentissageAutomatique, Apprentissage Profond.

ABSTRACT

Social networks today represent an essential means of communication. However, they can also be a reliable source for spreading news and propaganda around the world in different languages, including the English language.

Indeed, some users are currently creating accounts to disseminate abusive content, moreover, this abusive language is prohibited and automatic detection of such language is a major challenge to overcome.

Our research suggests one of the basic programs for automatic language processing. In this work we used machine learning and deep learning in order to propose a program for automatic detection of abusive language in English language posts on twitter.

The proposed solution allows detecting and classifyingan abusive post according to five different algorithms and the analysis of the results obtained on a Dataset constituted from a scrapping on Twitter allowed us to determine the most precise algorithm according to f-score and time complexity.

Keywords:Social Network Analysis, Abusive Language Detection, Automatic Classification, Machine Learning, Deep Learning.

ملخص

تمثل الشبكات الاجتماعية اليوم وسيلة اتصال أساسية. ومع ذلك، يمكن أن تكون أيضًا مصدرًا موثوقًا به لنشر الأخبار والدعاية حول العالم بلغات مختلفة، بما في ذلك اللغة الإنجليزية.
في الواقع، يقوم بعض المستخدمين حاليًا بإنشاء حسابات لنشر محتوى مسيء، علاوة على ذلك، هذه اللغة المسيئة محظورة والكشف التلقائي عن هذه اللغة يمثل تحديًا كبيرًا يجب التغلب عليه.

يقترح بحثنا أحد البرامج الأساسية لمعالجة اللغة التلقائية. استخدمنا في هذا العمل التعلم الآلي والتعلم العميق لاقتراح برنامج للكشف التلقائي عن اللغة المسيئة في منشورات اللغة الإنجليزية على تويتر.

يسمح الحل المقترح باكتشاف وتصنيف المنشورات المسيئة وفقًا لخمسة خوارزميات مختلفة وتحليل النتائج التي تم الحصول عليها في مجموعة بيانات تم إنشاؤها من شبكة Twitterوالتيسمحت لنا بتحديد الخوارزمية الأكثر دقة وفقًا لدرجةf-score والوقت المستغرق.

<u>الكلمات الرئيسية:</u> تحليل الشبكات الاجتماعية، كشف اللغة المسيئة، التصنيف التلقائي، التعلم الآلي، التعلم العميق.

Table de matières

Liste des figures

Liste des tableaux

Liste d'acronymes

AA : Apprentissage Automatique
API : Application Programming Interfaces
ARS: Analyse des Réseaux Sociaux
Bi-GRU: Bi GatedRecurrent Unit
CBOW: Continuous Bag-of-Words
CLI: Interface de ligne de commande
CNN: Convolutional neural network (Réseau neuronal convolutif)
COOPOL: Cooperative Politique
CRM: Customer Relationship Management
CT: Classification de Text
iLCD: intrinsic Longitudinal Community Detection
JASSS: Journal of Artificial Society and Social Simulation
LSMC: Lottery Syndicate Manager Complete
LSTM: Long Short Term Memory
OSNs: Online Social Network
PV-DBOW: Distributed bag of words paragraph vector
PV-DM: Paragraph Vector: A distributed memory model
QDA : Qualitative Date Analysis
RNA : Les réseaux neuronaux artificiels
RNN : Recurrent neural network (Réseau de neurones récurrents)
RS : Réseau Sociaux
SVM : Support Vecteur Machine
TAL : Traitement Automatique de la Langue
TALN : Traitement Automatique du Langage Naturel
TF-IDF: Term Frequency-Inverse Document Frequency
VDCNN: Very Deep Convolutional Neural Networks
XGboost: Gradient boosting

Introduction Générale

1. Contexte globale

Les réseaux sociaux sont de nos jours,l'outil le plus populaire pour diffuser des pensés et partager des photos et des vidéos, etc. Pour toutes ces raisons, ils ont pris une place importante de notre vie quotidienne. Bien qu'ils contiennent de nombreux avantages, ces nouveaux moyens d'expressionet de communication ne sont pas sans risque.Parmi ces risques, on trouve celui de contenir des contenus obscènes et offensants voire même des fois dangereux et pour cause, les réseaux sociaux ne sont pas complètement contrôlés.

2. Problématique

Plusieurs utilisateurs des réseaux sociaux créent maintenant des comptes abusifs pour distribuer du contenu violant dans différentes publications, et les adolescents sont les premières victimes de ce contenu car ils passent la plupart de leur temps sur les différents sites de réseaux sociaux.

Pour pallier à ce nouveau fléau,la détection automatique des obscénités et les contenus indécents, est devenue très utile pour protéger les consommateurs de contenus multimédia, qu'ils soient grands ou petits et aussi pour mettre en place des consensus et des bases standard permettant de détecter voire intercepter les contenus abusifs.

3. Objectifs

Dans notre travail, nous nous intéressons à la conception et au développement d'une solution de classification du contenu des tweets (entre contenu abusif et contenu non-abusif).Dans un premier temps, il faudra passer par une étape de parcours bibliographique afin d'assimiler les différentes notions relatives au domaine. Puis, il est question de faire une sélection d'un nombre restreint d'algorithmes qui peuvent s'avérer intéressants pour résoudre notre problème et de faire des tests sur une collection obtenue à partir d'un scrapping[1]sur Twitter qui est devenu actuellement, l'une des mines renfermant des données intéressantes pour faire différents traitement relevant du domaine de Traitement Automatique de la Langue.

4. Organisation du mémoire

Afin d'atteindre l'objectif susmentionné, nous avons structuré notre mémoire en deux parties comme suit :

La première partie concerne l'étude bibliographique qui comprend deux chapitres :

[1] Le scrapping est un terme emprunté de l'Anglais et utilisé ces dernières années en Informatique pour désigner le processus permettant l'extraction de grandes quantités d'informations d'un site Web.

Introduction générale

- Un premier chapitre intitulé « **Généralités sur les réseaux sociaux** » est consacré à la présentation de certains concepts fondamentaux relatifs aux réseaux sociaux numériques.

- Un second chapitre nommé «**La détection automatique du langage agressifau fil du temps** », nous intéresserons dans ce chapitre, à la détection automatique du langage agressif dans les réseaux sociaux où nous allons présenter quelques travaux récents qui nous ont semblé les plus intéressants et faire une étude comparative entre eux.

Une deuxième partie qui concerne les algorithmes que nous avons utilisés, les tests que nous avons effectués et leur analyse. Elle comprend à son tour, deux chapitres :

- Un troisième chapitre intitulé « **Algorithmes d'apprentissage automatique**» où nouspassons en revue un peu de théories sur l'apprentissage automatique, ses modèles, et quelques algorithmes.

- Un dernier chapitre « **Conception et Implémentation de la solution** » où nous avons proposé une méthode pour détecter le langage agressif dans les post Twitter en langue Anglaise. Notre méthode utilise l'apprentissage automatique et profond, elle nous permet de classifier un poste selon qu'il soit abusif ou non et ce, en utilisant cinq déférents algorithmes afin de nous permettre de déterminer dans un second temps l'algorithme le plus précis selon f-score.A la fin,où nous avons présenté l'étude de cas sur le réseau social Twitter et évaluer les performances de l'algorithme proposé.

Enfin, nous terminerons ce mémoire par une conclusion générale et quelques perspectives.

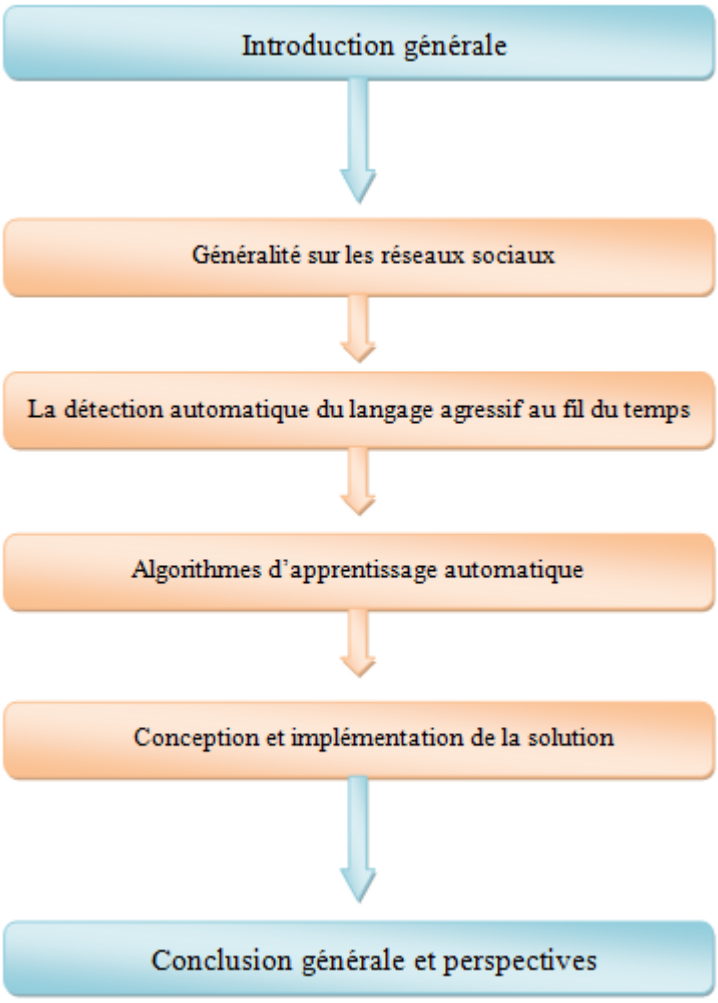


Figure 1 : L'organisation du mémoire.

Chapitre I :
Généralités sur les Réseaux Sociaux

1. Introduction

De nos jours, les Réseaux Sociaux (RS) sont devenus une partie très importante et omniprésente de notre vie quotidienne. Ce terme de RS est souvent utilisé pour référer à différents moyens permettant la distribution, la diffusion ou la communication avec possibilité de partage des contenus ou des expressions. Comme tout un moyen de communication le contenu peut être propre, obscène, cordial, offensif, etc. Les utilisateurs des réseaux sociaux trouvent une liberté d'expression, à tel point qu'ils expriment leurs opinions d'une manière violente dans des commentaires envers divers statuts et publications.

Ainsi, l'utilisation de langage agressif lors des échanges communicationnels se fait d'une manière explicite ou bien directement envers le récepteur du message, elles entrainent des conflits flagrants entre les personnes impliquées, malgré que la victime ne ressentira pas des coups physiques mais les répercutions morales et internes sont des fois irréversibles.

Dans ce premier chapitre, nous allons introduire des notions théoriques et des concepts clés relatifs à notre champ d'étude à savoir: la notion de réseaux sociaux, leurs caractéristiques, leurs types, leur intérêts, quelques principaux réseaux sociaux ainsi que l'analyse des réseaux sociaux et l'analyse des données collectées.

2. Définition des Réseaux Sociaux

Dans la littérature, il existe plusieurs définitions sur les réseaux sociaux, nous présentons quelques-unes ci-dessous:

Un réseau social est constitué à la fois d'un ensemble de personnes liées entre elles et par la force de leurs liens. On peut aussi dire qu'un réseau social est un ensemble d'individus liés entre eux par des liens caractérisés par un degré de familiarité variable qui va de simple connaissance aux liens familiaux les plus étroits[1].La figure 2 représente un exemple de réseau social.

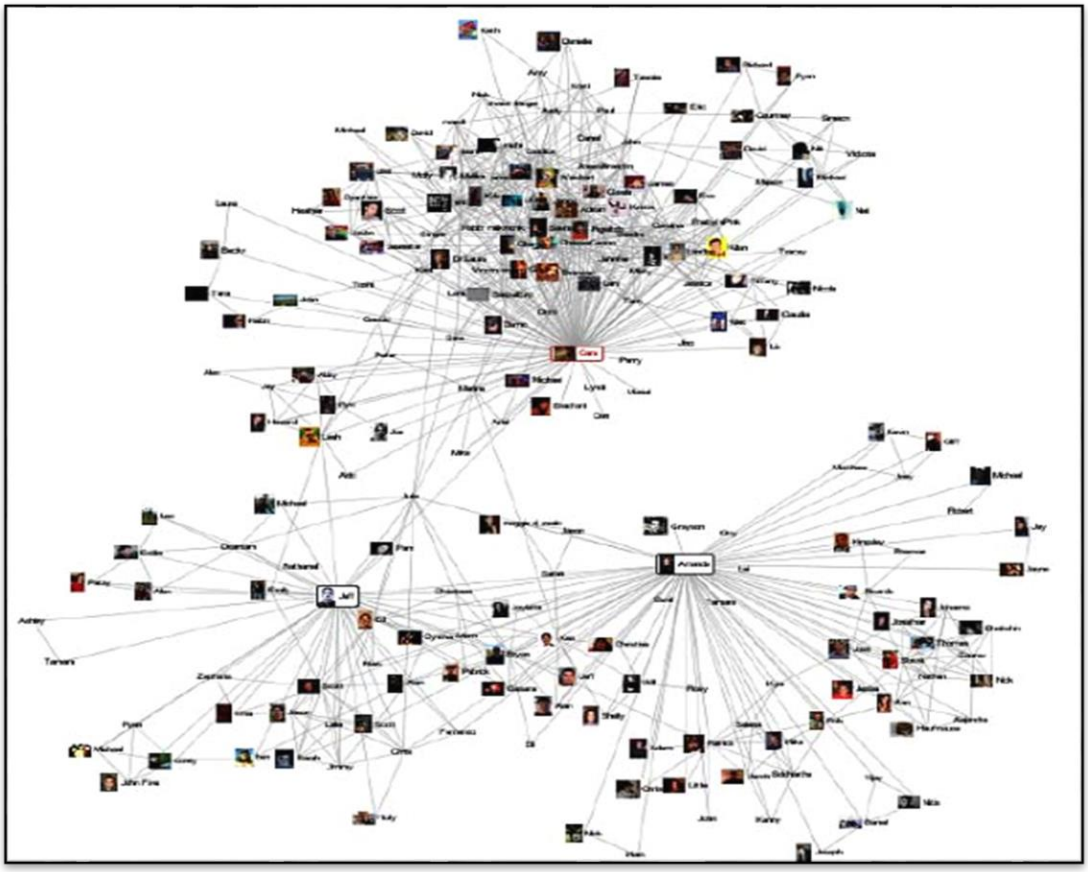


Figure 2:Représentation d'un réseau social[2].

Par réseau social, nous entendons donc toute plateforme en ligne dont la finalité est de mettre en relation des membres, et sur laquelle un individu peut s'inscrire librement, construire son propre réseau, produire du contenu, le partager et interagir avec les membres de son réseau. Un membre peut y créer un profil public visible par tous ou privé, visible par ses contacts uniquement. C'est ce profil qui servira de carte de visite à l'internaute sur le réseau social dont il est membre. L'intérêt de telles plateformes est notamment de pouvoir suivre l'actualité des membres de son réseau et d'éventuellement la commenter[2].

Par ailleurs, le réseautage social[3] (distinct du concept de réseau social en sociologie) se rapporte à une catégorie des applications d'Internet pour aider à relier des amis, des associés ou d'autres individus employant ensemble une variété d'outils. Ces applications, connues sous le nom de «*service de réseautage social en ligne*» (en anglais Social Networking) deviennent de plus en plus populaires. Elles peuvent aussi permettre une meilleure distribution artistique, en favorisant la formation de contacts, et en invitant des artistes à assurer une visibilité de leur travail (ex. musique, vidéo, photographie).

Selon la définition proposée dans [4][5] les réseaux sociaux sont des espaces d'échange sur Internet qui permettent aux individus de construire des profils publics ou semi publics, associés à une liste de contacts inscrits sur le même site.

Actuellement, Les OSNs (Online Social Network) intéressent différents domaines de recherche tels que : sociologie, psychologie, communication, informatique, marketing, …etc.

3. Les typologies des Réseaux Sociaux

Dans le monde des réseaux sociaux, on trouve plusieurs classifications suivant plusieurs critères, ces typologies sont organisées selon[1] :

3.1 L'évolution et l'apparition

Il y a une tentative de chronologisation des réseaux sociaux numériques, selon laquelle les auteurs dressent l'évolution et l'apparition des réseaux. Pour eux, il existe huit types de réseaux : *les réseaux généralistes, les réseaux politiques, les réseaux hyper locaux, les réseaux d'universités, d'entreprises, les réseaux associatifs,* ceux *des communautés d'intérêts et les réseaux de passionnées*. Leurs typologies synthétisées sous forme d'un tableau 1 :

Tableau 1: Typologie des réseaux sociaux[1].

Type de RS	But	Exemple
Réseaux généralistes	- Ces sites permettent indirectement de nouer des affinités personnelles, sans pour autant avoir vocation unique d'être des sites de rencontre - Ces sites permettent de créer et d'agrandir son cercle d'ami	- Facebook[2] - Meetic[3]
Réseaux politiques	- Ces sites permettent d'interroger le monde politique dans les réseaux sociaux	- Créateurs du possible de l'UM[4]P - lecoopol[5] (Coopérative politique)

[2]http://www.facebook.com
[3]http://www.meetic.fr/
[4]http://www.lescreateursdepossibles.com/
[5]https://www.lacoopol.fr/

Réseaux hyper locaux	- Renforcer les liens au niveau local - Mieux connaître ses voisins - Promouvoir et d'encourager les solidarités	- Voisinéo[6] - Peuplade[7] - la Ruche à Rennes[8]
Réseaux d'universités	- Un regroupement d'universités, dans le but de coopérer dans les domaines de la mobilité estudiantine, des enseignements ou de la recherche scientifique par exemple.	- Zeeya[9], - Réseau campus[10], - etnoka[11]
Réseaux d'entreprises	- Ces sites ontpour vocation de permettre à leurs adhérents de réaliser des objectifs dans l'intérêt commun de tous	
Réseaux associatifs	- Ces sites permettent de présenter l'activité de certaines associations et aussi de donner les différents renseignements afin de faciliter l'intervention en temps réel ou en différé.	- l'Association Française de Sociologie - les réseaux des créatifs culturels.
Réseaux des communautés d'intérêts	- Ces sites proposent des types de relations beaucoup plus spécifiques et qui pour certains s'apparentent à des communautés d'intérêts.	- baby-boom[12]-Globe2child [13] - Memoree[14]
Réseaux de passionnées	- Traite de l'image sous toutes ses formes.	- Knowtex[15]

3.2 La fonctionnalité

D'autres chercheurs ont tenté de classer et de catégoriser les réseaux sociaux selon un autre critère. Dans[6][7]l'auteur a essayé dans un billet intitulé : « *Ras le bol des réseaux sociaux ?* », d'élaborer un classement suivant la fonctionnalité [1].

[6]http://www.voisineo.com
[7]http://www.peuplade.fr/home/nHome.php
[8]http://www.ruche.org
[9]http://zeeya.net/
[10]http://réseaucampus.com/
[11]http://etnoka.fr/
[12]Le baby-boom ou « pic de natalité » est une augmentation importante du taux de natalité dans certains pays, juste après la fin de la seconde guerre mondiale.

[13]http://www.globe2child.org/xwiki/bin/view/main/webhome.

[14]http://www.memoree.fr/

[15]http://www.knowtex.com/

- *Networkings:* permettent les échanges entre les professionnels.
- *Bloglikes:* ressemblent vaguement aux blogs et sont souvent le refuge d'ados en mal de reconnaissance.
- *Spécialisés :* regroupent des communautés autour d'un thème bien précis.
- *Micro-blogging* : chat publique instantané.
- *Fourres-tout :* ce sont les inclassables qui se servent du collaboratif ou du participatif pour alimenter leur service.
- *Open-sources :* plateformes qui permettent aux utilisateurs de créer leurs propres réseaux.

3.3 Du point de vue des chercheurs

Dans[8]l'auteur catégorise les OSNs(Online Social Network) selon leurs trois objectifs qu'il nomme respectivement : *socialisation, réseautage* et *navigation sociale*[1].

- *Réseaux sociaux de socialisation :* Cette catégorie se caractérise par son aspect récréatif et elle conçue pour les loisirs de communication sociale entre les membres. Les connexions sont souvent utilisées pour trouver et afficher des listes d'«*amis* » existants d'ores et déjà. De plus, les connexions sont souvent utilisées pour trouver des « *ami*s » existants hors ligne, comme par exemple : MySpace et Facebook et Cyworld[16] (un monde visuel coréen lancé en 2001).

- *Réseaux sociaux de réseautage :*Utilisés davantage pour trouver de nouveaux contacts et peuvent servir pour entrer en connexion avec des personnes inconnues auparavant comme c'est le cas de LinkedIn ou Viadeo. Ce sont des sites de réseautage à caractère professionnel.

- *Réseaux sociaux de navigation :* Comme Digg[17] ou Del.icio.us[18], qui sont des sites de partage de liens Internet (connu sous le social bookmarking). Ce type de réseaux est un moyen pour aider les utilisateurs à trouver une information ou des ressources. Autrement dit, nous trouvons des listes de contacts, listes permettant l'accès à l'information et aux ressources associés à ceux-ci. Les membres peuvent soit lire les propositions mises en avant en page d'accueil, soit utiliser la

[16]http://us.cyworld.com
[17]http://digg.com
[18]http://delicious.com

navigation sociale en lisant les informations postées ou recommandées par leur amis, ou bien pour certains, recourir plusieurs objectifs.

Par ailleurs, il existe un classement de réseaux sociaux selon trois catégories : *Réseaux ouverts*, *Réseaux sur invitation* et *Services en ligne de réseautage professionnels* qui favorisent les rencontres professionnelles, les offres de poste et la recherche de profils.

4. Intérêts des réseaux sociaux

D'après [9], les services des réseaux sociaux sont classés selon les six grands usages illustrés dans la figure 3 comme suit :

Figure 3: Panorama des médias sociaux 2019 [9].

- **Les médias sociaux principaux :** ***Facebook, Twitter, LinkedIn, YouTube*** **et** ***Instagram*** (au centre du graphique).
- **Les médias sociaux de partage (sharing) :** ils servent à partager tout type de contenu, en public ou à son réseau (photo et vidéo, musique…).
- **Les médias sociaux de réseautage (networking) :** ils servent à créer et développer un réseau. Linkedin et Viadeo permettent par exemple de se créer un réseau professionnel. Ils sont ainsi très utiles dans le cadre d'une activité B to B.

- **Les médias sociaux de discussion (messaging) :** ils permettent la discussion, instantanée ou non, entre leurs membres. Le plus connu d'entre eux est sans doute Skype. Les forums sont également des médias de discussion.
- **Les médias sociaux de publication (publishing) :** ils servent à publier du contenu original, des articles, des rapports, des tests… il s'agit essentiellement des plateformes de blogging (WordPress, Blogspot…).
- **Les médias sociaux de collaboration (collaborating) :** ces médias sociaux comme Slack permettent de collaborer à distance. Ils sont notamment très utilisés en gestion de projet.

5. Principaux réseaux sociaux

Dans ce qui suit, nous allons présenter les différents types de réseaux sociaux :

5.1 Exemple deréseaux sociaux grands publics

Ci-dessous des exemples de quelques réseaux sociaux grands publics[10]:

- ***Facebook :*** le réseau social incontournable, créé en 2004 par Mark Zuckerberg, le réseau social s'est imposé dans le monde entier avec 1,3 milliards d'utilisateurs. Il vous propose 3 manières d'établir votre présence : **le profil, la page et le groupe**
  - ✓ ***Le profil*** est un compte personnel représentant une personne physique. Il permet de partager des contenus avec vos amis ainsi qu'à toute personne ayant accepté de partager ses informations avec vous.
  - ✓ ***La page*** : permet aux utilisateurs de suivre vos actualités sans partager ses contenus personnels. De plus vous pourrez visualiser rapidement les statistiques de vos publications : personnes atteintes, nombre de clics, nombre de fans, etc.
  - ✓ ***Le groupe*** : facebook se situe à mi chemin entre la page et le profil. Il rassemble des membres (profils) autour d'un centre d'intérêt commun, et leur permet d'échanger des informations (texte, photos, vidéos, liens, fichiers). Les groupes sont peu utilisés par les marques alors qu'ils représentent un moyen efficace de toucher une cible précise. Les membres sont regroupés autour d'un sujet et sont plus engagés que sur une simple page.

Facebook possède la plus grande base d'utilisateurs à l'heure actuelle.

- ***Twitter*** : Le réseau social du temps réel, est né le 21 mars 2006 *à San Francisco. Jack Dorsey, Evan Williams, Biz Stone et Noah Glass* ont souhaité créer une plateforme où les utilisateurs pourraient facilement partager des moments de vie avec leurs amis par l'intermédiaire de messages courts du type « SMS ».C'est pour cette raison que les tweets sont sous forme de messages courts allant jusqu'à 140 caractères. Aujourd'hui, Twitter compte près de 284 millions d'utilisateurs actifs.

- ***Google+*** : Créé en 2011, Google+ est un réseau social qui fonctionne sur le principe de « cercles » de contacts, permettant de choisir facilement avec quel « cercle » de contacts vous souhaitez partager vos contenus (amis, famille, collègues, clients, prospects, etc.). Les cercles vous permettent de segmenter vos contacts et donc d'adapter vos messages en fonction de votre cible. L'utilisation de Google+ peut également améliorer votre référencement. En effet, Google affiche par défaut depuis fin 2013 des résultats personnalisés en fonction de votre activité sur Google+. Vos abonnés sur Google+ verront vos posts plus souvent sur les pages de résultats Google, ce qui permettra votre Authorrank et donc votre référencement naturel. Par conséquent, plus vous serez actifs sur Google+, plus vos posts seront visible sur les pages de résultats Google.

5.2 Exemple de réseaux sociaux professionnels

Ci-dessous des exemples de quelques réseaux sociaux professionnels [11] :

- *Linkedin :* Un réseau professionnel international permet la mise en relation entre des professionnels. Il offre un espace de présentation de sescompétences et expériences qui peuvent être consultable par le public. Très utile pour le recrutement mais aussi pour apprendre de nouvelles compétences.
- *Viadeo:*Il est le pendant français du réseau social Linkedin. Il permet lui aussi de construire et de gérer son réseau professionnel. Viadeo est plus populaire et plus connu en France que Linkedin mais Il offre à peu près les mêmes possibilités.

- *Xing :*C'est une plateforme allemande qui permet de construire et d'agréger son réseau professionnel. Il possède 16 millions d'utilisateurs répartis sur plus de 190 pays[19].

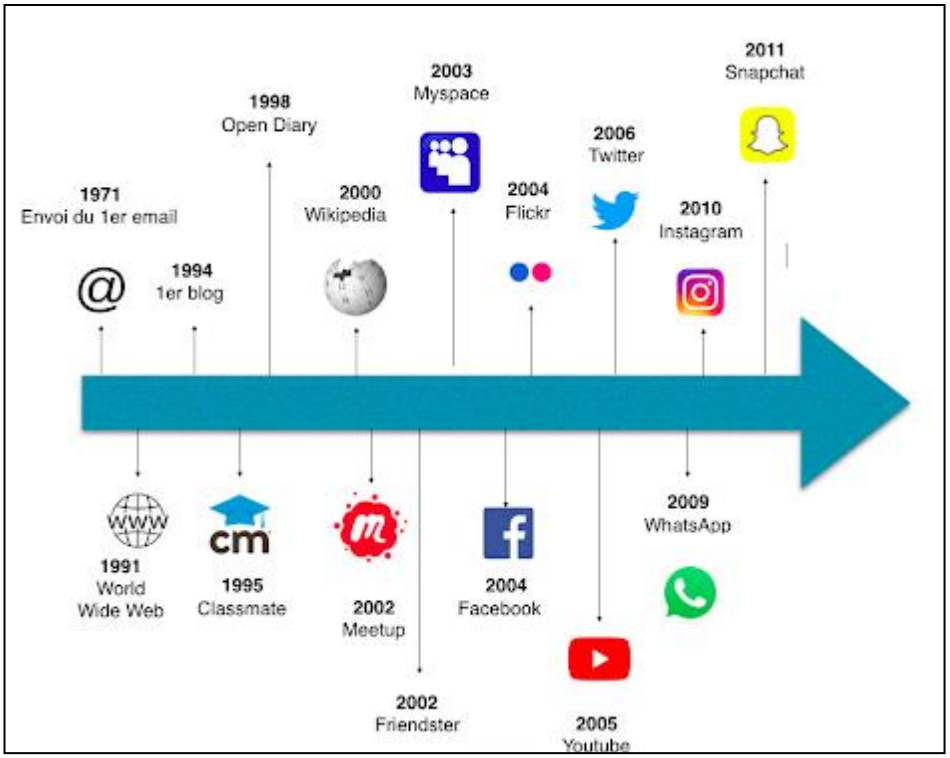


Figure 4: Chronologie des réseaux sociaux [12].

Le tableau suivant décrit les avantages et les inconvénients des réseaux sociaux présentés précédemment:

	Facebook	**Twitter**	**Linkedin**
Avantages	Créer une page Facebook simple. La plateforme offre de plus en plus d'outils qui permettent notamment aux entreprises et marques de suivre la progression de leur nombre d'adeptes ainsi que leurs données démographiques.	On peut s'informer ou diffuser une information importante et pertinente en temps réel. Il n'y a aucun problème à publier fréquemment puisque les tweets se suivent rapidement.	C'est le réseau social professionnel par excellence, qui permet de bâtir un réseau à partir de ses expériences de travail mais également autour de ses intérêts et compétences.

[19]https://www.statista.com/

Inconvénients	Facebook demande une interaction avec les adeptes. On ne peut pas diffuser une information sur Facebook et ignorer la réaction des adeptes.	Sur Twitter, on ne peut pas envoyer un message qui contient plus de 140 caractères. Twitter oblige les entreprises à être concises et claires.	L'interaction sur Linkedin, à part dans certains groupes, est vraiment limitée. Une interface un peu moins accessible ne rend pas les choses faciles.

Tableau 2: Comparaison entre les réseaux sociaux [11].

6. Analyse des réseaux sociaux

La première personne à avoir représenté un réseau social est Jacob Levy Moreno[20] au début des années 1930. Son objectif étant de visualiser graphiquement un réseau social, il a représenté les personnes par des points et une relation entre deux personnes par des flèches. Cette représentation est désignée par le terme *Sociogramme (*figure 5). Mais on parlait également de *toiles* en raison de leur aspect en toile d'araignée (figure 6). Cette forme de visualisation fut un premier outil d'identification rapide des caractéristiques d'un réseau social [13].

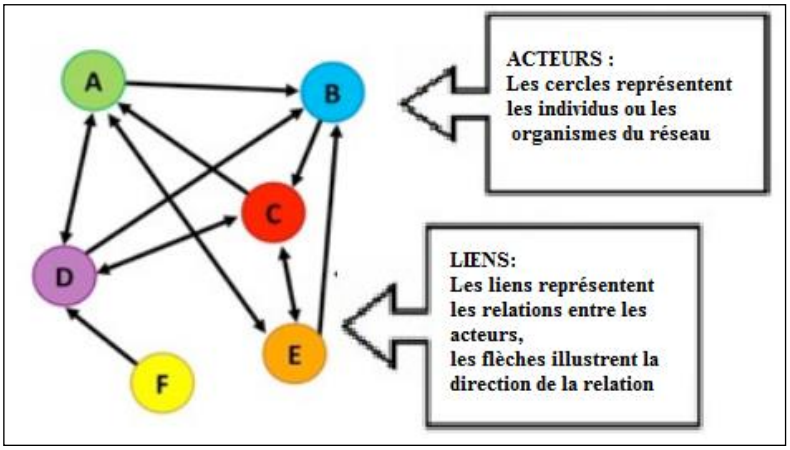


Figure 5: Exemple d'un sociogramme[13].

[20] Jacob Levy Moreno né le 18 mai 1889 à Bucarest, et mort le 14 mai 1974 à Beacon (États-Unis), est un médecin américain d'origine roumaine.

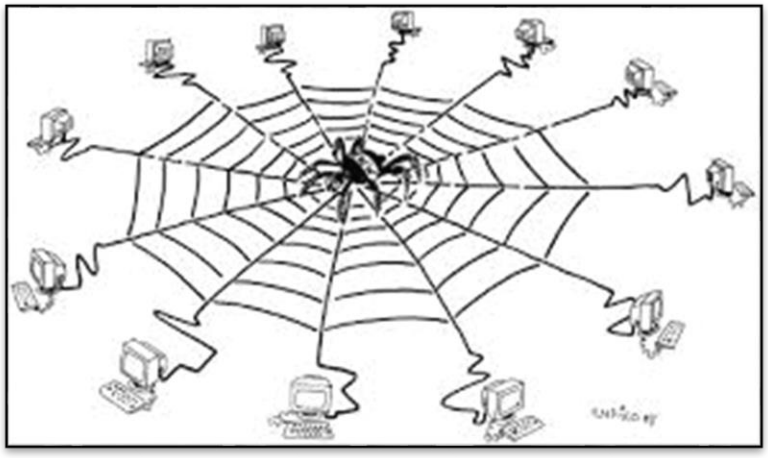


Figure 6 : Représentation en toile d'araignée d'un réseau social [13].

6.1 Définition

L'Analyse des Réseaux Sociaux (ARS) est la cartographie et la mesure des relations et des flux entre personnes, groupes, organismes, ordinateurs, sites web ou toute autre entité de traitement d'information ou de connaissances [13].Les nœuds dans le réseau sont les personnes ou les groupes tandis que les liens montrent les relations ou les flux entre les nœuds.

L'analyse des réseaux sociaux est principalement une analyse visuelle et mathématique des relations humaines.

6.2 Les approches de l'analyse des réseaux sociaux

Il y a trois approches de l'ARS[8] :

- *L'approche réseau complet :* Cela implique qu'avant de commencer l'ARS, tous les acteurs sont connus. Il est donc possible de recueillir toutes les données facilement. Un exemple d'un réseau complet est le réseau au sein d'un milieu de travail où il est facile d'obtenir une liste de toutes les personnes qui y travaillent. Dans ce cas-ci, les liens du réseau seraient fondés sur une relation professionnelle. Les données du réseau proviendraient de tous ses membres.
- *L'approche réseau égocentré:* Cette approche analyse le réseau social d'un individu ou d'une organisation. Les réseaux égocentrés individuels peuvent être joints ou

croisés pour avoir une meilleure compréhension du réseau global. Les réseaux égocentrés sont particulièrement utiles lorsque tous les acteurs du réseau ne sont pas connus au début du processus. Le réseau d'amis d'un individu est un exemple de réseau égocentré. Dans ce cas-ci, les liens du réseau seraient fondés sur une mesure ou un niveau d'amitié (i.e. comment les amis se parlent entre eux, depuis combien d'années ils sont amis, etc.) Toutes les données du réseau proviendraient d'une seule personne.

- *L'approche par archives:* Cette approche est utilisée lorsqu'il y a des documents historiques qui indiquent qu'il y a des relations entre des acteurs. Un exemple d'un réseau par archives est le cas d'une personne qui essaie de reconstruire son arbre généalogique à partir de registres historiques ou d'informations disponibles sur des sites internet tels qu'ancestry.com. Toutes les données du réseau proviennent de documents ou de sites internet.

6.3 La Représentation de l'analyse des réseaux sociaux

Les mathématiciens ont fait le rapprochement entre les représentations sociogrammes et la théorie des graphes au sens mathématique. Le graphe est devenu par la suite la représentation adoptée par toutes les sciences manipulant l'analyse des réseaux sociaux, dont la Sociologie, les Mathématiques et l'Informatique.

Un réseau social est souvent représenté sous forme de graphe. Ce dernier est composé de *nœuds* (sommets) qui décrivent les personnes et les *liens* (arêtes) qui décrivent les connexions/relations sociales entre ces personnes.

Un graphe sert à identifier les personnes selon différents critères à savoir[13]

- Les personnes les plus reliées entre elles,
- Les amis explicites d'un utilisateur, ou même
- Les personnes partageant des caractéristiques en communs.

Figure 7 : Représentation graphique de réseau social de media [13].

Dans ce contexte, il existe des travaux qui visent à analyser certains types de réseau.

Dans ce qui suit, nous nous focalisons sur les communautés (centrés sur un ensemble d'utilisateurs).

6.3.1 Les Communautés

L'analyse d'un réseau social peut être centrée sur un ensemble d'utilisateurs ayant des caractéristiques en commun. Cet ensemble d'utilisateurs est appelé communauté. Le problème de la détection de communautés dans les réseaux sociaux est un sujet relativement récent, mais qui a très rapidement conduit à une grande quantité de travaux.

La figure 8 montre une visualisation des communautés dans le réseau de citation de JASSS (Journal of ArtificialSocietie and Social Simulation), avec la première version d'iLCD (intrinsic Longitudinal Community Detection), un méta-algorithme permettant de traiter des réseaux temporels directement, sans passer par un prétraitement).

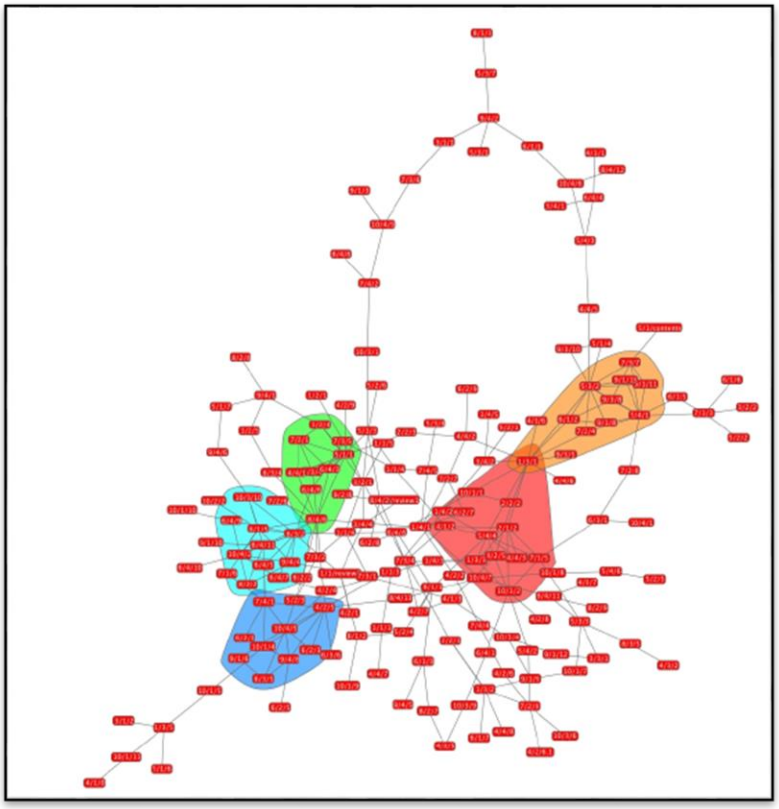


Figure 8 : Représentation de communautés par l'algorithme d'iLCD[13].

6.4 Accès aux données des réseaux sociaux

Avec la popularité des réseaux sociaux, les analyses sont devenues de plus en plus fréquentes et nécessitent la collecte des données du réseau à savoir les données sur les utilisateurs (les informations personnelles, les amis, etc.), ressources(le type, la date de création, la popularité, etc.), les interactions dans ces réseaux (le partage des ressources, les tags, les commentaires, etc.).

Selon le type du réseau, il est possible d'extraire des données de différentes manières, à savoir[8] :

- *Réseaux d'intérêt :* Ces réseaux sont créés à partir de forums et impliquent les utilisateurs et leurs sujets d'intérêt. Par projection, ce réseau peut ensuite être utilisé pour générer un réseau d'utilisateurs, dans lequel, deux utilisateurs sont connectés s'ils partagent les mêmes sujets d'intérêt.

- *Réseaux de blogs :* Les articles publiés sur les blogs font souvent référence à d'autres blogs par le biais des liens hypertextes. Ces informations peuvent être utilisées pour générer des réseaux de blogs connectant deux blogs quand l'un fait référence à l'autre.
- *Réseaux d'amitié* : À l'origine, les sites communautaires tels que Facebook, Twitter, Google+, Delicious ou Flickr, permettaient d'obtenir des informations sur les liens d'amitiés ou de connaissances entretenus par les individus. Aujourd'hui, face à la diversité d'informations présentes sur ces sites (individu, entreprise, parti politique, artiste, produit, évènement, etc.), nous observons que la sémantique du lien s'assimile désormais à "porter un intérêt à".
- *Réseaux de co-auteurs :* Le cas le plus répandu concerne les bases de données d'articles scientifiques. Par exemple, en utilisant les données issues de DBLP[21] Computer Science Bibliography, un réseau de co-auteurs peut être obtenu, dans lesquels deux auteurs sont connectés s'ils ont collaboré sur un même article.

L'accès aux données des réseaux sociaux, permet d'analyser les différentes caractéristiques relatives aux informations présentes. Les analyses peuvent être orientées vers l'adaptation de l'information à l'utilisateur.

6.5 L'analyse des données dans les réseaux sociaux

De nos jours, les données sont qualifiées de l'or noir du XXI e siècle. Il s'agit en effet d'une ressource précieuse qui dans bien des cas est sous-exploitée. Sa simple extraction n'est pas suffisante, encore faut-il avoir un moyen de « raffiner » ces données brutes pour en extraire les informations utiles. Les données issues des médias sociaux constituent une grande partie des Bigdata[22].

Les réseaux sociaux sont largement utilisés dans différents domaines de la recherche universitaire, notamment en sciences politiques, communications, journalisme et management. Autant l'influence des médias sociaux sur la politique, et particulièrement son impact dans les campagnes électorales que ses interactions avec les médias traditionnels sont des sujets d'actualités à l'échelle planétaire. De plus, les données des médias sociaux peuvent

[21]https://dblp.uni-trier.de/

[22]Le terme Big Data décrit des ensembles de très gros volumes de données – à la fois structurées, semi-structurées ou non structurées – qui peuvent être traitées et exploitées dans le but d'en tirer des informations intelligibles et pertinentes.

jouer un rôle de baromètre pour suivre l'évolution des attitudes sur des sujets d'actualité ou des questions controversés.

L'analyse des données textuelles issues des médias sociaux peut être considérablement simplifiée si 1l'on dispose des bons outils. Parmi les outils existants *QDA Miner*(14).et*WordStat*(15).

7. Conclusion

De nos jours, les réseaux sociaux numériques sont devenus des outils de communication incontournables. Dans ce premier chapitre, nous avons exposé les principaux axes de notre travail en commençant par la notion de réseaux sociaux, leurs types, quelques principaux réseaux sociaux ainsi que l'analyse des réseaux sociaux et l'analyse des données collectées.

Cependant, bien que les RS contiennent de nombreux avantages, ils ne sont pas sans risque. Comme tout un moyen de communication, le contenu peut être propre, cordial, haineux, ou agressif. Malheureusement, les réseaux sociaux ne peuvent pas contrôler tout le contenu que les utilisateurs publient, c'est pour cette raison, il existe une demande de détection automatique de langage agressif via les RS.

Dans le chapitre suivant, nous intéressonsà la notion de langage agressif, aussi nousallons présenter quelques travaux récents qui nous ont semblé les plus intéressants et faire une étude comparative entre eux.

Chapitre II :
La détection automatique du langage abusif au fil du temps

1. Introduction

Dans les plateformes de médias sociaux, le langage agressif (Hate Speech) peut être une cause de «cyber conflit» qui peut affecter la vie sociale à la fois au niveau individuel et au niveau du pays. Les contenus haineux propagés via les réseaux sociaux ont le potentiel de causer du tort et de la souffrance sur une base individuelle et de conduire à des tensions et des troubles sociaux au-delà du cyberespace. Cependant, les réseaux sociaux ne peuvent pas contrôler tout le contenu que les utilisateurs publient. Pour cette raison, il existe une demande de détection automatique des langages agressifs.

Dans un premier temps nous allons présenter quelques définitions de ce langage, ensuite nous intéresserons à la détection automatique du langage agressif dans les réseaux sociaux où nous allons présenter les travaux récents existants dans la littérature et nous finalisons par une étude comparative entre eux.

2. Le langage Agressif

Le langage agressif ou la violence verbale (Hate Speech) est un phénomène de la société, elle fait partie du langage commun chez les individus et on le retrouve un peu partout dans leurs discours et interactions avec autrui. La violence verbale se manifeste lors des conflits entre les individus ou chacun cherche à exprimer et imposer son point de vue et opinions dans divers sujet de discussion. La violence verbale fait partie des pratiques langagières, elle est définie comme un acte« menaçant » et «blessant » qui se manifeste lors des interactions entre les individus. La violence verbale se trouve dans plusieurs milieux (familial, professionnel, réseaux sociaux...etc.).

La violence ne peut s'imaginer qu'au sein d'une interaction entre deux individus ou bien tout au long d'un échange communicatif quelconque, les différents participants que l'on dira donc des interactants exercent les uns sur les autres un réseau d'influences mutuelles; parler c'est échanger et c'est changer en échangeant.

Dans ses communications, des fois les commentaires s'embrasent rapidement, les échanges laissent la place à des débats ou chaque intervenant veut défendre son idée sans même chercher à comprendre celle de l'autre. L'auteur de la violence verbale cherche à déstabiliser sa victime, nuire et provoquer.

Nous résumons les principales définitions du langage agressif provenant de diverses sources.

- ***Encyclopédie de la Constitution américaine:*** "Le discours de haine est un discours qui attaque une personne ou un groupe sur la base d'attributs tels que la race, la religion, l'origine ethnique, l'origine nationale, le sexe, le handicap, l'orientation sexuelle ou l'identité de genre." [16]
- ***Facebook:*** «Nous définissons le discours de haine comme une attaque directe contre les personnes en fonction de ce que nous appelons des caractéristiques protégées: race, ethnie, origine nationale, appartenance religieuse, orientation sexuelle, caste, sexe, genre, identité de genre et maladie grave ou handicap. Nous offrons également des protections pour le statut d'immigration. Nous définissons l'attaque comme un discours violent ou déshumanisant, des déclarations d'infériorité ou des appels à l'exclusion ou à la ségrégation. » [17].
- ***Twitter:*** «Conduite haineuse: Vous ne pouvez pas promouvoir la violence contre ou attaquer directement ou menacer d'autres personnes sur la base de la race, de l'ethnie, de l'origine nationale, de l'orientation sexuelle, du sexe, de l'identité de genre, de l'appartenance religieuse, de l'âge, du handicap ou d'une maladie grave.» [18].
- ***Davidson et al. :*** « Langage utilisé pour exprimer la haine envers un groupe ciblé ou destiné à être désobligeant, à humilier ou à insulter les membres du groupe. » [19].

3. Les méthodes de détection du Langage Agressif dans les Réseaux Sociaux

L'une des principales applications de l'exploration des médias sociaux est la détection automatique des événements et des comportements qui comprend l'identification du comportement des personnes dans des événements du monde réel en surveillant leurs interactions les unes avec les autres. Cette tâche dépend principalement des approches d'exploration de texte telles que la TAL et les algorithmes d'apprentissage automatique. Pour accomplir cette tâche, plusieurs caractéristiques et modèles communs doivent être identifiés. Ensuite, des algorithmes d'apprentissage automatique sont appliqués pour effectuer la tâche de classification afin d'obtenir le résultat ciblé des données.

3.1 Apprentissage automatique

Les algorithmes de classification sont utilisés pour effectuer la tâche de détection. En termes de classificateurs, les approches d'apprentissage automatique peuvent être classées en: approches supervisées, semi-supervisées et non supervisées[20] .

- **Apprentissage supervisé :** Cette approche dépend du domaine car elle repose sur un étiquetage manuel d'un grand volume de texte. La tâche d'étiquetage demande du temps et des efforts, mais elle est plus efficace pour les événements dépendants du domaine.
- **Apprentissage semi-supervisé :** Dans ce paradigme, les algorithmes sont entraînés à l'aide de données étiquetées et non étiquetées. L'utilisation de données étiquetées conjointement avec des données non étiquetées peut améliorer efficacement les performances.
- **Apprentissage non supervisé :** Il s'agit d'une approche indépendante du domaine et capable de gérer une diversité de contenus tout en conservant l'évolutivité. Il ne s'appuie pas sur le travail humain pour étiqueter un ensemble de formation à grand volume, mais extrait dynamiquement les termes clés liés au domaine.

Nous allons voir cette partie en détail dans le chapitre suivant.

3.2 Apprentissage Profond

L’apprentissage profond [21]est l’une des nombreuses approches de l’apprentissage automatique. Il a été inspiré par la structure et la fonction du cerveau, à savoir l’interconnexion de nombreux neurones. Les réseaux neuronaux artificiels (RNA) sont des algorithmes qui imitent la structure biologique du cerveau.

Dans les RNA, il y a des « neurones » qui ont des couches discrètes et des connexions avec d’autres « neurones ». Chaque couche sélectionne une caractéristique spécifique à apprendre, telle que des courbes / bords dans la reconnaissance d’image. C’est cette stratification qui donne son nom à l’apprentissage profond, la profondeur est créée en utilisant plusieurs couches par opposition à une seule couche.

4. Travaux connexes

Durant ce projet, nous avons eu l'occasion d'aborder plusieurs projets liés à notre travail de recherche. Nous présentons dans ce qui suit les différents travaux récents qui s'intéressent à la détection du langage agressif dans les réseaux sociaux en langue Anglaise. Nous allons présenter ces travaux en trois parties, ceux qui ont appliqué l'apprentissage supervisé, ceux qui ont appliqué l'apprentissage non supervisé et ceux qui ont appliqué l'apprentissage profond.

4.1 Les travaux en apprentissage supervisé

L'utilisation de méthodes de classification de l'apprentissage supervisé pour la détection du langage abusif n'est pas nouvelle. Dans ce qui suit nous allons voir quelques travaux qui ont appliqué ce modèle.

a. Une représentation en N-grammes

Davidson et al.[22]Proposent d'étiqueter les tweets en trois catégories : *discours de haine, langage offensant ou aucun des deux.* Ils ont utilisés une régression logistique avec régularisation L1 pour réduire la dimensionnalité des données. Ils ont testés ensuite une variété de modèles qui ont été utilisés dans des travaux antérieurs : *régression logistique, Bayes naïves, arbres de décision, forêts aléatoires et SVM linéaires.* Ils ont testé chaque modèle en utilisant une validation croisée 5 fois, en retenant 10% de l'échantillon pour évaluation afin d'éviter tout sur ajustement. Après avoir utilisé une recherche de grille pour itérer sur les modèles et les paramètres, ils ont constaté que la régression logistique et la SVM linéaire ont eu tendance à fonctionner nettement mieux que les autres modèles. C'est pour cela ils ont décidé d'utiliser une régression logistique avec régularisation L2 pour le modèle final.

Leur résultat montre que le modèle le plus performant a eu une précision globale de 0,91, rappel de 0,90 et score F1 de 0,90.

b. Une représentation en lexiques incorporant des mots

Del Vigna et al.[23]Ont utilisés un classificateur LSTM (Long Short Term Memory), ils ont évalués sur un petit échantillon de données Facebook pour seulement 2 classes (Haine,

Non-Haine), et 3 niveaux différents de force de haine. Leur résultat a montré que les classificateurs n'ont pas été en mesure de distinguer entre trois niveaux de haine.

c. Une représentation en N-grammes

Malmasi et Zampieri [24]ont classés les données dans l'une des trois classes:

- (HAINE): contient un discours de haine;
- (OFFENSIVE): contient un langage offensant mais pas de discours de haine
- (OK): aucun contenu offensant du tout.

Ils ont appliqué des méthodes de classification de texte pour distinguer le discours de haine, le blasphème et les autres textes. Ils ont appliqué des caractéristiques lexicales standard et un classificateur SVM linéaire pour établir une base de référence pour cette tâche. Leur meilleur résultat a été obtenu par un modèle de 4 grammes avec une précision de 78%. Les résultats présentés dans leur travail ont montré que distinguer le blasphème du discours de haine est une tâche très difficile.

d. Une représentation en Linguistiques

Par ailleurs, **Wiegand et al**[25]supposaient qu'ils pouvaient filtrer les mots abusifs des expressions polaires négatives. Ils ont profité d'un lexique de base en prenant un petit sous-ensemble d'expressions polaires négatives, puis via le crowdsourcing[23], les mots abusifs ont été étiquetés.

e. Une représentation en Caractère word2vec

Park et Fung[26] ont comparé les performances des classificateurs en une et deux étapes en utilisant l'ensemble de données fourni par **Waseem et Hovy**[27]. Sur la base de leurs résultats, ils pensent que la combinaison de 2 classificateurs (par exemple CNN et régression logistique) peut augmenter les performances.

f. Une représentation Aléatoire

Badjatiya et al.[28]ont abordé la question avec un modèle d'apprentissage supervisé basé sur un réseau neuronal. Leur méthode a obtenu un score plus élevé sur le même ensemble de données de tweets que toute solution d'apprentissage non supervisée connue à ce jour. Cette

[23] Le crowdsourcing est une pratique qui consiste à faire participer des consommateurs ou le grand public à la création d'un produit ou à la réalisation d'un service

solution utilise un modèle LSTM, avec des fonctionnalités extraites par caractère n-grammes, et assisté par Gradient BoostedDecisionTrees.

4.2 Les travaux en apprentissage non supervisé

Les approches d'apprentissage non supervisées sont assez courantes pour détecter des messages offensifs dans le texte en appliquant des concepts du TAL pour exploiter les caractéristiques syntaxiques et lexicales des phrases, dans ce qui suit nous allons présenter quelques travaux récents.

a. Une représentation en Lexique et syntaxique

Chen et al.[29]Ont utilisé les commentaires YouTube comme un ensemble de données pour détecter un langage offensant. Ils ont utilisé une combinaison de fonctionnalités lexicales et syntaxiques et ont incorporé le style d'écriture de l'utilisateur pour prédire le comportement de l'utilisateur à l'avenir.

b. Une représentation en Modélisation de sujets

Une approche similaire a été proposée par **Xiang et al.**[30]pour détecter le contenu offensant sur Twitter. Leurs caractéristiques étaient principalement basées sur les régularités linguistiques des termes profanes également basées sur la modélisation de sujets statistiques sur un ensemble de données relativement volumineux pour un scénario d'apprentissage en profondeur.

c. Une représentation en N-grammes

Waseem et Hovy[27]ils ont présenté une liste de critères basés sur la théorie critique de la race pour identifier les insultes racistes et sexistes.

Ils ont voulu d'évaluer l'influence de différentes caractéristiques sur la prédiction dans une tâche de classification. C'est pour cela ils ont utilisé un classificateur de régression logistique et une validation croisée 10 fois pour tester l'influence de diverses caractéristiques sur les performances de prédiction.

Leur résultat montre que l'utilisation de caractères n-grammes de longueurs jusqu'a 4 a fourni les meilleurs résultats.

d. Une représentation en Intégration de mots

Pour finir, **Chen et al.**[31]Ont utilisé FastText comme leur classificateur de réseau neuronal pour détecter les textes abusifs de diverses plateformes de réseaux sociaux. Ils ont constaté que les performances de FastText sont inférieures à celles de l'utilisation de SVM comme classificateur.

4.3 Les travaux en apprentissage profond

Des travaux récents se sont tournés vers l'emploi de l'apprentissage en profondeur pour de telles tâches.

a. Une représentation enword2vec

Gambäck et Sikdar[32]ont appliqué une approche d'apprentissage en profondeur sur l'ensemble de données de **Waseem et Hovy**[27]. Leurs résultats ont été impressionnants en termes de précision et de rappel.

b. Une représentation en Enrobage

Le même corpus a également été utilisé par **Badjatiya et al.**[33]pour comparer différentes combinaisons de modèles d'apprentissage en profondeur.

c. Une représentation en N-grammes

Dans leur travail**Koratana& Hu** [34], les chercheurs ont montré que les modèles d'apprentissage en profondeur peuvent être performants dans la tâche de détection de commentaires toxiques. Les deux modèles Bi-GRU / LSMC et VDCNN ont pu produire une bonne F1 et une plus grande précision par rapport à une base de référence de régression logistique assez forte (quoique simples). Ils ont abordé la question du temps de calcul et le coût pour les modèles d'apprentissage en profondeur en analysant leurs vitesses d'inférence. En effet, un énorme inconvénient de CNNs et RNNs sont là formationlente et temps des tests par rapport aux «*méthodes classiques*». En tant que tel, ils ont proposé d'adopter un modèle en cascade, ce qui a été rendu possible par la performance solide.

Une limitation de leur projet est que leurs modèles sont formés sur un ensemble de données de commentaires toxiques Google Jigsaw.

5. Discussion

Le tableau suivant présente un résumé de tous les travaux cités ci-dessus, ils sont organisés selon leurs séries chronologiques respectives. Ce tableau peut servir de référence rapide pour les travaux clés effectués dans la détection automatique dans les médias sociaux. Toutes les approches et leurs résultats d'expériences avec les métriques: Précision (P), Rappel (R)[24] , F1-Score (F), sont répertoriés de manière concise. Pour la colonne des résultats, les meilleurs résultats de chaque travail sont indiqués.

[24]**Précision** est la proportion des items pertinents parmi l'ensemble des items proposés ; le **rappel** est la proportion des items pertinents proposés parmi l'ensemble des items pertinents

Auteur	année	Plateforme	ML approche	Représentation	Algorithme	P	R	F
APPRENTISSAGE SUPERVISÉ								
Davidson et al. [22]	2017	Twitter	Supervisé	N-grammes	SVM	0.91	0.90	0.90
Del Vigna et al. [23]	2017	Facebook	Supervisé	Morphosyntaxique, lexiques incorporant des mots	SVM	0.75	0.68	0.71
					RNN & LTSM	0.70	0.75	0.72
Malmasi et all [24]	2018	Twitter	Supervisé	N-grammes, Skip grammes, groupes de mots hiérarchiques	RBF kernel , SVM	0.78	0.80	0.79
Wiegand et al [25]	2018	Twitter, wikipedia,useNet	Supervisé	Linguistiques et intégration de mots	SVM	0.82	0.80	0.81
Park et Fung [26]	2017	Twitter	Supervisé	Caractère word2vec	CNN	0.71	0.75	0.73
Badjatiya et al. [28]	2017	Twitter	Supervisé	Aléatoire, Enrobage	LSTM et GBDT	0.93	0.93	0.93
APPRENTISSAGE NON SUPERVISÉ								
Chen et al. .[29]	2012	Youtube	Non supervisé	Lexique et syntaxique	Règles de match	0.98	0.94	-
Xiang et a.[30]	2016	Twitter	Semi-supervisé	Modélisation de sujets	Régression logistique	-	-	0.84
Waseem et Hovy(30	2016	Twitter	Non supervisé	n-grammes	Régression logistique	0.72	0.77	0.73
Chen et al.[31]	2017	Youtube, Myspace, SlashDot	Non supervisé	Intégration de mots	FastText	-	0.76	-
APPRENTISSAGE EN PROFONDEUR								
Gambäck et Sikdar[32]	2017	Twitter	Algorithme en profondeur	Caractère N-grammes, word2vec	CNN	0.85	0.72	0.78
Badjatiya et al. .[33]	2017	Twitter	Algorithme en profondeur	Enrobage	LSTM et GBDT	0.93	0.93	0.93
Koratana& Hu [34]	2018	Google Jigsaw,	Algorithme en profondeur	Caractère N-grammes	CNN & RNN	-	-	-

Tableau 3 : Résumé de l'état actuel de la détection des comportements antisociaux.

Les travaux cité ci-dessous ont été réalisé dans le but de détecter le langage agressif dans les médias sociaux, nous constatons que:

- La détection du langage abusif n'est pas un simple repérage de mots clés, c'est une tâche complexe avec de nombreux défis.
- Certains travaux classent le langage abusif comme : *abusif ou propre*. d'autres en : *obscène, offensant ou propre* ainsi que *discours de haine, langage offensant ou aucun des deux*.
- Le choix de l'approche d'apprentissage machine le plus approprié est une décision difficile.
- La majorité des chercheurs se sont appuyés sur des approches d'apprentissage automatique supervisé dans leur tâche de détection automatique.
- Les approches non supervisées viennent au prochain rang de popularité.
- Les approches semi-supervisées sont les techniques les moins utilisées.
- Les recherches sont orientées vers l'apprentissage en profondeur pour résoudre des tâches d'apprentissage complexes.
- Les chercheurs ont affirmé que l'apprentissage en profondeur est puissant lorsqu'il s'agit de trouver une représentation des données pour la classification et, évidemment, il a un avenir prometteur dans le domaine de la détection automatique.
- Il existe deux architectures principales pour les réseaux de neurones profonds qui sont généralement utilisées pour les tâches TAL, ces modèles sont: RNN et CNN.
- Les travaux proposés ne comparent pas leurs résultats aux travaux qui les précèdent. On n'a donc pas une idée sur la meilleure solution proposée jusque-là.
- Les données d'expérimentation sont généralement des données issues de réseaux sociauxréels. Seulement, la taille des réseaux testés reste limitée.
- Il y a des bons résultats dans les langues étrangères, car il existe plusieurs efforts et des outils qui facilitent la continuité de ces travaux dans ce domaine.
- Choisir d'adopter le Deep Learning nécessite un engagement à la fois dans la préparation et la formation du modèle avec une grande quantité de données.

En conclusion, nous devons considérer tous les facteurs qui peuvent influer sur notre décision dans le bon choix de l'approche. Par exemple, un facteur majeur est la taille du corpus, car certains algorithmes ML (Machine Learning) fonctionnent assez bien avec de

petits ensembles de données. D'autres, comme les réseaux neuronaux, nécessitent une formation plus intensive et complexe.

6. Conclusion

Dans ce chapitre, nous avons commencé par définir le langage agressif, nous nous sommes intéressés aux méthodes automatiques de détection du langage agressif sur les réseaux sociaux, tout en montrant les théories d'apprentissage profond et automatique où nous allons les voir en détails dans le chapitre suivant, ensuite nous avons fini par citer quelques travaux récents dans ce domaine avec une comparaison entre eux.

Dans le prochain chapitre, on va voir quelques études théoriques sur l'apprentissage automatique, leurs modèles et nous allons décrireles algorithmes que nous allons utiliser dans notre travail.

Chapitre III :
Algorithmes d'apprentissage automatique

1. Introduction

La détection automatique du langage abusif est une tâche difficile mais importante pour les médias sociaux. Malgré le nombre d'approches récemment proposées dans le domaine de recherche du Traitement Automatique du Langage Naturel (TALN) pour détecter ces formes de langage abusif, la question de l'identification du langage abusif à grande échelle reste un problème non résolu. L'objectif de notre travail est de proposer et expérimenter des approches pour la détection des expressions dans les postes des utilisateurs qui expriment des insultes véhémentes ou de l'agressivité quel quesoit son sujet.

Dans ce chapitre nous allons présenter l'apprentissage automatique, ses modèles ainsi pour faciliter la compréhension du concept utilisé, nous allons expliquer, en détail, les algorithmes d'apprentissage automatique supervisé que nous allons utiliser.

2. Apprentissage automatique

L'apprentissage automatique fait référence au développement, l'analyse et l'implémentation de méthodes qui permettent à une machine (au sens large) d'évoluer et de remplir des tâches associées à une intelligence artificielle grâce à un processus d'apprentissage. Cet apprentissage permet d'avoir un système qui s'optimise en fonction de l'environnement, les expériences et les résultats observés[35].

En générale, l'apprentissage automatique c'est la capacité de la machine à apprendre à faire mieux à l'avenir sur la base de ce qui a été connu dans le passé.

«L'apprentissage dénote des changements dans un système qui ... lui permettent de faire la même tâche plus efficacement la prochaine fois» ***Herbert Simon***.

2.1 Définition

Un programme d'ordinateur est capable d'apprendre à partir d'une expérience E et par rapport à un ensemble T de tâches et selon une mesure de performance P, si sa performance à effectuer une tâche de T, mesurée par P, s'améliore avec l'expérience E [35].

3. Les piliers de l'apprentissage automatique

L'apprentissage automatique (Le machine Learning) repose sur deux piliers fondamentaux :

- d'une part, **les données**, qui sont les exemples à partir desquels l'algorithme va apprendre ;
- d'autre part, **l'algorithme d'apprentissage**, qui est la procédure que l'on fait tourner sur ces données pour produire un modèle. On appelle entraînement le fait de faire tourner un algorithme d'apprentissage sur un jeu de données.

Ces deux piliers sont aussi importants l'un que l'autre. D'une part, aucun algorithme d'apprentissage ne pourra créer un bon modèle à partir de données qui ne sont pas pertinentes. D'autre part, un modèle appris avec un algorithme inadapté sur des données pertinentes ne pourra pas être de bonne qualité.

4. Types d'apprentissage automatique

Les algorithmes d'apprentissage peuvent se catégoriser selon le mode d'apprentissage qu'ils emploient : ***apprentissage supervisé, apprentissage semi-supervisé, apprentissage non supervisé, apprentissage par renforcement***, etc. On trouvera dans la figure9, une représentation graphique des différents types.

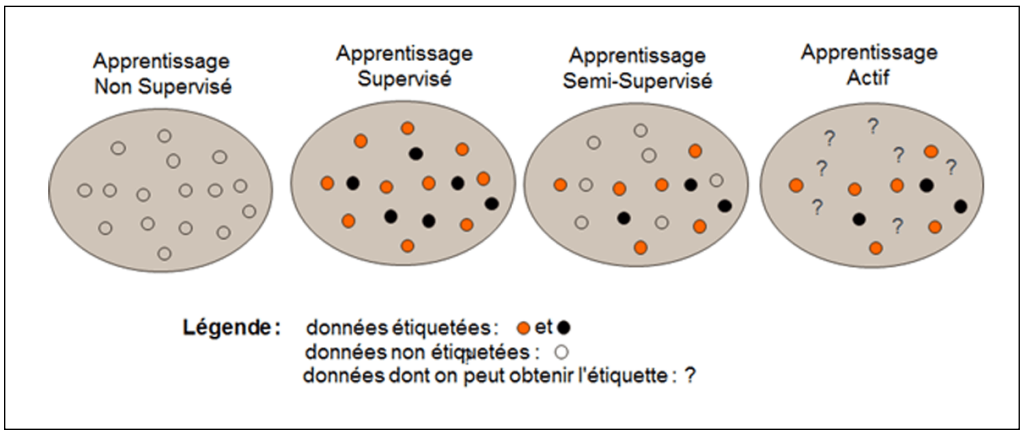


Figure 9 : Les différents modes d'apprentissage[36] .

4.1 Apprentissage supervisé

La formulation du problème de l'apprentissage supervisé est simple : on dispose d'un nombre fini d'exemples d'une tâche à réaliser, sous forme de paires (entrée, sortie désirée), et

on souhaite obtenir, d'une manière automatique, un système capable de trouver de façon relativement fiable la sortie correspondante à toute nouvelle entrée qui pourrait lui être présentée. Un expert (ou oracle) doit préalablement correctement étiqueter des exemples. L'apprenant peut alors trouver ou approximer la fonction qui permet d'affecter la bonne « étiquette » à ces exemples[37].

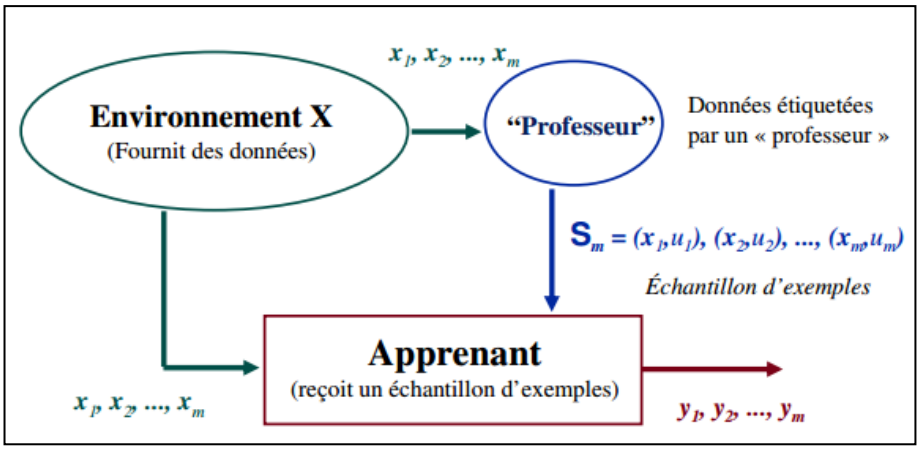


Figure 10 : Apprentissage supervisé[37].

La figure 10 montre qu'un expert est employé pour étiqueter correctement des exemples.L'apprenant doit alors trouver ou approximer la fonction qui permet d'affecter la bonne étiquette à ces exemples.

En résumé :

- **Source d'apprentissage:** des données annotées (nous avons les résultats attendus)
- **Retour d'information:** direct; à partir des résultats attendues.
- **Fonction:** prédire les futurs résultats.

On distingue en général deux types de problèmes auxquels l'apprentissage supervisé est appliqué. Ces tâches diffèrent essentiellement par la nature des paires (entrée, sortie) qui y sont associées :

- **Classification :** Dans les problèmes de classification, l'entrée correspond à une instance d'une classe, et la sortie qui y est associée indique la classe. Par exemple pour un problème de reconnaissance de visage, l'entrée serait l'image bitmap d'une personne telle que fournie par une caméra, et la sortie indiquerait de quelle personne il s'agit (parmi l'ensemble de personnes que l'on souhaite voir le système reconnaître).

- **Régression :** Dans les problèmes de régression, l'entrée n'est pas associée à une classe, mais dans le cas général, à une ou plusieurs valeurs réelles (un vecteur). Par exemple, pour une expérience de biochimie, on pourrait vouloir prédire le taux de réaction d'un organisme en fonction des taux de différentes substances qui lui sont administrées.

4.2Apprentissage semi-supervisé

L'apprentissage non-supervisé est une classe de techniques d'apprentissage automatique qui utilise un ensemble de données étiquetées et non étiquetées. Il se situe ainsi entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non supervisé qui n'utilise que des données non étiquetées. Il a été démontré que l'utilisation de données non étiquetées, en combinaison avec des données étiquetées, permet d'améliorer significativement la qualité de l'apprentissage.

Il est effectué de manière probabiliste ou non, il vise à faire apparaitre la distribution sous-jacente des « exemples » dans leur espace de description. Il est mis en œuvre quand des données (ou « étiquettes ») manquent... Le modèle doit utiliser des exemples non-étiquetés pouvant néanmoins renseigner [37].

C'est un mélange entre le principe de base de l'apprentissage superviser et non superviser utiliser lorsque l'étiquetage des exemples est couteux ce type d'apprentissage qui a connu beaucoup d'importance au cours des dernières années a permet d'améliorer les résultats d'apprentissage**.**

Un superviseur est partiellement disponible et étiquette quelques données. L'algorithme va alors apprendre la tâche de classification en se basant sur les étiquettes posées par le superviseur et en découvrant par lui-même les informations manquantes (les exemples non étiquetés).

Les algorithmes semi-supervisés fonctionnent sur les deux même phases mais acceptent en plus des données non étiquetées pendant la phase d'entraînement comme la figure suivante ou l'ensemble d'apprentissage (x1 non étiqueté et x2 étiqueté).

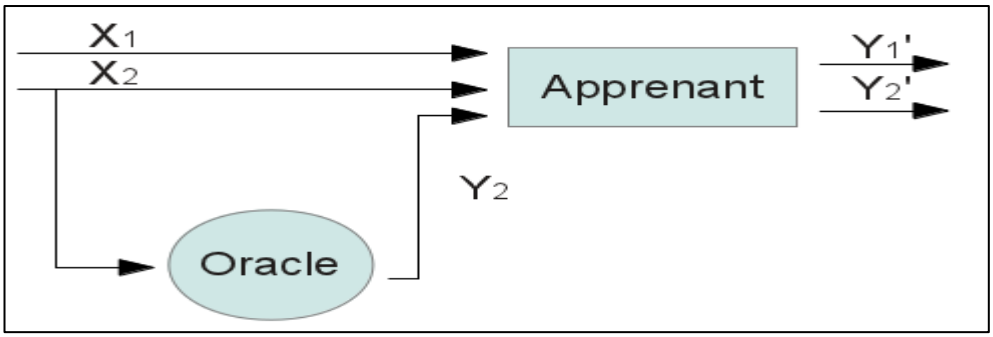


Figure 11 : Fonctionnement d'apprentissage semi supervisé[37].

4.3 Apprentissage non-supervisé :

Quand le système ou l'opérateur ne dispose que d'exemples, mais non d'étiquettes, et que le nombre de classes et leur nature n'ont pas été prédéterminés, on parle d'apprentissage non supervisé (ou clustering). Aucun expert n'est disponible ni requis. L'algorithme doit découvrir par lui-même la structure plus ou moins cachée des données [37].

En résumé:

- **Source d'apprentissage:** des données non étiquetées
- **Retour d'information:** pas de retour; on dispose seulement des données en entrée.
- **Fonction:** rechercher les structures cachées dans les données.

5. Apprentissage profond

L'apprentissage profond est un nouveau domaine de recherche du l'apprentissage automatique, qui a été introduit dans le but de rapprocher le MLde son objectif principal : l'intelligence artificielle. Il concerne les algorithmes inspirés par la structure et le fonctionnement du cerveau. Ils peuvent apprendre plusieurs niveaux de représentation dans le but de modéliser des relations complexes entre les données. Il est basé sur l'idée des réseaux de neurones artificielles et il est taillé pour gérer de larges quantités de données en ajoutant des couches au réseau.[38]. La figure suivante montre la relation entre l'apprentissage automatique, profond et l'IA.

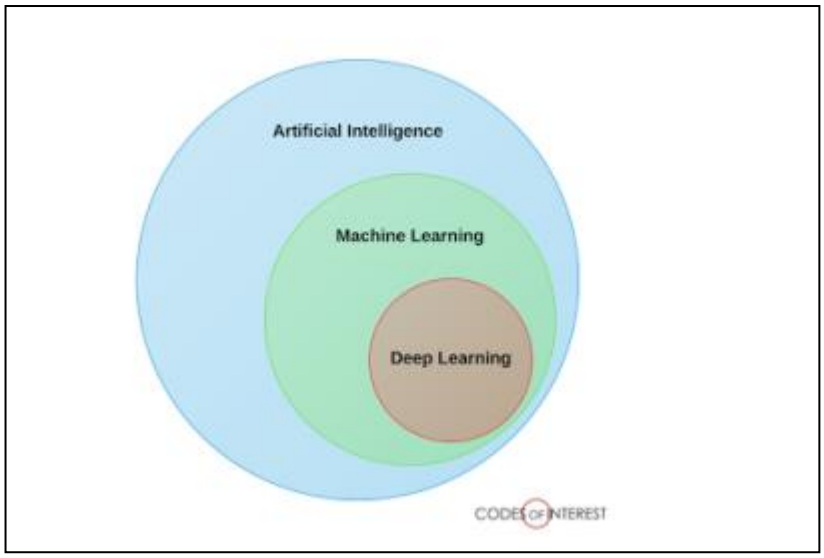


Figure 12 : Relation entre Machine learning, Deep learning et IA [38]

6. Algorithmes de classification

Ci-après, nous présentons quelques algorithmes de classification.

6.1 Machine à vecteurs supports

Une machine à vecteurs de support, traduction littérale pour Support Vector Machine, est un algorithme d'apprentissage automatique supervisé qui peut être utilisé à des fins de classification et de régression. Les SVM sont plus généralement utilisés dans les situations de classification.[39].

Les SVM reposent sur l'idée de trouver un hyperplan qui divise au mieux un jeu de données en deux classes, comme le montre l'image ci-dessous.

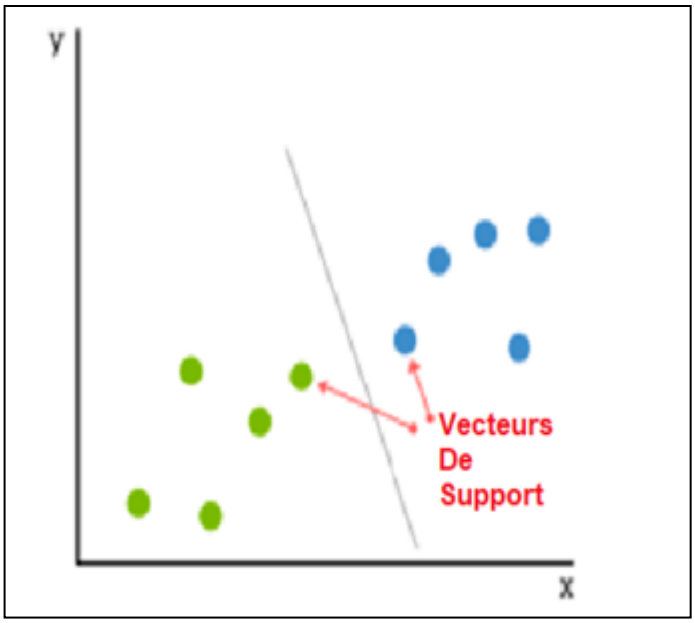


Figure 13 : Machine à vecteurs de supports[39].

- **Les Avantages**

  • Sa grande précision de prédiction.

  • Fonctionne bien pour de plus petits datasets.

  • Ils peuvent être plus efficaces car ils utilisent un sous-ensemble de points d'entraînement.

- **Les inconvénients**

  • Ne convient pas à des jeux de données plus volumineux, car le temps d'entraînement avec les SVM peut être long

  • Moins efficace sur les jeux de données contenant des bruits.

6.2 Régression Logistique

La régression logistique est un algorithme de classification utilisé pour attribuer des observations à un ensemble discret de classes. Contrairement à la régression linéaire qui génère des valeurs numériques continues, la régression logistique transforme sa sortie à l'aide de la fonction sigmoïde logistique pour renvoyer une valeur de probabilité qui peut ensuite être mappée sur deux ou plusieurs classes discrètes.[40]

Type de régression logistique :

- Binaire (réussite/ échec)
- Multi (chats, chiens, moutons)
- Ordinal(faible, moyen, élevé).

Dans notre travail, nous allons utiliser la régression logistique binaire. Pour mieux la comprendre nous allons présenter un exemple en dessous.

Supposons que nous recevons des données sur les résultats des examens des étudiants et que notre objectif est de prédire si un étudiant réussira ou échouera en fonction du nombre d'heures passées endormi et des heures passées à étudier. Nous avons deux caractéristiques (heures de sommeil, heures étudiées) et deux classes: réussi (1) et échoué (0)[40].

Etudiées	Sommeil	Classe
4.85	9.36	1
8.62	3.32	0
5.43	8.23	1
9.21	6.43	0

Tableau 4 : Exemple d'un résultat d'examen des étudiants[40]

Graphiquement, nous pouvons représenter ces données avec un nuage de points comme illustré dans la figure suivante :

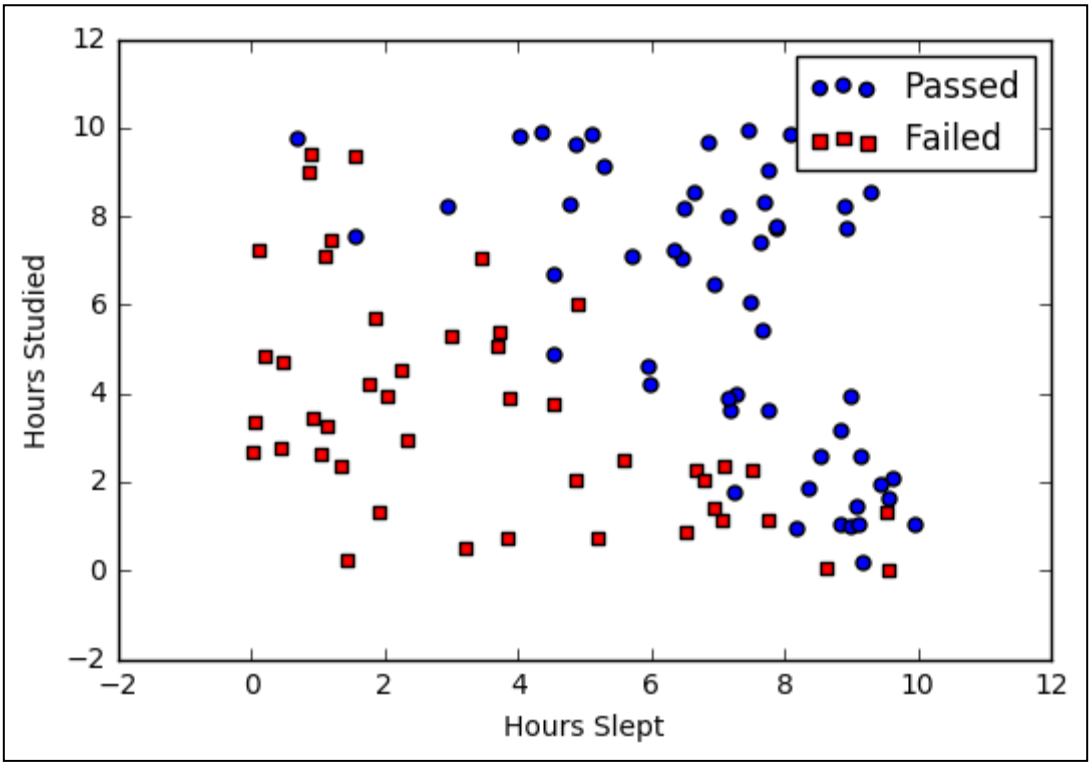


Figure 14 : Nuage de points[40].

- **Activation sigmoïde :**

Afin de mapper les valeurs prédites aux probabilités, nous utilisons la fonction sigmoïde. La fonction mappe toute valeur réelle dans une autre valeur entre 0 et 1.

Dans l'apprentissage automatique, nous utilisons sigmoïde pour mapper les prédictions aux probabilités[40].

La forme Mathématique de la fonction sigmoïde est :

$$S(z) = \frac{1}{1 + e^{-z}}$$

Telle que :

S (z) = sortie entre 0 et 1 (estimation de probabilité)

z = entrée dans la fonction (prédiction de l'algorithme, par exemple mx + b)

e = base du logarithme naturel Graphique.

Dans la figure ci-dessous un graphe qui illustre la fonction sigmoïde

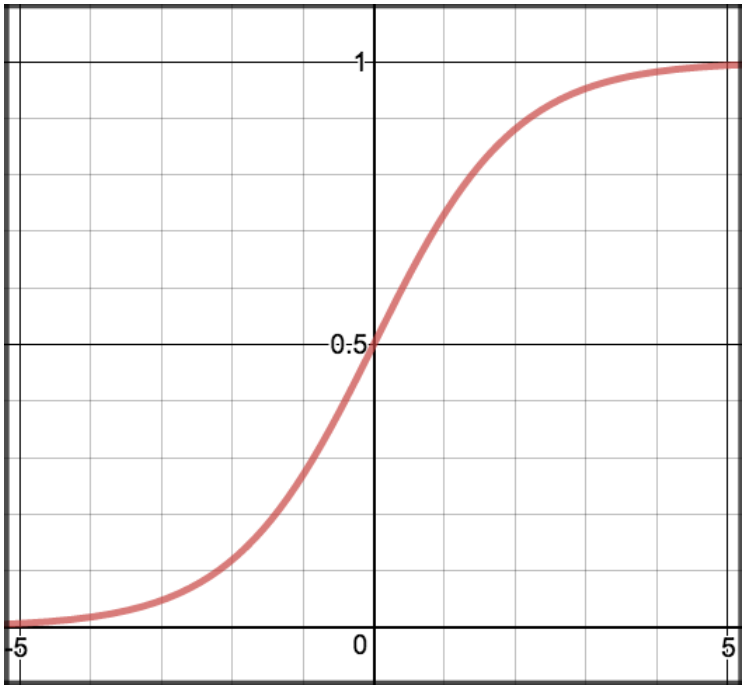


Figure 15 : Graphe de la fonctionsigmoïde[40].

- **Avantage :**
  - Le modèle est facile à interpréter
- **Inconvénients** :
  - Sensible aux bruits
  - Négligence des interactions entre les variables prédictives

6.3 RandomForest

Les forêts aléatoires est un algorithme d'apprentissage supervisé. Il peut être utilisé à la fois pour la classification et la régression. C'est également l'algorithme le plus flexible et le plus facile à utiliser. Une forêt est composée d'arbres. On dit que plus il y a d'arbres, plus la forêt est robuste. Les forêts aléatoires créent des arbres de décision sur des échantillons de données sélectionnés au hasard, obtiennent des prévisions à partir de chaque arbre et sélectionnent la meilleure solution au moyen du vote. Il fournit également un assez bon indicateur de l'importance des fonctionnalités[41].

6.3.1 Fonctionnement de l'algorithme de RandomForest :

L'algorithme des Random Forest fonctionne en quatre étapes:

- Sélection des échantillons aléatoires dans un ensemble de données donné.
- Constructiond'un arbre de décision pour chaque échantillon et obtentiond'un résultat de prédiction à partir de chaque arbre de décision.
- -Etablissement d'un vote pour chaque résultat prévu.
- -Sélectiondu résultat de la prédiction avec le plus de votescomme prédiction finale.

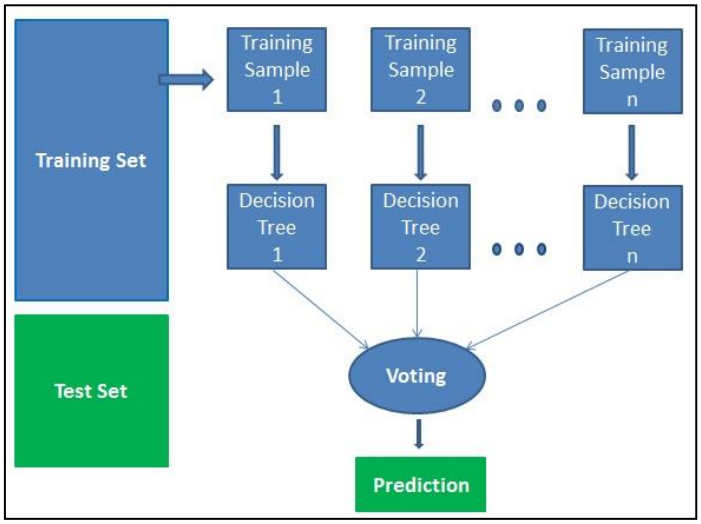


Figure 16 : Fonctionnement de l'algorithme RandomForest[41].

- **Avantages :**
  - C'est un des meilleurs algorithmes pour ce qui est de la précision.

- Incorporation de la validation croisée.
- Il conserve une bonne puissance de prédiction même si il y a des données manquantes.
- Il ne souffre pas du sur-apprentissage.

- **Inconvénients**
  - Une implémentation difficile.

6.4 XGBoost

XGBoost est devenu un outil largement utilisé parmi les concurrents de Kaggle et les Data Scientists de l'industrie, car il a été testé pour la production sur des problèmes à grande échelle. Il s'agit d'un outil très flexible et polyvalent qui peut gérer la plupart des problèmes de régression, de classification et de classement ainsi que les fonctions d'objectif créées par l'utilisateur.

En tant que logiciel open source, il est facilement accessible et peut être utilisé via différentes plates-formes et interfaces. La portabilité et la compatibilité incroyables du système permettent son utilisation sur les trois systèmes d'exploitation Windows, Linux et OS X. Il prend également en charge la formation sur des plates-formes cloud distribuées comme AWS, Azure, GCE entre autres et il est facilement connecté à des systèmes de flux de données cloud à grande échelle tels que comme Flink et Spark.[42].

6.4.1 Fonctionnement de l'algorithme XGBoot

Avant de passer à un exemple de l'algorithme, définissons quelques définitions de base pour obtenir une compréhension intuitive et complète de cet outil populaire.

Tout d'abord, il s'agit d'une méthode d'ensemble qui cherche à créer un classificateur fort (modèle) basé sur des classificateurs «faibles». Dans ce contexte, faible et fort fait référence à une mesure de la corrélation des apprenants avec la variable cible réelle. En ajoutant les modèles les uns sur les autres de manière itérative, les erreurs du modèle précédent sont corrigées par le prédicateur suivant, jusqu'à ce que les données d'apprentissage soient prédites ou reproduites avec précision par le modèle.
Maintenant, le renforcement du gradient comprend également une méthode d'ensemble qui ajoute séquentiellement des prédicateurs et corrige les modèles précédents.

Cependant, au lieu d'attribuer des poids différents aux classificateurs après chaque itération, cette méthode adapte le nouveau modèle aux nouveaux résidus de la prédiction précédente, puis minimise la perte lors de l'ajout de la dernière prédiction.

Donc, à la fin, vous mettez à jour votre modèle en utilisant la descente de gradient et donc le nom, boosting de gradient. Ceci est pris en charge pour les problèmes de régression et de classification.

Dans qui suit une illustration simple de XGboost.

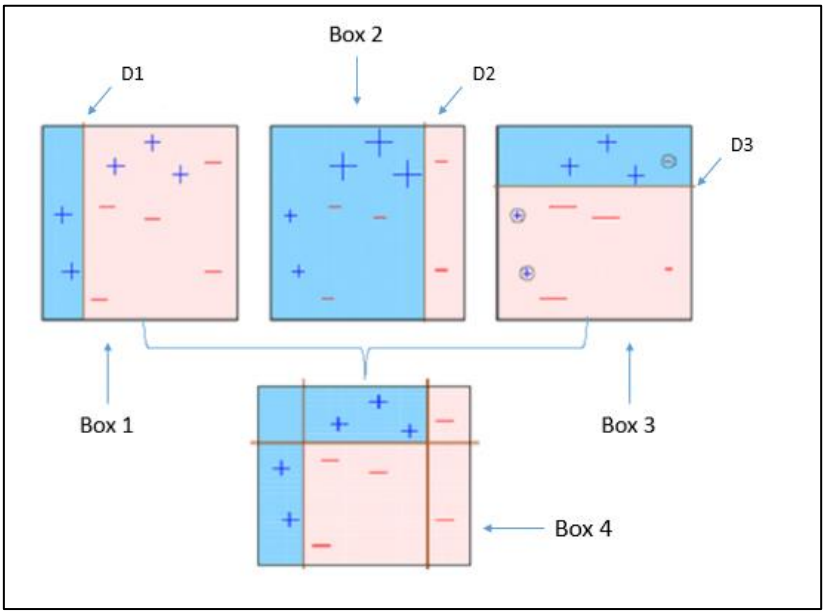


Figure 17 : Une illustration simple de XGboost[43].

La figure 17 montre quatre classificateurs (dans 4 cases), illustrés ci-dessus, tentent de classer les classes + et - de manière aussi homogène que possible.

Dans **le premier encadré** : crée une ligne verticale (divisée) à D1. Il indique que tout ce qui se trouve à gauche de D1 est + et tout ce qui se trouve à droite de D1 est - . Cependant, ce classificateur classe mal trois points +.

Dans **le deuxième encadré :**donne plus de poids aux trois points + mal classés (voir la plus grande taille de +) et crée une ligne verticale à D2. Encore une fois, tout ce qui se trouve à droite de D2 est - et à gauche est +. Pourtant, il fait des erreurs en classant incorrectement trois points.

Dans **le troisième encadré :** Encore une fois, donne plus de poids aux trois points mal classés et crée une ligne horizontale à D3. Pourtant, ce classificateur ne parvient pas à classer correctement les points (dans les cercles).

Et dans **le dernier encadré :** Il s'agit d'une combinaison pondérée des classificateurs faibles (encadrés 1,2 et3). Comme vous pouvez le voir, il classe bien tous les points.

6.5 MLPClassifier

Un MLP (pour Multi-Layer Perceptron) ou un réseau neuronal multicouche définit une famille de fonctions. Considérons d'abord le cas le plus classique d'un réseau neuronal à couche cachée unique, cartographiant un d-vecteur à un m-vecteur (par exemple pour la régression):

$$g(x) = b + W \tanh(c + V_x)$$

Où*x*est un d-vecteur (l'entrée),*V*est une matrice$K \times D$ (appelée poids entrée-caché),*c*est un*k*-vecteur (appelé décalages d'unités cachées ou biais d'unités cachées),*b*est un*m*-vecteur (appelé décalage des unités de sortie ou biais des unités de sortie), et*W e*st une matrice $m \times h$ (appelée poids caché à la sortie)[44].

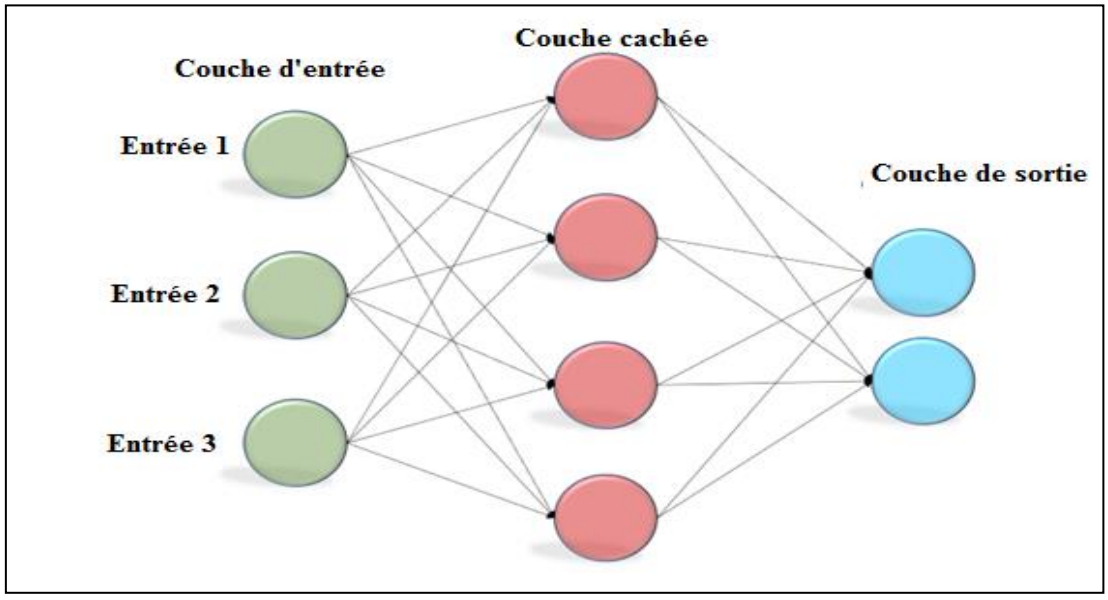


Figure 18 : MLPCLASSIFIER [45]

- **Les avantages :**

- Capacité d'apprendre des modèles non linéaires.

- Capacité d'apprendre des modèles en temps réel (apprentissage en ligne) à l'aide de partial_fit.[25]

- **Les inconvenants :**

- Les MLP avec des couches cachées ont une fonction de perte non convexe où il existe plus d'un minimum local. Par conséquent, différentes initialisations de poids aléatoires peuvent conduire à une précision de validation différente.

- Le MLP nécessite le réglage d'un certain nombre d'hyperparamètres tels que le nombre de neurones cachés, de couches et d'itérations.

- Le MLP est sensible à la mise à l'échelle des fonctionnalités.

7. Conclusion

Dans ce chapitre, nousavons passé en revue un peu de théories sur l'apprentissage automatique, Ses modèles et quelques algorithmes qui nous semblons capables de résoudre notre problème.

Dans le dernier chapitre, nous allons décrire notre propre système, où nous allons collecter et étiqueter un ensemble de données à partir des réseaux sociaux, puis appliquer un ensemble d'algorithmes d'apprentissage automatique et les évaluer tout en montrantles étapes du processus de traitement de façon détaillée. Pour terminer avec les tests et validation du système.

[25]Ajustement incrémentiel sur un lot d'échantillons, le partial fit est une méthode qui doit être appelé plusieurs fois consécutivement sur différents morceaux d'un ensemble de données afin de mettre en œuvre un apprentissage hors cœur ou en ligne.

Chapitre IV :
Conception et Implémentation de la solution

1. Introduction

Ce chapitre est consacré à la partie conception et implémentation de notre solution. Dans un premier temps, nous allons présenter les données utilisées ainsi que l'environnement de développement ensuite nous montrons notre architecture proposée en détaillantles différentes étapes de déroulement duprogramme et enfin, nous allons présenter les résultats obtenus.

Notre travail est basé sur l'exploitation de plusieurs méthodes de l'apprentissage automatique, consistant à effectuer une classification sur un corpus pour détecter le langage agressif. Pour cela nous avons utilisé un dataset collectées par des experts via le réseau social twitter, ce dataset contient plus de 49 000 tweets.

2. Description des données

Cette section est consacrée à la description des données utilisées en entrée et à celles des résultats produits.

2.1 Données en entrée

Nous avons choisi, comme mentionné précédemment, de nous concentrer sur le réseau social Twitter. Donc l'entrée est un texte court (tweet).

- **Tweet** : Les données en entrée que nous exploitons sont des textes qui représentent des statuts en anglais postés sur le réseau social Twitter.

- **Caractéristiques d'un Tweet :** Un tweet est composé de plusieurs informations :
    - Le nom de l'utilisateur, son pseudo et son image de profil.
    - La date et l'heure d'émission du tweet.
    - Le nombre de "j'aime" et de retweet.
    - Le lieu où a été posté le tweet.

2.2Données en sortie

En sortie, nous avonsun fichier CSV qui contient des tweets étiquetés par (0 ou 1). Tels que :

- 0 signifie que ce tweet est non abusif.
- 1 signifie que ce tweet est abusif.

3. Matériels utilisés

Le matériel hardware utilisé pour les tests est résumé dans le tableau suivant :

Machine 1	Machine 2
- Processeur : Intel(R) Core (TM) i5-4200U CPU @ 1.60GHz 2.30 GHz - Mémoire installée (RAM): 6.00 Go - Système d'exploitation : Système d'exploitation 64 bits Windows 10 professionnel.	- Processeur : Intel(R) Core(TM) i5-6705U CPU @ 1.70GHz 1.70 GHz - Mémoire installée (RAM):8.00 Go - Système d'exploitation : Système d'exploitation 64 bits Windows 10.

Tableau 5: Matériel utilisé.

4. Outils utilisés

Pour l'implémentation de notre solution, nous avons opté pour le langage ***Python 3.7*** qui est l'un des langages de programmation les plus intéressants du moment, facile à apprendre, il s'agit d'un langage de programmation interprété, qui ne nécessite donc pas d'être compilé pour fonctionner.

De plus nous avons utilisé ***Spyder*** comme environnement de développement qui est (nommé Pydee dans ses premières versions) est un environnement de développement pour Python. Libre (Licence MIT) et multiplateforme (Windows, Mac OS, GNU/Linux), il intègre de nombreuses bibliothèques d'usage scientifique : *Matplotlib, NumPy, SciPy et IPythn*.
Il permet de compléter le code de manière intelligente, d'inspecter le code, de mettre en évidence à la volée les erreurs et de les corriger rapidement, ainsi que la refacturation du code automatiquement et offre des fonctionnalités de navigation avancées.

Il nous permet d'utiliser différentes API (Application Programming Interfaces) à partir de plusieurs langages de programmation comme Java et C++...etc. Nous avons utilisé dans notre application quelques bibliothèques pour assurer certaines fonctionnalités. Citons :

- **La bibliothèque « re » :** qui fournit des opérations de correspondance d'expressions régulières similaires à celles trouvées en Perl. Les modèles et les chaînes à rechercher peuvent être des chaînes Unicode ainsi que des chaînes 8 bits.
- **La bibliothèque « nltk » :** qui fournit des interfaces faciles à utiliser à plus de 50 corpus et ressources lexicales telles que WordNet, ainsi qu'une suite de bibliothèques de traitement de texte pour la classification, la tokenisation, le stemming, le balisage, l'analyse et le raisonnement sémantique, des wrappers pour les bibliothèques NLP de qualité industrielle, et un forum de discussion actif.
- **La bibliothèque « NumPy » :** La bibliothèque **NumPy** permet d'effectuer des calculs numériques avec Python. Elle introduit une gestion facilitée des tableaux numériques.
- **La bibliothèque « Pandas » :** Pandas est une librairie python qui permet de manipuler facilement des données à analyser :

    - Manipuler des tableaux de données avec des étiquettes de variables (colonnes) et d'individus (lignes).
    - Ces tableaux sont appelés DataFrames, similaires aux dataframes sous R.
    - On peut facilement lire et écrire ces dataframes à partir ou vers un fichier tabulé.
    - On peut facilement tracer des graphes à partir de ces DataFrames grâce à matplotlib.

5. Architecture proposée

Les processus composant notre système sont présentés dans la figure 19, représentant le processus général de la méthodologie du système.

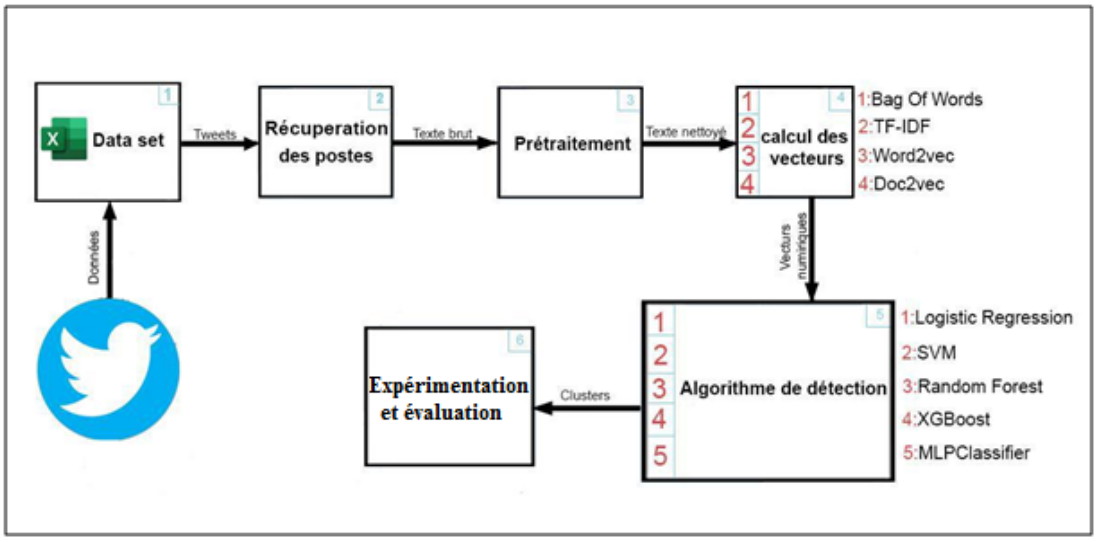


Figure 19 : Processus général de la méthodologie du système.

Dans ce qui suit, nous allons décrire chacune des phases de ce processus :

5.1 Phase 1 : le dataset

Pour ce projet nous avons travaillé sur plusieurs collections de tweets, déjà collectées par des experts[26]. La figure 20 montre le siteoù nous avons téléchargé le dataset qui était publié par ***ALI TOOSI***[46].

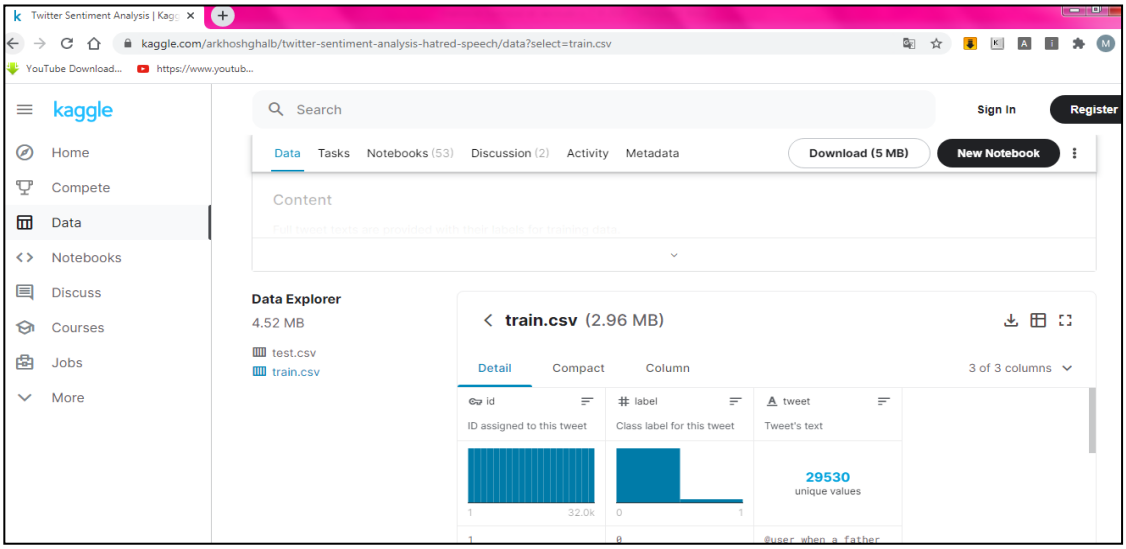
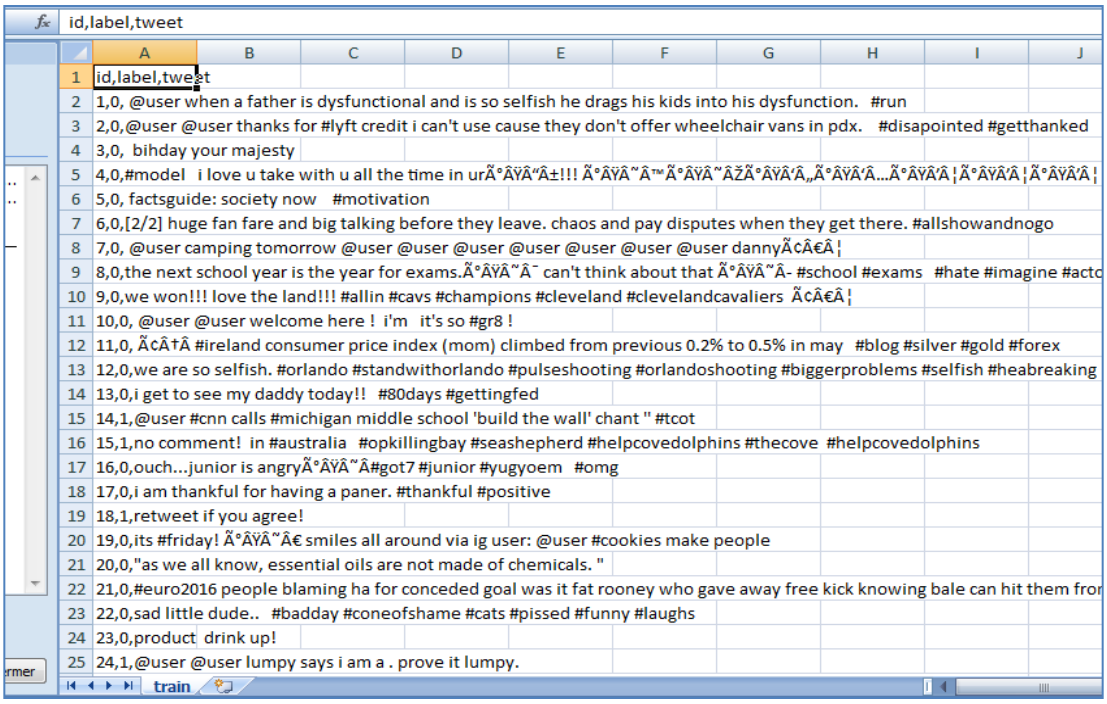


Figure 20: Capture d'écran du dataset téléchargé via le web.

Le data set est organisé sous forme de deux fichier .CSV (test.csv et train.csv) :

[26]https://www.kaggle.com/arkhoshghalb/twitter-sentiment-analysis-hatred-speech/data?select=train.csv

- Train.csv : un ensemble qui contient des tweets déjà jugés (abusifs ou non- abusifs) pour entrainer l'algorithme, il a trois colonnes « id », « label », « tweet » :
    - « id » : contient les id de tweet.
    - « label » : contient l'étiquette des tweets déjà analysés (0ou1).
    - « tweet » : contient le tweet en question.



	A	B	C	D	E	F	G	H	I	J
1	id,label,tweet									
2	1,0, @user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run									
3	2,0,@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked									
4	3,0, bihday your majesty									
5	4,0,#model i love u take with u all the time in urÃ°ÂŸÂ“Â±!!! Ã°ÂŸÂ™Ã°ÂŸÂ˜ÂŽÃ°ÂŸÂ‘Â„Ã°ÂŸÂ‘Â…Ã°ÂŸÂ‘Â¦Ã°ÂŸÂ‘Â¦Ã°ÂŸÂ‘Â¦									
6	5,0, factsguide: society now #motivation									
7	6,0,[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo									
8	7,0, @user camping tomorrow @user @user @user @user @user @user @user dannyÃ¢Â€Â¦									
9	8,0,the next school year is the year for exams.Ã°ÂŸÂ˜Â¯ can't think about that Ã°ÂŸÂ˜Â- #school #exams #hate #imagine #actc									
10	9,0,we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers Ã¢Â€Â¦									
11	10,0, @user @user welcome here ! i'm it's so #gr8 !									
12	11,0, Ã¢Â†Â #ireland consumer price index (mom) climbed from previous 0.2% to 0.5% in may #blog #silver #gold #forex									
13	12,0,we are so selfish. #orlando #standwithorlando #pulseshooting #orlandoshooting #biggerproblems #selfish #heabreaking									
14	13,0,i get to see my daddy today!! #80days #gettingfed									
15	14,1,@user #cnn calls #michigan middle school 'build the wall' chant " #tcot									
16	15,1,no comment! in #australia #opkillingbay #seashepherd #helpcovedolphins #thecove #helpcovedolphins									
17	16,0,ouch...junior is angryÃ°ÂŸÂ˜Â#got7 #junior #yugyoem #omg									
18	17,0,i am thankful for having a paner. #thankful #positive									
19	18,1,retweet if you agree!									
20	19,0,its #friday! Ã°ÂŸÂ˜Â€ smiles all around via ig user: @user #cookies make people									
21	20,0,"as we all know, essential oils are not made of chemicals. "									
22	21,0,#euro2016 people blaming ha for conceded goal was it fat rooney who gave away free kick knowing bale can hit them fror									
23	22,0,sad little dude.. #badday #coneofshame #cats #pissed #funny #laughs									
24	23,0,product drink up!									
25	24,1,@user @user lumpy says i am a . prove it lumpy.									

Figure 21 : Capture d'écran de dataset "Train.cvs" collectée

- Test.csv :un ensemble qui contient des tweets non analysés pour pouvoir tester les différents algorithmes. Il a 2 colonnes « id », « tweet » :
    - « id » : contient l'id du tweet.
    - « tweet » : contient le tweet en question.

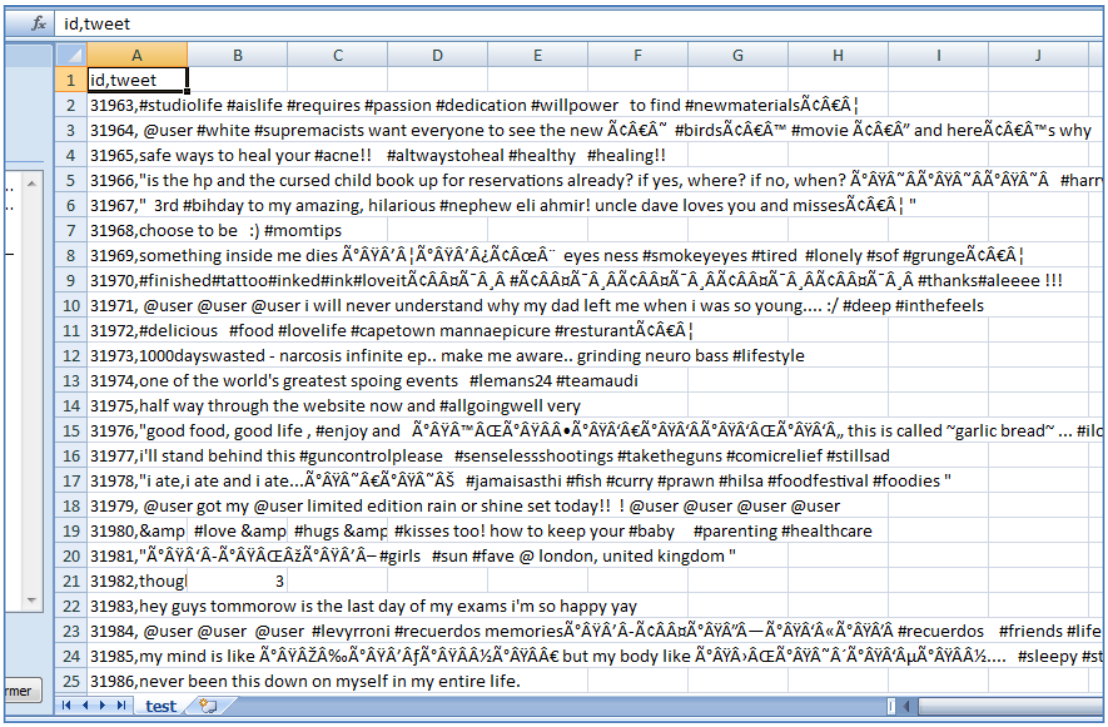


Figure 22 : Capture d'écran de dataset "test.cvs" collectée.

Chaque Tweet dans le dataset possède des attributs tels que :

- Le contenu (du texte) ;
- Les hashtags s'ils existent ;
- La date de la publication ;
- Les propriétés textuelles des mots et hashtags ;
- L'URL du tweet.

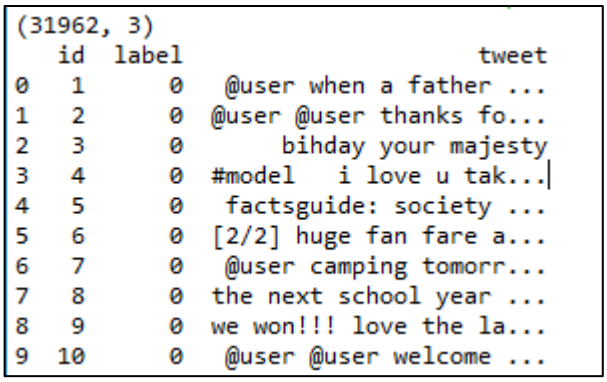
5.2 Phase 2 : Récupération des tweets

Afin de pouvoir commencer la manipulation des données présentes dans le dataset, nous devons tout d'abord procéder par la lecture de notre dataset grâce à l'instruction pd.read_csv :

```
25
26 train  = pd.read_csv('train.csv')
27 test = pd.read_csv('test.csv')
28
```

Figure 23: Lecture des ensembles 'Train' et 'Test'.

La figure 24 montre un fragment des dix premières linges de l'ensemble d'apprentissage :

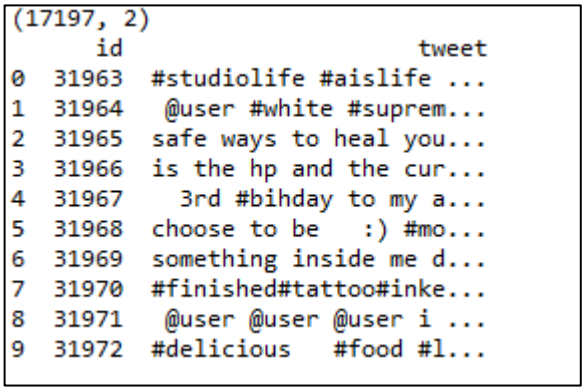


```
(31962, 3)
   id  label                      tweet
0   1      0   @user when a father ...
1   2      0  @user @user thanks fo...
2   3      0       bihday your majesty
3   4      0  #model   i love u tak...
4   5      0   factsguide: society ...
5   6      0  [2/2] huge fan fare a...
6   7      0   @user camping tomorr...
7   8      0  the next school year ...
8   9      0  we won!!! love the la...
9  10      0   @user @user welcome ...
```

Figure 24: Un fragment de la collection d'apprentissage.

Le fichier global contient en tout 31962 linges et 3 colonnes.

La figure 25 montre un fragment des dix premières linges de l'ensemble de test :



```
(17197, 2)
      id                      tweet
0  31963  #studiolife #aislife ...
1  31964   @user #white #suprem...
2  31965  safe ways to heal you...
3  31966  is the hp and the cur...
4  31967    3rd #bihday to my a...
5  31968  choose to be   :) #mo...
6  31969  something inside me d...
7  31970  #finished#tattoo#inke...
8  31971   @user @user @user i ...
9  31972  #delicious   #food #l...
```

Figure 25: Affichage du programme - dix lignes de test.

Le fichier global contient en tout 17197 linges et 2 colonnes.

Détail du data set :

a) Train :L'ensemble d'apprentissage contient 31 962 tweets.Dans l'ensemble de données d'apprentissage, nous avons 2242 (~ 7%) tweets étiquetés comme agressifs et 29720 (~ 93%) tweets étiquetés comme non agressifs. Il s'agit donc d'un défi de classification déséquilibré.

b) Test : L'ensemble de test contient*17 197* tweets.

5.3 Phase 3 : Prétraitement

Le prétraitement (preprocessing en Anglais) est une étape très importante qui consiste en plusieurs techniques visant à traiter les messages pour les structurer et faciliter leurs utilisations.

5.3.1 Les phases de prétraitement proposés

Après avoir lu beaucoup d'articles et des rapports concernant les problèmes de la détection de langage agressif dans les réseaux sociaux, nous sommes arrivés à la conclusion que les étapes de prétraitement illustrés dans la figure 26 ont le plus d'impact et d'influence pour l'obtention d'un meilleur résultat et nous les avons de ce fait retenu pour nos tests.

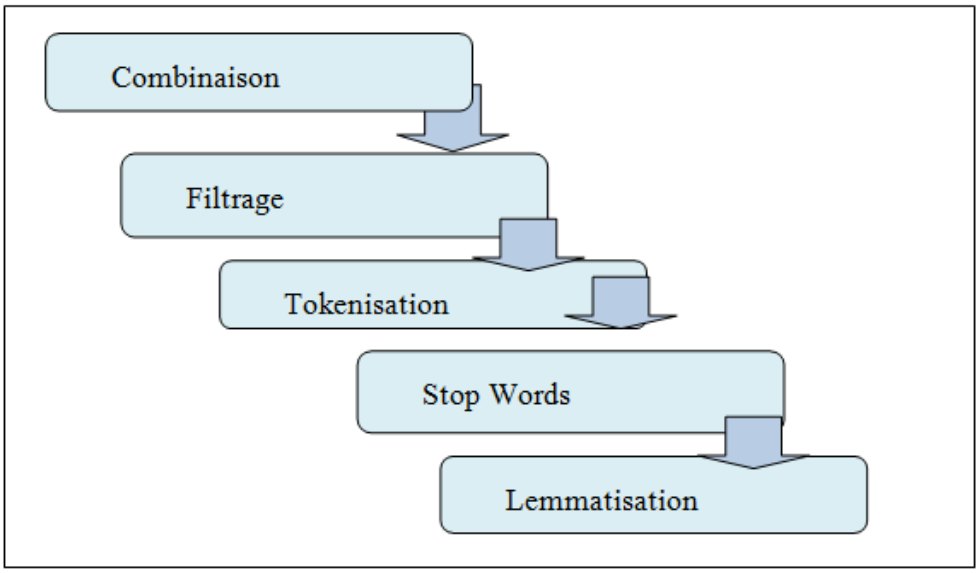


Figure 26 : Les phases de prétraitement

Dans ce qui suit,nous allons décrire le nettoyage (prétraitement)des deux ensembles (Train et Test). Pour cela nous allons dans un premier temps, les combiner puis nous appliquons les déférentes procédures de nettoyage présentées dans la figure 27.

a) Combinaison

Avant de commencer le prétraitement,nous allons combiner l'ensemble du dataset (les deux fichiers .CSV ensemble de train et ensemble de test) pour nous faciliter leur prétraitement.Une fois le prétraitement effectué,nousallons les diviser de nouveaux pour pouvoir procéder au training et tests.

```
#prétraitement (nettoyage du texte brut)
combi = train.append(test, ignore_index=True)
```

Figure 27: Combinaison des deux ensembles (train et test).

b) Filtrage

L'opération de filtrage a pour but la suppression des métadonnées contenues dans les tweets.

- **Suppression des noms d'utilisateurs :**

```
combi['tidy_tweet'] = np.vectorize(remove_pattern)(combi['tweet'], "@[\w]*")
```

Figure 28: Suppression des mots précédés par '@'.

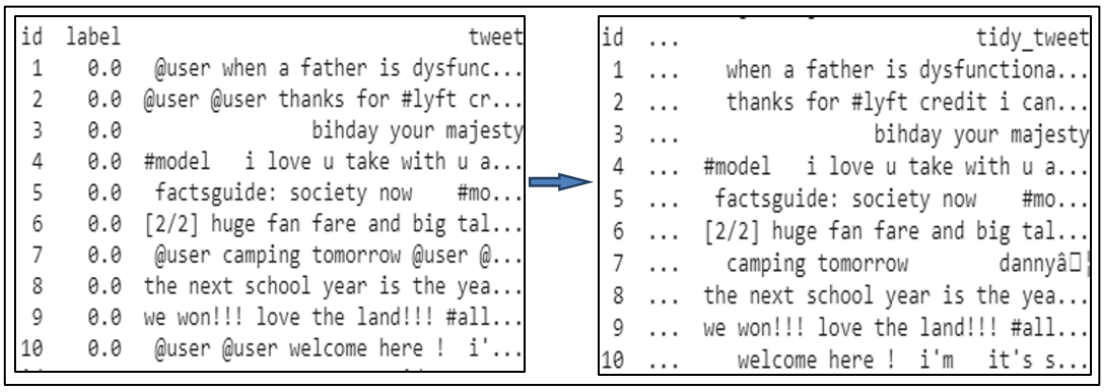


Figure 29: Affichage après la suppression des noms d'utilisateurs

- **Suppression des ponctuations, nombres et caractères spéciaux :**

```
combi['tidy_tweet'] = combi['tidy_tweet'].str.replace("[^a-zA-Z#]", " ")

```

Figure 30 : Remplacementdes caractères spéciaux par un ''vide''.

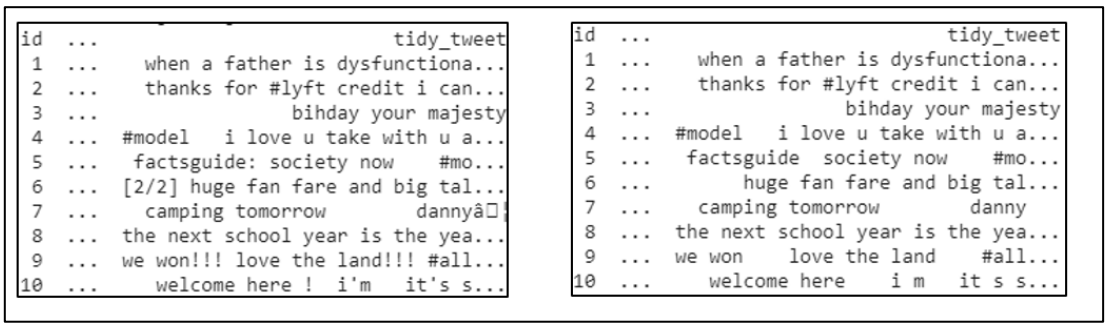
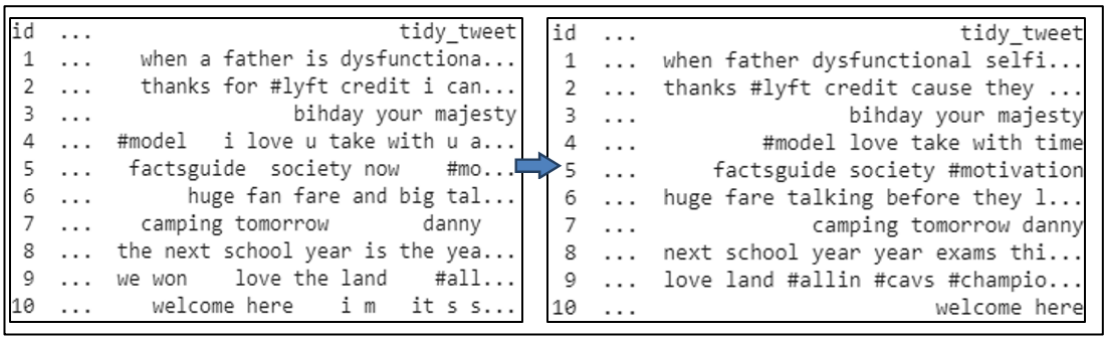


Figure 31 : Affichage après la suppression des caractères spéciaux.

- **Suppression des mots courts (taille < 4 lettres) :**

```
71 combi['tidy_tweet'] = combi['tidy_tweet'].apply(lambda x: ' '.join([w for w in x.split()
72 if len(w)>3]))
```

Figure 32 : Suppression des mots<4.



id	...	tidy_tweet
1	...	when a father is dysfunctiona...
2	...	thanks for #lyft credit i can...
3	...	bihday your majesty
4	...	#model i love u take with u a...
5	...	factsguide society now #mo...
6	...	huge fan fare and big tal...
7	...	camping tomorrow danny
8	...	the next school year is the yea...
9	...	we won love the land #all...
10	...	welcome here i m it s s...

id	...	tidy_tweet
1	...	when father dysfunctional selfi...
2	...	thanks #lyft credit cause they ...
3	...	bihday your majesty
4	...	#model love take with time
5	...	factsguide society #motivation
6	...	huge fare talking before they l...
7	...	camping tomorrow danny
8	...	next school year year exams thi...
9	...	love land #allin #cavs #champio...
10	...	welcome here

Figure 33: Affichage après la suppression des mots<4.

c) Tokenizationdes tweets :

Dans cette partie, l'opération de Tokenization [47] est l'acte de décomposer une séquence de chaînes en morceaux tels que des mots, des mots-clés, des phrases, des symboles et d'autres éléments appelés jetons (tokens).

Extractiondes mots du texte un par un, et leur classement dans une liste :

```
80 tokenized_tweet = combi['tidy_tweet'].apply(lambda x: x.split()) # tokenizing
```

Figure 34 : Extraction des mots.

d) Stop words :

Dans cette étape, nous effectuons une suppression des mots quiaugmente considérablement et inutilement le nombre de mots dans le vocabulaire. Ces mots peuvent inclure :

- Les conjonctions de coordination (for, and, nor, but,or ,yet ,so).
- Les déterminants (a/an, the, this, that, these, those).
- Les prépositions (at, in, to)

Nous avons appliqué dans cetteétapel'algorithmestop words (ou 'Porter stemmer')qui est un processus permettant de supprimer les terminaisons morphologiques les plus courantes des mots en Anglais. Son utilisation principale fait partie d'un processus de normalisation des

termes généralement effectué lors de la configuration de systèmes de recherche d'informations[48].

```
from nltk.stem import PorterStemmer
stemmer = PorterStemmer()
tokenized_tweet = tokenized_tweet.apply(lambda x: [stemmer.stem(i) for i in x])
```

Figure 35 : Exemple du porter Stemmer

- **Reformuler les tweets après le nettoyage**

```
for i in range(len(tokenized_tweet)):
    tokenized_tweet[i] = ' '.join(tokenized_tweet[i])
combi['tidy_tweet'] = tokenized_tweet
```

Figure 36 : Reformulation des tweets**.**

id	...	tidy_tweet
1	...	when father dysfunct selfish drag kid into dysfunct #run
2	...	thank #lyft credit caus they offer wheelchair van #disap...
3	...	bihday your majesti
4	...	#model love take with time
5	...	factsguid societi #motiv
6	...	huge fare talk befor they leav chao disput when they the...
7	...	camp tomorrow danni
8	...	next school year year exam think about that #school #exa...
9	...	love land #allin #cav #champion #cleveland #clevelandcavali
10	...	welcom here

Figure 37 : Affichage après le nettoyage.

- **Affichage du temps de prétraitement (secondes):**

```
le temp de pretraitement de dataset est: 9.20022702217102
```

Figure 38 : Affichage du temps de prétraitement en secondes.

5.4 Phase 4 : transformation des textes vers des vecteurs numériques

En général, les algorithmes de classification ne savent pas manipuler les données brutes (non structurées). C'est pourquoi, après la phase de prétraitement, il faut représenter les données sous forme de matrice. Cette représentation est nommée ***plongement lexical des mots*** (en anglais word embedding) qu'est une manière de représenter des mots comme des vecteurs, typiquement dans un espace de quelques centaines de dimensions. Le vecteur représentant chaque mot est appris via un algorithme itératif, à partir d'une grande quantité de

texte. Cette technique a été massivement exploitée ces dernières années, notamment pour l'apprentissage d'une représentation de mots[49].

Il existe plusieurs modèles pour la représentation des corpus, Au cours de notre expérimentation, nous avons utilisé quatre Modèle, le premier modèle **Sac de mots**, le deuxième modèle **TF-IDF**, le troisième nommé **word2vecteur** et le dernier modèle appliqué est **doc2vecteur**. Nous allons décrire et montrer l'affichage de chaque modèle par la suite.

5.4.1 Modèle de sac de mots

Généralement, dans les processus de traitement statistique des textes, le modèle basé sur la présence de mots d'un certain vocabulaire communément appelé modèle de sac de mots (BOW bag of Words) est le plus utilisé.

On peut le considérer comme étant une suite de caractères appartenantà un dictionnaire, et le message électronique représente l'expérience de tirer avec possibilité de replacement des mots dans un sac. L'ensemble des messages du corpus sera représenté par un vecteur de la même taille que le dictionnaire, dont la composante k indique le nombre d'occurrences du $k^{ème}$mot du dictionnaire dans le document. L'idée est de transformer les messages en vecteurs dont chaquecomposantereprésente un mot. Cette représentation des messages exclut toute analyse grammaticale et toute notion de distance entre les termes, d'où l'appellation « sac de mots »[50].

Le résultat obtenu après avoir appliqué ce modèle est montré dans la figure qui suit :

```
(0, 738)      1
(0, 463)      1
(0, 302)      1
(1, 600)      1
(1, 131)      1
(1, 879)      1
(2, 79)       1
(3, 888)      1
(3, 518)      1
(3, 557)      1
(4, 566)      1
(5, 485)      1
(5, 69)       1
(5, 858)      1
(5, 423)      1
```

Figure 39:Affichage des vecteurs BOW.

5.4.2 Le model TF-IDF

La formule TF*IDF permet de déterminer dans quelles proportions certains mots d'un document texte, d'un corps de document ou d'un site web peuvent être évalués par rapport au reste du texte.

Avec tf-idf, au lieu de représenter un terme dans un document par sa fréquence brute (nombre d'occurrences) ou sa fréquence relative (nombre de termes divisé par la longueur du document), chaque terme est pondéré en divisant la fréquence des termes par le nombre de documents dans le corpus contenant le mot[51].

Le résultat obtenu après avoir appliqué ce modèle est montré dans la figure qui suit :

```
(0, 302)      0.46871135687055143
(0, 463)      0.5896490179849546
(0, 738)      0.6577413621857342
(1, 879)      0.3938403468675415
(1, 131)      0.6145629375807904
(1, 600)      0.6835218920644196
(2, 79)       1.0
(3, 557)      0.7040384885805177
(3, 518)      0.44016705593507854
(3, 888)      0.5572995329862621
(4, 566)      1.0
(5, 423)      0.586120951905663
(5, 858)      0.4743403266916206
(5, 69)       0.4637175931713698
(5, 485)      0.4652198168550412
```

Figure 40 : Affichage des vecteursTF-IDF.

5.4.3 Le model word2vecteur

Word2vec est un concept bien connu, utilisé pour générer des vecteurs à partir des mots. C'est une technique de traitement du langage naturel. L'algorithme word2vec utilise un modèle de réseau neuronal pour apprendre les associations de mots à partir d'un grand corpus de texte. Une fois formé, un tel modèle peut détecter des mots synonymes ou suggérer des mots supplémentaires pour une phrase partielle. Comme son nom l'indique, word2vec représente chaque mot distinct avec une liste particulière de nombres appelés vecteur. Les vecteurs sont choisis avec soin de telle sorte qu'une simple fonction mathématique (la similitude cosinus entre les

vecteurs) indique le niveau de similitude sémantique entre les mots représentés par ces vecteurs[52].

Le résultat obtenu après avoir appliqué ce modèle est montré dans la figure qui suit :

	0	1	2	...	197	198	199
0	-0.023232	-0.012501	-0.164656	...	0.050263	-0.143882	-0.094754
1	-0.002420	-0.140419	0.031167	...	-0.148072	-0.071873	0.113702
2	-0.225845	-0.165452	-0.133372	...	-0.116459	-0.208738	-0.407700
3	-0.103515	-0.298436	-0.320728	...	-0.118708	0.244061	-0.012515
4	-0.140276	-0.328294	-0.195215	...	0.227797	-0.291108	-0.148256
5	0.071023	-0.207297	-0.142079	...	-0.025057	-0.239282	-0.199337
6	-0.309155	-0.405162	-0.202114	...	-0.025929	0.134261	-0.336942
7	0.138757	0.060875	-0.322381	...	0.067695	-0.057567	0.045136
8	-0.355881	0.059108	-0.184552	...	-0.009708	0.287072	-0.253826
9	-0.533821	-0.603070	-0.570437	...	-0.243162	-0.036268	-0.125669
10	-0.339078	-0.032360	0.004775	...	-0.369855	0.034818	-0.646408
11	0.092885	-0.022576	0.040500	...	-0.060665	-0.196367	-0.388858
12	-0.325369	-0.428833	-0.518056	...	-0.253840	0.163572	-0.329119
13	0.152231	-0.022508	-0.348707	...	0.160073	-0.130128	-0.332924

Figure 41: Affichage des vecteurs word2vecteur.

5.4.4 Le model doc2vecteur

Le but de doc2vec est de créer une représentation numérique d'un document, quelle que soit sa longueur. Il peut être utilisé de la manière suivante: pour la formation, un ensemble de documents est requis. Un vecteur de mot W est généré pour chaque mot, et un vecteur de document D est généré pour chaque document. Le modèle entraîne également des poids pour une couche cachée softmax. Au stade de l'inférence, un nouveau document peut être présenté, et tous les poids sont fixés pour calculer le vecteur de document[52].

Le résultat obtenu après avoir appliqué ce modèle est montré dans la figure qui suit :

	0	1	2	...	197	198	199
0	0.069738	-0.005422	0.276049	...	-0.237911	0.330360	-0.061608
1	0.047542	-0.207964	-0.018937	...	0.132484	0.133047	0.274900
2	0.026137	0.010740	0.076969	...	-0.039241	0.194507	-0.107435
3	-0.014307	0.020198	-0.034525	...	-0.004986	0.149467	-0.115106
4	-0.037708	0.091862	0.121662	...	0.034578	0.138780	-0.118242
5	-0.205321	0.240782	-0.145337	...	-0.178670	-0.253617	-0.445587
6	-0.098654	0.033835	-0.115626	...	-0.108361	0.143807	-0.077760
7	-0.028123	-0.100906	-0.222663	...	0.036298	0.330650	-0.699158
8	-0.177077	0.272121	0.003470	...	-0.117470	0.083462	-0.301241
9	0.076336	0.053299	-0.004665	...	-0.083085	0.130630	-0.193329
10	-0.141481	-0.200569	-0.096905	...	-0.095681	-0.269015	0.085407
11	-0.222107	0.009066	0.011334	...	0.120066	-0.042381	-0.139934
12	0.101023	-0.140993	-0.120539	...	-0.184138	0.080323	-0.153676
13	0.009420	-0.103806	-0.092960	...	-0.309221	-0.077474	-0.060736

Figure 42 : Affichage de vecteur doc2vecteur.

5.5 Phase 5 : Algorithme de classification

Après avoir transformé les tweets en des vecteurs numériques, les algorithmes de classification peuvent prendre place pour effectuer la tâche de détection. Dans cette phase, nous avons appliqué cinq algorithmes différents qui sont : *Régression Logistique, SVM (Support Vector Machine), Forêt aléatoire, MLP classifier et XGBoost.*

- Dans un premier temps, nous avons divisé les données en ensemble de training et validation, 70% pour l'ensemble de train et 30% pour l'ensemble de validation. La figure 43 suivante montre la division du dataset.

```
# division des données en ensemble de training et validation set
xtrain_bow, xvalid_bow, ytrain, yvalid = train_test_split(train_bow, train['label'],
                                                          random_state=42,
                                                          test_size=0.3)

```

Figure 43 : Division de dataset.

Par la suite, nous avons fait un appel aux cinq modèles utilisés comme suit :

```
lreg = LogisticRegression()
svc = svm.SVC(kernel='linear', C=1, probability=True)
rf = RandomForestClassifier(n_estimators=400, random_state=11)
mlp = MLPClassifier(solver='lbfgs',hidden_layer_sizes=(80,40,40,10),
activation='relu', random_state=1,learning_rate='adaptive', alpha=1e-6)
xgb_model = XGBClassifier(max_depth=6, n_estimators=1000)

```

Figure 44 : Appel aux cinq modèles de classification.

À savoir que :

- Pour *la régression logistique*, qu'est un algorithme de classification Machine Learning utilisé pour prédire la probabilité d'une variable dépendante catégorielle. Cette variable est une variable binaire qui contient des données codées 1 (oui, succès, etc.) ou 0 (non, échec, etc.). En d'autres termes, le modèle de régression logistique prédit P (Y = 1) en fonction de X.

- Pour *SVM(Support Vector Machine)* qui est en général, considérées comme une approche de classification, mais elles peuvent être utilisées dans les deux types de problèmes de classification et de régression. Il peut facilement gérer plusieurs variables continues et catégorielles.il y a beaucoup de paramètres que nous pouvons choisir de définir et de personnaliser ici, mais nous ne définirons que

Paramètre de régularisation C=1 , kernel=linear qui Spécifie le type de noyau à utiliser dans l'algorithme, et probality=true.

- Pour *Forêt aléatoire*: qui est un modèle composé de nombreux arbres de décision. Ce modèle utilise deux concepts clés qui lui donnent le nom aléatoire :
  - Échantillonnage aléatoire des points de données d'entraînement lors de la création d'arbres
  - Sous-ensembles aléatoires d'entités pris en compte lors de la division des nœuds

  Pour les paramètres, nous définirons le nombre d'arbres dans la forêt n_estimators=400 et random_state=11 qui contrôle à la fois le caractère aléatoire du bootstrap des échantillons utilisés lors de la construction des arbres (if bootstrap=True) et l'échantillonnage des entités à prendre en compte lors de la recherche de la meilleure division à chaque nœud.

- Pour *Mlpclassifier* : signifie classificateur perceptron multicouche qui se connecte à un réseau neuronal, il s'appuie sur un réseau neuronal sous-jacent pour effectuer la tâche de classification. Pour les paramètres, nous définirons les hidden_layer_sizes et Le solveur pour l'optimisation du poids. Pour ces paramètres, nous passons dans un tuple constitué du nombre de neurones que nous voulons à chaque couche, où la nième entrée dans le tuple représente le nombre de neurones dans la nième couche du modèle MLP. Il existe de nombreuses façons de choisir ces nombres, mais pour plus de simplicité, nous choisirons 4 couches (80 ,40,40,10) et le solveur 'lbfgs' est un optimiseur de la famille des méthodes quasi-Newton.

- Pour *XGboost*est un algorithme d'apprentissage automatique puissant, en particulier en ce qui concerne la vitesse et la précision. Nous devons considérer différents paramètres et leurs valeurs à spécifier lors de la mise en œuvre d'un modèle XGBoost. Pour les paramètre, nous ne définirons que paramètre de profondeur maximale des estimateurs de régression individuels max_depth=6 , et Le nombre d'étapes de boost à effectuer n_estimators=1000.

- Après avoir fait l'appel, nous avons entrainé les modèles comme montré dans la figure 45.

```
293 lreg.fit(xtrain_bow, ytrain)
294 svc.fit(xtrain_bow, ytrain)
295 rf.fit(xtrain_bow, ytrain)
296 mlp.fit(xtrain_bow, ytrain)
297 xgb_model.fit(xtrain_bow, ytrain)
298
```

Figure 45 : Entrainement des modèles.

- Après l'entrainement des modèles, nous avons fait une prédiction sur l'ensemble de validation comme montré dans la figure 46.

```
302 prediction = lreg.predict_proba(xvalid_bow)
303 prediction = svc.predict_proba(xvalid_bow)
304 prediction = rf.predict_proba(xvalid_bow)
305 prediction = mlp.predict_proba(xvalid_bow)
306 prediction = xgb_model.predict_proba(xvalid_bow)
307
```

Figure 46: Prédiction sur l'ensemble de validation**.**

- Dans la figure 47, montre comment nous avons étiqueté l'ensemble de validation par 0 ou 1 :

  Si la prédiction est supérieure ou égale à 0.3 alors 1 sinon 0

```
313
314 prediction_int = prediction[:,1] >= 0.3
315
```

Figure 47 : Condition d'étiquetage**.**

- Enfin, nous avons calculé les f-scores pour la validation comme il est illustré dans la figure 48.

```
318 wlrbow=f1_score(yvalid, prediction_int)
```

Figure 48: Calculer de f-score.

A titre d'indication, le f-score ou bien la f-mesure est une mesure globale de la qualité des classifications produites. Cette métrique d'évaluation est basée sur les métriques connues de « Rappel » et de « Précision » auxquelles elle alloue le même degré d'importance. Elle se calcule ainsi :

$$F - mesure = 2 \times \frac{(Précision \times Rappel)}{(Précision + Rappel)}$$

Où la *Précision* concerne une mesure d'exactitude, qui varie entre 0 et 1 et est calculée de la manière suivante (entre autres) :

$$Précision = \frac{|TP|}{|TP + FP|}$$

Et le *Rappel concerne une mesure de perfection qui varie elle aussi entre 0 et 1*et qui est calculée ainsi :

$$Rappel = \frac{|TP|}{|TP + FN|}$$

Tels que :

- Les correspondances correctement trouvées sont appelées « True Positives » (TP).
- Les correspondances incorrectes et détectées par le système comme étant correctes sont appelées « False positives » (FP).
- Les correspondances correctes omises par le système sont appelées « False negatives » (FN).

6. Résultats des tests et discussion

Il est maintenant temps de conclure les choses. Revoyons rapidement ce que nous avons fait, d'abord nous avons nettoyé nos données de texte brutes, puis nous avons pris 5 types différents d'algorithmes de classification pour construire des modèles de détection de langage agressif sur la base d'un dataset fragmenté en collection d'apprentissage et de test. La figure ci-dessous, récapitule les scores F1- obtenus pour les différents algorithmes.

```
!algorithme!    BOW   !    TF-IDF   !    W2V    !    D2V    !
logregf-sc [0.5303408146300915, 0.5451327433628319, 0.6216017634092579, 0.372940156114484]
logregtime [0.262986421585083, 0.18440485000610352, 0.5862982273101807, 0.8419134616851807]
_____________________________________________________________________
SVMf-sc    [0.5092936802973977, 0.5105215004574565, 0.6240487062404871, 0.20871559633027525]
SVMtime    [34.18232083320618, 26.464191913604736, 133.19585371017456, 229.4269049167633]
_____________________________________________________________________
ranforf-sc [0.5529225908372828, 0.562152133580705, 0.5058949624866024, 0.056737588652482275]
ranfortime [44.05534338951111, 50.97785782814026, 122.69449782371521, 126.33569169044495]
_____________________________________________________________________
XGboostf-sc [0.5130687318489837, 0.5185891325071497, 0.6583407671721677, 0.3612040133779264]
XGboosttime [12.614444255828857, 21.471285820007324, 372.52190136909485, 402.65181374549866]
_____________________________________________________________________
resneurf-sc [0.0, 0.4843137254901961, 0.6529680365296804, 0.4182509505703422 3]
resneurtime [2.414398431777954, 7.603217601776123, 40.55584168434143, 43.182392835617065]
_____________________________________________________________________
le temp total d'execution est: 1865.2300534248352
```

Figure 49:Résultats de f1 score.

Pour mieux analyser cet affichage, nous avons synthétisé ces informations en addition aux temps d'exécution des algorithmes dans le tableau suivant (en rouge les F-score et en noir les temps d'exécution):

		Modèle de Vecteur			
Algoritmes		*Bag ofwords*	*Tf-idf*	*Word2vec*	*Doc2vec*
	Logistique Regession	***0.530*** ***0.262***	***0.545*** ***0.184***	***0.621*** ***0.586***	***0.372*** ***0.841***
	Support Vector Machine	***0.509*** ***34.182***	***0.510*** ***26.464***	***0.624*** ***133.195***	***0.208*** ***229.426***
Machine learning	Random Forest	***0.552*** ***44.055***	***0.562*** ***50.977***	***0.505*** ***122.694***	***0.056*** ***126.335***
	Xgboost	***0.513*** ***12.614***	***0.518*** ***21.471***	***0.658*** ***372.521***	***0.361*** ***402.651***
Deep Learning	*MLP-Classifier*	***0.0*** ***2.414***	***0.484*** ***7.603***	***0.652*** ***40.555***	***0.418*** ***43.182***

Tableau6: Résultats d'évaluation des cinq algorithmes declassification.

Nous avons trouvé un très bon résultat de (F-mesure de 0.658), où le modèle **XGboost** avec le vecteur **Word2vec**ont surclasser tous les autres classificateurs, mais l'inconvénient est que ça prend beaucoup de temps pour exécuter la tache par contre, nous pouvons remarquer que le classificateur **MLPClassifier** avec le vecteur **Word2vec**, met 9 fois moins de temps avec une infime différence de précision par rapport au premier (F-mesure de 0.652).

Donc selon les résultats cités plus haut, nous jugeons que pour la tâche de détection de langage abusif, il est préférable d'utiliser le classificateur d'apprentissage profond **MLPClassifier** qui donne des résultats incontestables selon le dataset utilisé.

Comme nous l'avons mentionné précédemment, divers travaux anciens et récents ont entamé ce sujet mais notre travail a l'avantage d'avoir tester un maximum d'algorithmes d'apprentissage.

7. Conclusion

Dans ce chapitre, nous avons proposé une approche pour détecter les contenus abusifs sur twitter en Anglais. Nous avons décrit brièvement le processus de réalisation de notre approche, en spécifiant l'environnement de développement, l'implémentation et la présentation des différentes étapes de l'approche.

En conclusion, nous avons achevé l'implémentation et les tests en respectant la conception élaborée et les tests effectués ont montré que l'algorithme MLPClassifier donne des résultats très intéressants pour notre problème.

Conclusion générale et perspectives

1. Conclusion

Les RS permettent de nos jours, à n'importe qui de partager de l'information avec un grand nombre de personnes, cette information pouvant aller du simple texte (documents, annotations, commentaires,…) aux images, vidéos, ou tout autre contenu multimédia. Bien qu'ils contiennent de nombreux avantages, ces derniers ne sont pas sans risque. Chaque utilisateur sur un réseau possède un profil qui lui est propre et qui guide ses choix sur le réseau. De plus il est relié à un certain nombre d'autres personnes selon des motifs personnels. Parmi les inconvénients des RS, on trouve celui du risque que les posts utilisateurs contiennent des mots obscènes et offensants voire dangereux. Les RS ne sont pas complètement contrôlés et de ce fait les solutions visant à détecter afin de prévenir ce danger sont devenues plus que souhaitables.

En effet, les adolescents passent la plupart de leur temps sur les différents sites de réseaux sociaux. Alors qu'ils sont les premières victimes de harcèlement et de contenu abusif ce qui peut affecter négativement leur moral et leur état mental. A cause de cela, les sites de réseaux sociaux sont soumis à une pression croissante pour s'attaquer aux langages abusifs au cours des dernières années.

Notre travail rentre dans le cadre du grand effort actuel pour essayer de lutter contre la propagation du langage abusif sur les réseaux sociaux. En outre, notre système permet de détecter la présence de langage abusif dans des messages échangés dans les réseaux sociaux (en l'occurrence Twitter dans notre cas).

Le travail a débuté par une étude axée sur les généralités des réseaux sociaux, ensuite un deuxième chapitre a été consacré à une brève définition du langage agressif en se basant sur les travaux antérieurs récents dans le domaine de la détection du langage abusif sur les réseaux sociaux. Un troisième chapitre a été consacré à l'apprentissage automatique, ses modèles, ainsi que quelques algorithmes de classifications que nous avons utilisées.

Enfin, un dernier chapitre a été entamé par une proposition d'une méthodologie de détection du langage agressif et suivi de l'implémentation et les résultats qui ont été présentés dans le dernier chapitre.

Les résultats obtenus ont démontré un très bon résultat de F-mesure (0.658), où la combinaison du modèle **XGboost** avec le vecteur **Word2vec** surclasser les autres classificateurs. Paradoxalement, cette combinaison met beaucoup de temps pour exécuter la tâche. En parallèle, nous avons constaté que le classificateur **MLPClassifier** combiné avec le vecteur **Word2vec** met 9 fois moins de temps avec une infime différence de précision par rapport au premier (F-mesure = 0.652). Donc selon les résultats obtenus, nous avons conclus qu'il était préférable d'utiliser le classificateur d'apprentissage profond **MLPClassifier** pour la tâche de détection du langage abusif dans les réseaux sociaux.

2. Perspectives

Les résultats que nous avons obtenus sont plutôt challengeant et encourageant et répondent aux objectifs initiaux de ce projet, il reste toutefois quelques perspectives d'amélioration qui peuvent être résumées dans ce qui suit :

- Améliorer le rappel en appliquant différents algorithmes et élargir l'ensemble de tests pour inclure des tweets d'autre langue voire différentes variétés dialectales d'une même langue.
- Aussi, le travail dans ce mémoire, s'est concentré sur l'identification des tweets obscènes, et nous jugeons qu'il serait intéressant de l'étendre pour couvrir d'autres fléaux comme le *racisme* et le *sexisme*.
- Une autre perspective consisterai à détecter d'autres formes de contenu offensant sur les médias sociaux tels que la vidéo ou l'audio contenant un discours offensant.
- De plus, il existe de nombreux modèles et combinaison de paradigmes d'apprentissage en profondeur que nous aurions aimé tester et explorer comme la combinaison des réseaux de neurones convolutionnels et récurrents, ainsi que des modèles SVM.
- Bien que nous ayons mis en œuvre une solution simple et efficace, il serait intéressant d'optimiser les critères pour déterminer la confiance de la ligne de base et d'améliorer à la fois l'efficacité et la précision du modèle.

Bibliographie

[1] Z. Nisrine, Réseaux Sociaux numériques : essai de catégorisation et cartographie des controverses, Université Rennes 2, Thèse de doctorat 2012.

[2] C. Marie, «Réseaux sociaux : quelles opportunités pour les services ,Le cas de l'assistance en ligne d'Orange, Mémoire de fin d'étude INTD,» Institut national des techniques de la documentation, 2010.

[3] «Bibliothèque municipale de Lyon,» [En ligne]. Available: http://www.pointsdactu.org/article. Php3? Id article=1293. [Accès le 21 1 2020].

[4] e. a. ELLISON Nicole, «Social network sites : definition, history, and scholarship,,» *Journal of Computer- Mediated Communication*, vol. 13, n° %11, pp. 210-230, 2007.

[5] T. Philippe, «Enjeux et perspectives des réseaux sociaux, mémoire de fin d'étude 2006,» Institut Supérieur du Commerce, paris.

[6] THELWALL Mike, ″Social network sites: Users and uses, ″M.Zelkowitz (Ed.) AdvancesIncomputers Elsevier, vol. 76, pp. 19-73, 2009.

[7] Les petits debrouillards. [En ligne]. Available: https://lespetitsdebrouillards-na.org/?s=r%C3%A9seaux+sociaux. [Accès le 20 1 2020].

[8] T. Mike, «Social network sites: Users and uses, M.Zelkowitz (Ed.) Advances Incomputers Elsevier,» p. 73, 2009.

[9] f. cavazza, «Usages numériques et transformation digitale,» [En ligne]. Available: https://fredcavazza.net/. [Accès le 23 1 2020].

[10] «Community management et réseaux sociaux,» [En ligne]. Available: https://www.vu-du-web.com/community-management/reseaux-sociaux/grands-reseaux-sociaux/. [Accès le 24 1 2020].

[11] K. S. S. Bensalem, «Détection des rumeurs dans les réseaux sociaux. (Master en informatique),» université Abderrahmene Mira - Bijaia, 2017.

[12] «Représentation graphique de réseau social de media,» [En ligne]. Available: https://fr.dreamstime.com/illustration-stock-repr%C3%A9sentation-graphique-r%C3%A9seau-social-media-image81266375. [Accès le 2 1 2020].

[13] R. Cazabet, «Détection de communautés dynamiques dans des réseaux temporels. Thèse de doctorat,» Toulouse, France, 2013.

[14] «QDA,» [En ligne]. Available: https://provalisresearch.com/products/qualitative-data-analysis-software/. [Accès le 29 1 2020].

[15] «wordStat» [En ligne]. Available: https://provalisresearch.com/products/content-analysis-software/wordstat-

whats-new/. [Accès le 30 1 2020].

[16] N. JT, Hate Speech. Encyclopedia of the American Constitution, 2003.

[17] «Community Standards,» [En ligne]. Available: https://www.facebook.com/communitystandards/objectionable_conten. [Accès le 27 2 2020].

[18] «Hateful conduct policy,» [En ligne]. Available: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy. [Accès le 29 2 2020].

[19] W. D. M. M. W. I. Davidson T, Automated Hate Speech Detection and the Problem of Offensive Language., ICWSM. 2017.

[20] H.-D. Wehle, «Machine Learning, Deep Learning, and AI: What's the Difference?. Overview about: What's the difference between:Machine Learning, Deep Learning and AI,» 2017. [En ligne].

[21] R. H. H. Rimouche Nour El Houda, Amélioration du produit scalaire via les mesures de similarités sémantiques dans le cadre de la catégorisation des textes, tlemcen, 2015.

[22] T. D. D. W. a. I. W. Zeerak Waseem, «Understanding abuse : A typology of abusive language detection subtasks,» 2017.

[23] F. C. A. D. F. P. M. a. T. M. Del Vigna, «Hate me, hate me not: Hate speech detection on Facebook,» In CEUR Workshop Proceedings., 2017.

[24] Z. M. A. Malmasi, «Detecting hate speech in social media .,» 2018.

[25] M. R. J. S. A. e. G. C. Wiegand, «Induire un lexique de mots abusifs a Feature-Based Approach », dans Actes de la Conférence de 2018 de la section nord-américaine de l'Association for Computational Linguistics: Human Language Technologies.,» 2018.

[26] J. F. P. .. Park, ««Classification en une étape et en deux étapes pour la détection de langage abusif sur Twitter ». dans AICS Conference : s.n.,» 2017.

[27] W. e. Hovy, «Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies.,» ACL, San Diego, California,, 2016.

[28] P. G. S. G. M. V. V. Badjatiya, «Deep Learning for Hate Speech Detection in Tweets,,» In Proceedings of ACM WWW'17 Companion, Perth, Western Australia., 2017.

[29] Y. Z. S. Z. e. H. X. Y. Chen, «Détecter le langage offensant dans les médias sociaux pour protéger la sécurité des adolescents en ligne »,. / IEEE 2012 sur la confidentialité, la sécurité, les risques et la confiance et Conférence internationale ASE / IEEE.,» 2018.

[30] B. F. L. W. J. H. e. C. R. G. Xiang, ««Détecter les tweets offensants via une fonction d'actualité découverte sur un corpus Twitter à grande échelle »,» 2012.

[31] H. M. S. e. D. S. Chen, «Détection abusive de texte à l'aide de réseaux de neurones». dans CEUR Actes de l'atelier : s.n.,» 2017.

[32] B. G. e. U. Sikdar, ««Utilisation de réseaux de neurones convolutifs

pour classer le discours de haine»,Assoc. Comput. Linguiste., Non.,» 2017.

[33] P. G. S. G. M. e. V. V. Badjatiya, «Deep Learning for Hate Speech Detection in Tweets », dans Actes de la 26e Conférence internationale sur le World Wide Web Companion. .,» 2017.

[34] A. K. T. H. Koratana, «Détection Discours toxiques. s.l. : Stanford University Department».

[35] L. M. Y. A. Cornuéjols, Apprentissage Artificiel, Concepts et algorithmes, 2002.

[36] [En ligne]. Available: https ://recherche.orange.com/learning-zoo/. [Accès le 26 06 2020].

[37] N. Marref, Apprentissage Incrémental et machine à vecteurs supports, batena, 2013.

[38] D. Moualek, Deep Learning pour la classification des images , tlemcen, 2017.

[39] «data scientist,» [En ligne]. Available: https://le-datascientist.fr/les-svm-support-vector-machine. [Accès le 02 07 2020].

[40] «readthedoc,» [En ligne]. Available: https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html. [Accès le 04 07 2020].

[41] «RandomForests,» [En ligne]. Available: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm. [Accès le 9 7 2020].

[42] «xgboost,» [En ligne]. Available: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/. [Accès le 12 07 2020].

[43] [En ligne]. Available: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost. [Accès le 12 07 2020].

[44] [En ligne]. Available: https://stackabuse.com/introduction-to-neural-networks-with-scikit-learn/. [Accès le 14 07 2020].

[45] «Medium,» [En ligne]. Available: https://becominghuman.ai/multi-layer-perceptron-mlp-models-on-real-world-banking-data-f6dd3d7e998f. [Accès le 15 07 2020].

[46] r. d. set. [En ligne]. Available: https://www.kaggle.com/arkhoshghalb/twitter-sentiment-analysis-hatred-speech/data?select=train.csv. [Accès le 01 03 2020].

[47] «Tokenization. (s.d.),» [En ligne]. Available: techopedia: https://www.techopedia.com. [Accès le 20 05 2020].

[48] «Dive Into NLTK,Part IV: Stemming and Lemmatization. (s.d.),» [En ligne]. Available: https://textminingonline.com/dive-into-nltk-part-iv-stemming-andlemmatization. [Accès le 22 05 2020].

[49] V. D. Connes, Apprentissage de plongements lexicaux par une approche réseaux, Toulouse, France, TALN 2019, Jul 2019.

[50] [En ligne]. Available: https://machinelearningmastery.com/gentle-introduction-bag-words-model/. [Accès le 01 06 2020].

[51] [En ligne]. Available: https://programminghistorian.org/en/lessons/analyzing-documents-with-tfidf. [Accès le 05 06 2020].

[52] [En ligne]. Available: https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e. [Accès le 05 06 2020].