République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Bordj Bou Arreridj Mohammed El Bachir El Ibrahimi

Faculté des Mathématiques et d'Informatique

Département d'Informatique

Mémoire

En vue de l'obtention du diplôme de Master

Domaine : Mathématiques et Informatique
Filière : Informatique
Option : Technologies de l'information et de la communication (TIC)

Intitulé

Le discours de haine sur le web et les médias sociaux

Présenté par :

- Fayssal Zerargui
- Omar Benzaoui

Devant le jury composé de :

-	Président
-	Examinateur
- Djamila MOHDEB. MCB, Université de BBA	Encadrante

Année Universitaire 2020/2021

Dédicace

Au nom de Dieu Clément et Miséricordieux

Je dédie ce modeste travail de fin de formation à mes très chers parents qui m'ont été un grand soutien moral dans les moments difficiles durant la formation

À tous mes enseignants pour leur contribution à ma formation

Ainsi qu'à mon frère et à toute ma famille

Sans oublier mon binôme et mes amies proches

A tous qui ont contribué de près ou de loin à la réalisation de ce travail, qu'ils trouvent ici la traduction de ma gratitude et ma reconnaissance.

Fayssal Zerargui

Dédicace

Au nom de Dieu Clément et Misércordieux

Je dédie ce modeste travail de fin d'études à mes grands-parents et mes très chers parents, particulièrement à mon père qui est mon idole dans la vie et ma mère qui m'a été un grand soutien moral dans les moments difficiles durant mes études

À tous mes enseignants pour leur contribution à ma formation

Ainsi qu'à ma sœur et mes frères et à toute ma famille

Sans oublier mes collègues étudiants et mes amies

A tous qui ont contribué de près ou de loin à la réalisation de ce travail, qu'ils trouvent ici la traduction de ma gratitude et ma reconnaissance.

Omar Benzaoui

Remerciements

C’est avec un immense plaisir que nous tenons à remercier très sincèrement toutes les personnes qui nous ont aidé et qui ont ainsi contribué à la réalisation de ce mémoire.

Nous tenons à remercier notre encadrante Dr. Mohdeb Djamila d’avoir dirigé ce travail.

Nos vifs remerciements vont également aux membres du jury pour l’intérêt qu’ils ont porté à notre projet de fin d’études en acceptant d’examiner ce travail et de l’enrichir par leurs propositions.

Nous souhaitons exprimer notre gratitude envers nos familles et tous nos amis pour leur soutien et encouragements tout au long de ce travail.

Enfin, nous voudrions également remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

ملخص

على مدى السنوات القليلة الماضية، ازداد عدد الأشخاص الذين يستخدمون وسائل التواصل الاجتماعي عبر الإنترنت مما أدى إلى توسع غير متوقع لمحتوى الكراهية و العنصرية الذي يستهدف الأشخاص على أساس العرق والجنس والدين وما إلى ذلك.

التحليل اليدوي لخطاب الكراهية على وسائل التواصل الاجتماعي غير عملي بسبب الحجم الهائل للبيانات، حيث إنه مكلف ويستغرق وقتًا طويلاً. لهذا السبب، من المهم اكتشاف وإزالة المحتوى الذي يحض على الكراهية عبر الإنترنت عن طريق المعالجة التلقائية لمحتوى المستخدم.

في هذا المشروع، نناقش دراسة حالة عملية للكشف التلقائي عن تعليقات الكراهية التي تستهدف بشكل خاص فئة اللاجئين الأفارقة في الجزائر على المنصة الاجتماعية YouTube باستخدام تقنيات التعلم الآلي والتعلم العميق ومن خلال احترام منهجية تصنيف النص من مجال المعالجة التلقائية للغة الطبيعية.

كلمات مفتاحية: الكراهية ، خطاب الكراهية ، اللاجئون الأفارقة ، التعلم الآلي ، التعلم العميق ، تصنيف النص ، المعالجة الآلية للغة الطبيعية.

Abstract

Over the past few years, the number of people using online social media has exploded leading to an unexpected expansion of hate and racial content that targets people on the basis of their race, gender, religion, etc.

Manual analysis of hate speech on social media is impractical due to the sheer volume of data, as it is both expensive and time consuming. For this reason, it is important to detect and remove online hate speech by automatically processing user content.

In this project, we discuss a practical case study for the automatic detection of hate comments that specifically target the category of African refugees in Algeria on the social platform YouTube using machine learning and deep learning techniques with respect to the methodology of text categorization resulting from the field of automatic natural language processing.

Keywords: hate, hate speech, Africans, machine learning, deep learning, text classification, NLP.

Résumé

Au cours des dernières années, le nombre de personnes utilisant les réseaux sociaux en ligne a explosé menant à une expansion inattendue du contenu haineux et racial qui cible les personnes sur la base de leurs races, sexes, religions, etc.

L'analyse manuelle du discours de haine sur les médias sociaux est peu pratique en raison du volume considérable de données, car cela est à la fois coûteux et laborieux. Pour cette raison, il est important de détecter et de supprimer les discours de haine en ligne en traitant automatiquement le contenu des utilisateurs.

Dans ce projet, nous abordons un cas d'étude pratique pour la détection automatique des commentaires haineux qui visent particulièrement la catégorie des réfugiés africains en Algérie sur la plateforme sociale YouTube à l'aide des techniques d'apprentissage automatique et d'apprentissage profond et en respectant la méthodologie de catégorisation de texte issue du domaine de traitement automatique du langage naturel.

Mots-clés : haine, discours de haine, haine, Africains, apprentissage automatique, apprentissage approfondi, classification de texte, TALN.

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

Introduction générale

1. Contexte

Toute interaction sociale, que ce soit dans les forums en ligne, les sections de commentaires ou les plateformes comme médias sociaux implique souvent un échange d'idées ou de croyances. Malheureusement, nous voyons souvent que les utilisateurs ont recours à la violence verbale pour gagner une dispute ou éclipser l'opinion de quelqu'un. [1]

Les médias sociaux sont le terrain fertile de cette nouvelle menace socio-virtuelle. Un rapide coup d'œil à travers la section commentaires d'une vidéo YouTube raciste démontre à quel point le problème est omniprésent. Bien que la plupart des grandes entreprises de médias sociaux comme Google, Facebook et Twitter aient leurs propres politiques quant à savoir quels types de discours haineux sont autorisés sur leurs sites, leurs politiques de contrôle sont souvent appliquées de façon incohérente et peuvent être difficiles à comprendre pour les utilisateurs. [2]

2. Problématique

Les contenus en ligne abusifs, tels que les discours haineux et le harcèlement, ont reçu une attention considérable de la part des universitaires, des décideurs et des grandes entreprises technologiques. Sans conteste, ce type de comportement anti-social risque de nuire à ceux qui sont ciblés, d'attiser le discours public, d'exacerber les tensions sociales et de menacer l'exclusion de groupes ciblés des espaces publics.

Malgré l'attention accrue portée à ce sujet dans la littérature scientifique, peu de choses sont connues sur la prévalence, les causes, ou les conséquences de différentes formes des discours haineux sur diverses plates-formes. De ce fait, mener plus d'études et de recherches en vue de détection et d'endiguement automatique du contenu abusif peut apporter une assistance considérable au traitement définitif de ce phénomène inquiétant dans le cyberspace.

3. Objectif

Dans le but d'identifier des solutions pour réduire les effets du langage abusif sur le Web, cette thèse explorera la portée et la nature du problème des discours de haine dans les médias sociaux d'aujourd'hui, en exploitant la plateforme YouTube à analyser et à déterminer la prévalence du discours de haine au sujet de la crise des réfugiés et migrants africains en Algérie.

Notre objectif consiste donc à mettre au point un système de détection automatique du langage abusif sur le YouTube en utilisant une approche d'apprentissage automatique basée essentiellement sur le traitement automatique du langage naturel et le '*deep learning'*.

4. Plan de mémoire

La suite de ce mémoire est organisée en quatre chapitres :

Le premier chapitre consistera en la présentation du discours de haine, ses types, ses acteurs et les motivations derrière la diffusion de ce type de discours. Le deuxième portera sur les fondements théoriques et techniques de la tâche de détection de discours de haine. Le troisième, fera l'objet de l'étude conceptuelle et technique du modèle proposé de détection en se basant sur une description, claire et précise, des techniques utilisées de l'apprentissage automatique et l'apprentissage approfondi. Enfin, dans le quatrième et dernier chapitre, nous présenterons les résultats obtenus par notre modèle après avoir mené une analyse exploratoire de données. Puis, nous dresserons une conclusion générale du projet.

Chapitre I
Le discours de haine

Introduction

Les géants des médias sociaux ont réussi à inscrire près de la moitié de la population mondiale à leurs services, un total si important que les taux de croissance commençaient à ralentir naturellement. Désormais, le harcèlement et les abus sur ces plateformes font des ravages. Certains utilisateurs fuient et les cours des actions ont chuté.

L'objet de ce chapitre est d'introduire les concepts à retenir de ce phénomène inquiétant qui se transforme parfois en violence réelle, mettant en péril la sécurité physique et le bien-être psychologique des victimes d'un côté, et détruisant les efforts des réseaux sociaux à la sociabilisation et à l'ouverture culturelle mondiale d'un autre côté. [3]

1. Qu'est-ce que le discours de haine ?

1.1. Définition 01

Il n'existe pas une seule définition internationalement acceptée du discours de haine. Mais en termes généraux, le discours de haine est une communication qui dénigre les gens en raison de leur appartenance à un **groupe particulier**. Cela peut inclure toute forme d'expression, comme des images, des pièces de théâtre et des chansons, ainsi que la parole.

Certaines définitions étendent même le concept de discours de haine aux communications qui favorisent un climat de préjugé et d'intolérance. L'idée ici est que ces types de communications peuvent alimenter plus tard la discrimination, l'hostilité et les attaques violentes. [4]

1.2. Définition 02

Les propos haineux peuvent être définis comme une expression d'hostilité à l'égard des individus ou des groupes sociaux en fonction de leur appartenance à un groupe, qui peut se référer à : leur race, leur origine ethnique, leur nationalité, leur religion, leur handicap, leur sexe ou leur orientation sexuelle.

A ce stade, le Conseil de l'Union Européen (UEC) définit le discours de haine comme toutes les formes d'expression qui propagent, incitent, promouvant ou justifient la haine raciale, la xénophobie, l'antisémitisme ou d'autres formes de haine fondées sur l'intolérance, y compris : intolérance exprimée par un nationalisme agressif et l'ethnocentrisme, la discrimination et l'hostilité à l'égard des minorités. [5]

2. Types de discours de haine sur les médias sociaux

On commence par les catégories principales d'incitation à la haine en ligne :

2.1. Le buzz émotionnel

Manque d'éthique des médias sociaux, manque de clarté sur les effets publics de l'activité en ligne.

2.2. La virulence personnelle

Ou collective déshumanisant l'autre, la propagande, la manipulation d'images ou des faits.

2.3. L'incitation

Organisée, ciblée, dirigée à la violence. Par exemple : utiliser les '*hoax'* (canulars) ou les rumeurs pour déclencher la violence ou l'action armée.

2.4. Le langage abusif

Peut-être de différents types. La littérature se concentre principalement sur :

- **Le racisme** : la conviction que la race est le principal déterminant des caractéristiques et des capacités humaines et que les différences raciales produisent une supériorité inhérente à une race particulière.
- **Le sexisme** : les préjugés ou la discrimination fondés sur le sexe ; en particulier discrimination à l'égard des femmes.
- **Le discours de haine** : est une langue utilisée pour exprimer la haine à l'égard d'un groupe ciblé ou qui vise à déroger, à humilier ou à insulter les membres d'un groupe (Davidson et al., 2017).
- **Le langage offensif (blessant)** : est une sorte d'abus qui provoque quelqu'un à se sentir blessé, en colère ou contrarié. Il est généralement grossier ou insultant et souvent très désagréable.
- **Le harcèlement** : est un type d'abus dont l'intention réelle est de causer des perturbations, de déclencher ou d'exacerber le conflit à des fins d'amusement.
- **Attaque personnelle** : est un type d'abus qui implique généralement d'insulter ou de rabaisser son adversaire pour invalider son argument. [6]

3. Les caractéristiques de discours de haine

3.1. La longévité

La longévité des propos haineux sur Internet est liée à leur faible coût et à leur potentiel de réapparition immédiate. Les discours de haine peuvent rester en ligne pendant longtemps et sous différents formats sur de nombreuses plates-formes.

3.2. L'itinérance

Les discours de haine en ligne sont aussi itinérants. Même si le contenu est supprimé, son auteur peut s'exprimer à nouveau ailleurs, éventuellement sur la même plate-forme mais sous un pseudonyme différent, ou bien sur d'autres cyberespaces.

3.3. L'anonymat

L'anonymat pose également problème lorsqu'on traite des discours de haine en ligne. L'Internet facilite l'expression anonyme et sous pseudonyme, qui peut aussi facilement accélérer les comportements destructifs qu'alimenter le débat public.

3.4. Le caractère transnational

La portée transnationale d'Internet constitue un obstacle supplémentaire à la lutte contre les discours de haine en ligne ; elle soulève en effet des problèmes de coopération trans-juridictionnelle quant aux mécanismes juridiques visant à combattre les discours de haine.

4. Les diffuseurs de discours de haine en-ligne

Le nombre de sites Web consacrés à la création de contenu haineux a augmenté au fil du temps. Les discours de haine sont courants sur les plateformes de médias sociaux telles que Twitter, Facebook, YouTube, Myspace, Tumblr, Whisper et autres. Parmi les acteurs haineux qui ont promu un discours de haine ciblé et dangereux ces dernières années, on distingue :

- **Les suprémacistes blancs** : la suprématie blanche est une idéologie fondée sur un système complexe de croyances sous-entendant la suprématie des valeurs culturelles et des normes des peuple d'origine européenne par rapport aux autres groupes humains.
- **Les racistes** : le racisme est l'attitude d'hostilité systématique à l'égard d'une catégorie déterminée de personnes. Il est inspiré par l'idéologie du racisme qui est fondée sur la croyance qu'il existe une hiérarchie entre les groupes humains.

- **Les sectaires** : le sectarisme est l'attitude intransigeante de partisans intolérants d'une secte religieuse, d'une opinion, d'un parti politique…etc. [larousse]
- **Les trolls** qui pratiquent la cyberintimidation et le harcèlement en ligne pour l'amusement personnel ou pour d'autres intérêts.
- **Les misogynes** qui éprouvent du mépris, voire de la haine pour les femmes.
- **Les islamophobes** qui éprouvent du mépris et de la haine à l'égard de l'Islam et des musulmans.
- **Les xénophobes** qui éprouvent du mépris et de la haine à l'égard des émigrés et des réfugiés.

Les créateurs de discours de haine sont progressivement bannis de Twitter et d'autres grandes plateformes sociales. Beaucoup de ces personnes créent simplement de nouveaux comptes après avoir été suspendues. D'autres se sont tournés vers des sites plus spécialisés où ils peuvent ouvertement exprimer leur haine. [7]

5. Les cibles de discours de haine en-ligne

Les recherches des dernières années suggèrent que le discours de haine a été ciblé de manière disproportionnée contre les journalistes, les politiciens, les artistes, les blogueurs et d'autres personnalités publiques. Une autre étude récente sur le réseau social Twitter [] a rapporté que les utilisateurs avec plus d'abonnés, de retweets et de listes sont plus susceptibles de devenir des cibles de haine.

Le discours de haine en ligne peut être plus visible dans les attaques coordonnées visant une cible bien déterminée (Mariconti et al. 2018). De telles attaques attirent beaucoup l'attention des médias, à la fois en ligne et dans les médias traditionnels, ce qui rend ces cibles stratégiques utiles pour les extrémistes et les trolls qui cherchent à atteindre un public plus large et élever leurs messages. [8]

6. Conséquences de discours de haine

Les individus et les groupes sont confrontés à de graves conséquences hors ligne en raison du discours de haine en ligne. Nous citons ci-après quelques cas réels de violence qui ont été alimenté et repoussé par des discours haineux contre certain groupes et minorités ciblées :

- Selon des recherches, il existe un taux élevé de discrimination en ligne contre les Afro-Américains, et cette exposition est associée à la dépression et à l'anxiété.

- Les discours de haine en ligne ont exacerbé les tensions intergroupes dans divers contextes, entraînant parfois de violents affrontements et sapant la cohésion sociale. Facebook, par exemple, a été réprimandé pour son rôle dans la mobilisation des foules antimusulmanes au Sri Lanka et l'incitation à la violence contre le peuple Rohingya au Myanmar (Vindu Goel et Frenkel 2018).
- L'auteur de l'attaque contre la synagogue de Pittsburgh en 2019 aurait été radicalisé sur Gab (service de réseau social connu pour sa base d'utilisateurs d'extrême droite), et l'auteur de la fusillade dans une mosquée en Nouvelle-Zélande aurait été radicalisé sur YouTube. Dans les deux cas, les auteurs de violence hors ligne citent fréquemment le rôle joué par les communautés en ligne pour les motiver à agir.

Conclusion

Le contenu haineux se disperse rapidement sur les pages, les chaînes ou les communautés manifestement racistes, misogynes ou discriminatoires sur les réseaux sociaux populaires tels que Facebook, Twitter et YouTube. Le discours de haine n'est qu'un des nombreux facteurs qui interagissent pour mobiliser les conflits intergroupes. L'indulgence envers ce phénomène très présent en ligne peut renforcer la subordination des groupes ciblés, les rendant vulnérables aux attaques et aux crimes de violence hors ligne.

Chapitre II

Détection automatique du discours de haine : Etat de l'art

Introduction

On assiste à un intérêt accru à la détection de discours de haine sur le Web ou sur les médias sociaux de la part de la communauté académique en Informatique. L'identification des fondements théoriques de base sur lesquels ces recherches sont fondées est l'objet de ce chapitre.

1. Problématique

La détection automatique du discours de haine est généralement définie en tant qu'un problème de classification binaire, multi-classe ou bien multi-label. La tâche de détection consiste à vérifier si un contenu donné (un article publié sur le Web ou les médias sociaux) contient des termes haineux à l'égard d'une cible bien déterminée.

2. Approches de détection de discours de haine en ligne

Pour détecter les discours haineux en-ligne, des approches variées ont été exploitées. A ce stade, l'approche principale est le « Data Mining », en particulier le NLP (traitement automatique de la langue naturelle).

Le traitement automatique de la langue naturelle offre plusieurs techniques pour résoudre la tâche de détection du langage haineux. L'apprentissage automatique et l'apprentissage profond sont les méthodes les plus investies par la communauté de recherche dans ce domaine.

2.1. Apprentissage automatique

L'apprentissage automatique, également appelé apprentissage machine ou apprentissage artificiel (*en anglais* Machine Learning), est une forme d'intelligence artificielle (IA) qui permet à un système d'apprendre à partir des données et non à l'aide d'une programmation explicite.

Pour la tâche de détection de discours de haine, l'apprentissage automatique est utilisé en entraînant un modèle de classification avec des données déjà classifiée.

Parmi les algorithmes d'apprentissage automatique, on cite :

2.1.1. **Algorithme des k-voisins les plus proches KNN**

L'algorithme des k-voisins les plus proches (« k-nearest neighbors » ou KNN) est une méthode d'apprentissage à base d'instances, le principe de fonctionnement de cet algorithme est de classifier des points cibles (classe méconnue) en fonction de

leurs distances par rapport à des points constituant un échantillon d'apprentissage (c'est-à-dire dont la classe est connue a priori). [9]

2.1.2. Les arbres de décision

Les arbres de décision sont les plus connues des méthodes d'apprentissage automatique, donc pour classer des données dans des catégories, il faut construire un arbre de décision par catégorie. Pour déterminer à quelles catégories appartient une nouvelle instance, on utilise l'arbre de décision de chaque catégorie auquel on soumet l'instance à classer. Chaque arbre répond Oui ou Non (pour prendre une décision). [10]

2.1.3. Machines à support de vecteurs (ou SVM)

Le SVM consiste en une stratégie de minimisation structurelle du risque mais le problème est de trouver une frontière de décision qui sépare l'espace en deux régions, à trouver l'hyperplan qui classifie correctement les données et qui se trouve le plus loin possible de tous les exemples. On dit qu'on veut maximiser la marge qui veut dire la distance du point le plus proche de l'hyperplan. [11]

2.1.4. Réseaux de neurones

Un réseau de neurones (ou Artificial Neural Network en anglais) est un modèle de calcul dont la conception est très schématiquement inspirée du fonctionnement de vrais neurones (humains ou non). Généralement le réseau de neurones repose sur un nombre élevé de processeurs opérant en parallèle et organisés en tiers. Le premier tiers reçoit les entrées d'informations brutes, un peu comme les nerfs optiques de l'être humain lorsqu'il traite des signaux visuels. [12]

2.1.5. Classification naïve bayésienne

La classification naïve bayésienne est un type de classification bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance des hypothèses. Elle met en œuvre un classifieur bayésien naïf, ou classifieur naïf de Bayes, appartenant à la famille des classifieurs linéaires. [13]

2.2. Deep learning

Deep Learning ou l'apprentissage profond est un type d'intelligence artificielle dérivé du *machine Learning* (apprentissage automatique) où la machine est capable d'apprendre par elle-même. Deep Learning s'appuie sur un réseau de neurones artificiels s'inspirant

du cerveau humain, ce réseau est composé de dizaines voire de centaines de « couches » de neurones, chacune recevant et interprétant les informations de la couche précédente. [14]

Quelques algorithmes d'apprentissage profond :

2.2.1. Les réseaux de neurones à convolution (CNN)

Le CNN est un type de réseau de neurones artificiels acycliques (feed-forward), dans lequel le motif de connexion entre les neurones est inspiré par le cortex visuel des animaux. Les neurones de cette région du cerveau sont arrangés de sorte qu'ils correspondent à des régions qui se chevauchent lors du pavage du champ visuel. Leur fonctionnement est inspiré par les processus biologiques. [15]

2.2.2. Les réseaux de neurones récurrents (RNN)

Un réseau de neurones récurrents est un réseau de neurones artificiels présentant des connexions récurrentes. Un réseau de neurones récurrents est constitué d'unités interconnectées interagissant non-linéairement et pour lequel il existe au moins un cycle dans la structure.

2.2.3. Le réseau de neurones Long Short Terme Memory (LSTM)

Est une architecture qui appartient aux réseaux de neurones récurrents (RNN) utilisé dans le domaine de l'apprentissage profond (Deep Learning).

Les réseaux de neurones récurrents sont conçus pour utiliser certains types de processus de mémoire artificielle qui peuvent aider ces programmes d'intelligence artificielle à imiter plus efficacement la pensée humaine.

2.2.4. Le réseau de neurones bidirectionnels (biLSTM)

En « bilSTM » les neurones comportent plusieurs mécanismes internes (une cellule mémoire et une porte d'entrée, une porte de sortie et une porte d'oubli) cela permettre de tenir compte à la fois des dépendances (courtes et longues) dans les séquences de données, donc il s'agit d'un mécanisme de « longue mémoire à court terme ».

3. Caractéristiques textuelles utilisées pour la détection automatique de discours de haine

Appliquer l'apprentissage automatique ou l'apprentissage profond requiert une étape d'extraction des caractéristiques principales à partir de données (*en anglais*, Feature

Extraction). Ces caractéristiques servent comme des prédicteurs efficaces qui peuvent simplifier la tâche de détection sans utiliser la totalité du vocabulaire des textes à traiter.

Pour la détection du langage haineux, nous divisons les caractéristiques habituellement utilisées en deux catégories : les caractéristiques générales utilisées dans le Text Mining, qui sont courants dans les autres domaines d'exploration de texte ; et les caractéristiques spécifiques de détection des discours de haine :

3.1. Caractéristiques générales

Généralement on utilise des stratégies déjà connues en text mining au problème spécifique de la détection automatique de discours de haine. Donc on peut définir les caractéristiques générales comme les caractéristiques couramment utilisées dans l'exploration de texte. On peut citer entre autres :

3.1.1. Les dictionnaires

C'est une stratégie dans le text mining qui utilise un dictionnaire de mot. Cette approche consiste à faire une liste de mots (dictionnaire) qui sont recherchés et comptés dans le texte. Ces fréquences peuvent être utilisés directement comme fonctionnalités ou pour calculer des scores. Dans le cas de la détection de discours de haine, cela a été menée à l'aide de :

- Mots de contenu (tels que les insultes et les jurons, les mots de réaction, les pronoms personnels).
- Nombre de mots grossiers dans le texte.
- Étiqueter les caractéristiques spécifiques qui consistaient à utiliser également des formes fréquemment utilisées d'abus verbal comme des énoncés stéréotypés largement utilisés.
- « Ortony Lexicon » a également été utilisé pour la détection des affects négatifs ; le lexique « Ortony » contient une liste de mots dénotant une connotation négative et peut être utile, car tout le commentaire contient nécessairement du blasphème et peut être tout aussi nuisible. [17]

3.1.2. Distance métrique

Dans les commentaires textuels il est possible que le les mots offensants sont masqués par une faute d'orthographe intentionnelle, souvent une substitution de caractère unique. Des exemples de ces termes sont « @ss », « sh1t », « nagger ». **La distance de Levenshtein**, c'est-à-dire le nombre minimum de modifications

nécessaires pour transformer une chaîne dans une autre, peut être utilisé à cette fin. La métrique de distance peut être utilisée pour compléter les approches basées sur les dictionnaires. [18]

3.1.3. Le sac de mots (Bag-of-words (BOW))

Sac-de-mots (BOW). Un autre modèle similaire aux dictionnaires est le sac de mots. Dans ce cas, un corpus est créé sur la base des mots qui sont contenu dans les données d'apprentissage, au lieu d'un ensemble prédéfini de mots, comme dans les dictionnaires. Après avoir collecté tous les mots, la fréquence de chaque mot est utilisée comme caractéristique pour entrainer un classificateur. L'inconvénient de ce type d'approches est que la séquence de mots est ignorée, ainsi que son contenu syntaxique et sémantique. Par conséquent, il peut conduire à une mauvaise classification si les mots sont utilisés dans des contextes différents. Pour surmonter cette limite Les N-grammes peuvent être adoptés.

3.1.4. Les N-grammes

L'approche des N-grammes la plus courante consiste à combiner mots séquentiels dans des listes de taille N. Dans ce cas, le but est d'énumérer toutes les expressions de taille N et compte toutes les occurrences. Cela permet d'améliorer les performances des classificateurs, car il incorpore dans une certaine mesure le contexte de chaque mot. Au lieu d'utiliser des mots, il est également possible d'utiliser des N-grammes avec des caractères ou des syllabes. [19]

3.2. Caractéristiques spécifiques

Les caractéristiques spécifiques de détection des discours haineux sont complémentaires aux approches couramment utilisées dans l'analyse de text mining. Plusieurs caractéristiques spécifiques sont utilisées pour s'attaquer au problème de détection automatique de la haine. Nous présentons brièvement les suivantes :

3.2.1. Langage de l'altérité (Othering Language)

Consiste à analyser le contraste entre les différentes expressions de groupes en examinant l'utilisation du « Nous » contre « Eux ». Le langage de l'altérité décrit l'autrui comme inférieur, indigne et incompatible. Par exemple, la phrase « Renvoyez-les tous dans leurs pays d'origine » exprime une séparation claire entre le « nous » et le « leurs ».

3.2.2. **Caractéristiques de l'acteur qui publie du contenu haineux**

Consiste à faire un lien entre le contenu disponible du même utilisateur en se concentrant sur les caractéristiques de l'utilisateur comme le sexe et la localisation géographique.

3.2.3. **Déclarations de supériorité**

Dans ce cas, le discours de haine peut également être présent lorsqu'il n'y a que des déclarations défensives ou des déclarations de fierté, plutôt que des attaques dirigées contre un groupe spécifique.

3.2.4. **Intersectionnisme de l'oppression**

L'intersectionnalité est un concept qui souligne le lien entre plusieurs types particuliers de discours de haine (par exemple, l'interdiction du voile peut être analysée comme un comportement islamophobe ou sexiste, puisque ce symbole est utilisé par les musulmans, mais uniquement par les femmes).

Conclusion

La détection du discours de haine repose particulièrement sur les techniques du text mining et du traitement automatique du langage naturel. Nous avons décrit dans ce chapitre les caractéristiques générales et spécifiques qui sont couramment utilisées pour l'identification automatique du langage abusif en ligne.

Chapitre III

Conception d'un modèle de détection du discours de haine

Introduction

Pour pratiquement comprendre le contenu haineux sur les médias sociaux, nous proposons dans ce chapitre l'étude d'un cas réel de discours de haine qui a marqué pour une longue période les discussions et les débats des communautés algériennes sur le réseau social YouTube.

Nous décrivons notre conception du problème et les techniques choisies pour la détection automatique en se basant sur l'approche de Deep Learning.

1. Description du cas d'étude

La question des réfugiés et des migrants africains qui ont quitté leur pays en raison des conditions de vie difficiles dans leur pays est considérée comme un sujet sensible qui provoque beaucoup de réactions de la part de l'opinion publique algérien. Sur les réseaux sociaux, des campagnes virales visant les migrants subsahariens se déclenchent de temps en temps, prenant par exemple l'hashtag « #لا_للافارقه_في_الجزاير » (#Non aux Africains en Algérie) » qui a envahi, en juin 2018, les pages et les publications des algériens en ligne sur les médias sociaux.

Ce projet vise à vérifier la présence du discours de haine dans le contenu publié par les communautés algériennes sur la plateforme sociale YouTube à l'égard des réfugiés et des émigrés africains en Algérie. Pour cet objectif, nous avons conçu une application de détection de discours de haine en exploitant une approche d'apprentissage automatique basée sur les modèles de Deep Learning.

Notre méthodologie de réalisation est illustrée dans la Figure 3.1. En résumé, nous avons suivi les étapes décrites ci-dessous :

1. Collection de données à l'aide de YouTube Scrapping API.
2. Annotation (étiquetage) manuelle des données collectées.
3. Nettoyage, prétraitement et préparation de données collectées pour la tâche de détection du discours haineux.
4. Classification des données à l'aide des modèles d'apprentissage profond.
5. Evaluation des résultats obtenus.

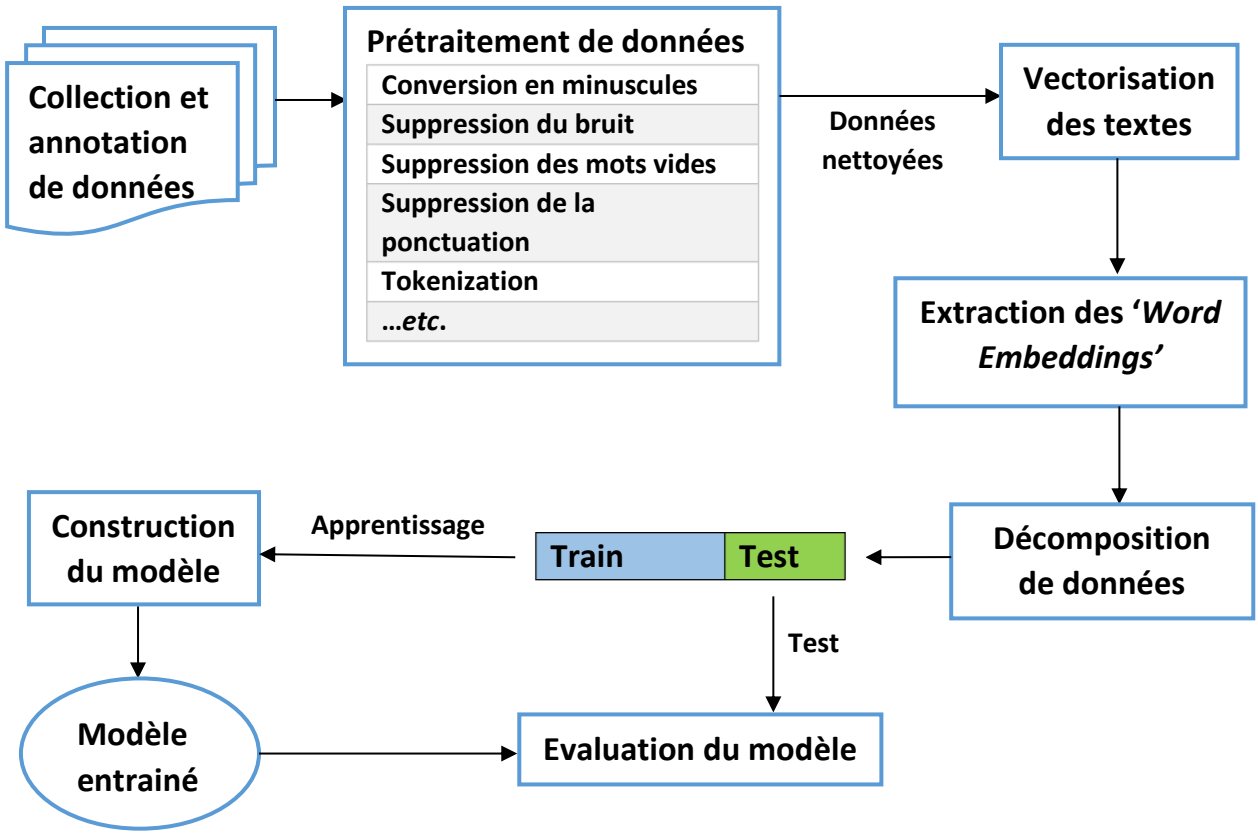


Figure 3.1. Architecture du système proposé

2. Collection de données

YouTube est un site web d'hébergement de vidéos sur lequel les utilisateurs peuvent envoyer, regarder, commenter, évaluer et partager des vidéos en streaming. Par rapport à d'autres plates-formes, YouTube est une plate-forme plus sujette aux discours de haine en raison du manque de mécanismes pour surveiller et bloquer le contenu haineux.

De cette plateforme, nous avons sélectionné 34 vidéos en utilisant YouTube Scrapping API pour collecter un ensemble de 4 730 commentaires qui abordent la question des réfugiés et des migrants africains durant la période entre septembre 2014 et Février 2021. Les titres de ces publications YouTube sont illustrés dans le Tableau 3.1.

Titre	Chaine	Nombre de 'Vues'	Nombre de 'Likes'	Nombre de 'Dislikes'
قناة البلاد داخل "معقل" اللاجئين الأفارقة في الجزائر .. شاهدوا ماذا يفعلون	ELBILADTV	929565	5800	620
هجرة: أفريكاتاون.. آخر خرجات اللاجئين الأفارقة في الجزائر	ENNAHAR tv	102850	442	52
مجتمع: يوميات اللاجئين الأفارقة بالجزائر.. وضعية تناقض التقارير الأجنبية المغلوطة	ENNAHAR tv	86331	578	97

اللاجئين الأفارقة...يفرضون منطقهم في عالم الشغل في الجزائر	ENNAHAR tv	19596	129	11
مناوشات خطيرة بين الافارقة والجزائريين:الدعارة الإفريقية في الجزائر وممارسات غير شرعية بدالي براهيم	Houssem eddine bediaoui	22222	204	23
الجزائر تشرع في نقل مهاجرين أفارقة إلى الجنوب تمهيدا لترحيلهم	DZAIR tv	4100	60	8
متسولون أفارقة يغزون المدن الجزائرية من جديد	ELBILAD TV officiel	7771	152	19
هجرة: أساليب الأفارقة لإجتياز الحدود الجزائرية.. السير على الأقدام ليلا	ENNAHAR tv	31044	208	17
ما وراء الحقيقة .. الأفارقة في الجزائر	Faycel Bouaroua	121015	378	122
هجرة: أفريكاتاون...أخر خرجات اللاجئين الأفارقة في الجزائر حصريا للنهار بتقنية 360	ENNAHAR tv	31148	127	41
خطر اللاجئين الأفارقة على أمن الجزائر و عائلتك -! Abderrahmane Chennafi	Abderrahmane channafi	81744	1100	256
اللاجؤون الأفارقة يساومون الحكومة الجزائرية ويطالبون بالجنسية وإلا..	Ana sohofi ana mawjoud	14736	17	86
مدينة افارقة في باتنة افارقة يؤسسون مدينة بالجزائر في ولاية باتنة 2017	Gouvernement-الحكومة	164814	674	148
تمنراست: تواصل عمليات ترحيل المهاجرين الأفارقة إلى بلدهم الأصلي في النيجر	ENNAHAR tv	8283	100	8
تبون: الجزائر ليست عنصرية و لا ترحيل للأفارقة اللاجئين	Med Midou	9239	47	66
الشيخ النوي: اللاجئون الأفارقة يعانون بسطيف.. الهلال الأحمر الجزائر أصبح أسود	ENNAHAR tv	10047	51	50
الجزائر تبدأ ترحيل المهاجرين الأفارقة	ELBILAD TV	19457	133	15
ممارسات الافارقة في الجزائر آفة تنخر المجتمع	Beur tv	7931	37	9
الجزائر/ المغرب يواصل ابتزاز الجزائر بإغراق الحدود الغربية بمئات المهاجرين الأفارقة	Centurion dz	109830	284	177
المهاجرون الأفارقة في الجزائر.. الخطر القادم من الجنوب	Dzair tv	725	15	2
المغاربي | تدفق المهاجرين الأفارقة على الجزائر	Alaraby tv	4597	22	19
الجزائر.. سياسة ترحيل المهاجرين الأفارقة	France 24 ARABIC	20257	83	38
الجزائر .. أويحي يثير الجدل بتصريحاته حول المهاجرين الأفارقة	Medi1TV	17442	77	22
الجزائر تطرد آلاف المهاجرين الأفارقة إلى النيجر.. توضيحات خبير في شؤون الهجرة والقانون الدولي	Medi1TV	6070	118	15
جزائريون : عودة قوية للمهاجرين الأفارقة الى المدن الجزائرية	ENNAHAR tv	15691	79	11
هكذا يواجه المهاجرون الأفارقة الرافضين لعمليات الترحيل قساوة الطقس	ENNAHAR tv	2929	17	4

من هناك | المهاجرون الأفارقة.. عبء أمني واقتصادي يثقل كاهل الجزائر	سكاي نيوز عربية	1986	33	4
مهاجرون أفارقة يعانون في المدن الجزائرية	France 24 ARABIC	5343	34	17
المهاجرون غير الشرعيين في الجزائر بين مطرقة الاعتقال وسندان الترحيل	France 24 ARABIC	13249	52	60
الأمم المتحدة تدعو الجزائر الى "التوقّف فوراً عن طرد المهاجرين" الأفارقة	Medi1TV	2204	30	25
المغاربي | المهاجرون الأفارقة في الجزائر... نصف حلم وقساوة واقع	Alaraby tv	10516	25	30
الجزائر.. تقارير عن استغلال المهاجرين الأفارقة كعبيد	Medi1TV	1652	38	27
عاجل مشادات دامية بين الجزائريين والافارقة بعدما حاول الافارقة اختطاف فتاة صغيرة بعمر9 في بشار	Houssem eddine bedjaoui	13891	158	8
تقرير| المهاجرون الأفارقة يعيشون أوضاعا متردية في الجزائر	ALghad tv-قناة الغد	2143	14	6
كارثة كارثة افارقة ماذا يصنعون في الجزائر .استعمار افريقي	Sohaib Anis	110307	437	146
العفو الدولة تنتقد حملة على الإنترنت لطرد المهاجرين الأفارقة من الجزائر	France 24 ARABIC	5300	30	47
جنيف: التنديد بقرار الجزائر ترحيل المهاجرين الأفارقة	Medi1TV	3496	52	31
تراجع انتشار الأفارقة في الجزائر	NumediaTV	126025	1100	70

Tableau 3.1. Les vidéos sélectionnées pour la collection de données

3. Annotation de données

Notre travail dans cette étape consiste à étiqueter chaque commentaire par rapport à la position exprimé du commentateur envers le contenu de la vidéo. Nous avons classé les commentaires manuellement, en utilisant un système de classification basée sur les cinq classes suivantes :

- **Incitation (I) :** Les commentaires qui incitent la violence contre les réfugiés. **Exemple** : "كونو جماعات في الأحياء وحازوهم ماتعولوش على الدولة. يديرو حتى االسحور اليهود "
- **Haine (H) :** Les commentaires qui refusent la présence des réfugiés et émigrés africains en utilisant un langage abusif (insultes, humiliation, intimidation, ...etc.). **Exemple :** " حنا بلادنا مغسولة بدم الشهداء الاحرار ماخصناش الزبل في بلدنا"
- **Refus Non-Haineux (RNH) :** Les commentaires qui refusent la présence des réfugiés et émigrés africains en utilisant un langage non haineux. **Exemple :** "حنا ماناش عنصريين بصح هادو كي يشبعو يحوزوني انا و انتم من بلادنا"
- **Sympathique (S) :** Les commentaires avec un contenu empathique pour la question de l'émigration africaine. **Exemple** : " حسبي الله ونعم الوكيل في الدولة الجزائرية لا رحمة لا شفقة ماشي قادرا تأوي لاجاءين الأفارقة لي تضمن لهم أكل وشرب ومبيت "

- **Commentaire (C) :** Les commentaires qui ne montrent pas une opinion spécifique à l'égard des migrants africains. **Exemple :** " مفهمتش الأفارقة ؟ العنوان غالط حنا تاني افارقة "

4. Nettoyage et prétraitement de données

Avant de travailler sur la base de données collectée, il faut la nettoyer et prétraiter, en effectuant les tâches suivantes :

- Supprimer les données bruit (commentaires vides, hors-sujet, URLs, etc.)
- Enlever les espaces supplémentaires.
- Supprimer les mots vides (Stopwords).
- Supprimer les caractères spéciaux.
- Transformer les chiffres en lettres.
- Supprimer la ponctuation.
- Convertir les lettres du texte non-arabe en minuscules.
- Suppression des hashtags, des mentions (@), et des Emojis.

Le texte arabe contient des particularités qui doivent être traitées par les fonctions supplémentaires suivantes :

- Conversion des textes écrits en Arabizi en textes écrits en arabe.
- Suppression de la ponctuation arabe.
- Suppression de diacritiques (*en arabe*, tashkeel).
- Normalisation des lettres longues (*en arabe*, tatweel).
- Normalisation de la ligature.
- Normalisation de la "hamza".

5. Vectorisation du texte avec GloVe

La vectorisation du texte (*en anglais* word embedding) est un ensemble de techniques de traitement du langage naturel où les mots ou les phrases sont représentés par des vecteurs numériques .Alors que les méthodes de vectorisation classiques utilisent des simples techniques telles que le BoW (Bag of Words) et le TF-IDF (Term Frequency-Inverse Document Frequency) pour extraire les représentations numériques des mots, les méthodes récentes utilisent des représentations prêtes à l'utilisation immédiat, parce qu'elles ont été entrainées auparavant sur des grands corpus des textes.

Il existe plusieurs modèles pré-entrainés pour la vectorisation du texte. Les plus populaires sont Word2Vec développé par Google, FastText développé par Facebook et GloVe développé par Stanford.

GloVe (**Glo**bal **Ve**ctors for Word Representation) est un algorithme d'apprentissage non supervisé pour obtenir les représentations pour les mots. L'idée de base de GloVe est de dériver la relation entre les mots de statistique globale. L'un des moyens les plus simples pour représenter ces statistiques globales est de regarder la matrice de cooccurrence. Une matrice de cooccurrence nous indique à quelle fréquence une paire particulière de mots se produit ensemble. Chaque valeur dans une matrice de cooccurrence est un décompte d'une paire de mots apparaissant ensemble. [21]

Par exemple : je suis raciste, je déteste les africains et je déteste les réfugiés.

	je	raciste	déteste	africains	refugiés
je	0.0	1.0	2.0	1.0	1.0
raciste	1.0	0.0	0.0	0.0	0.0
déteste	2.0	0.0	0.0	1.0	1.0
africains	1.0	0.0	1.0	0.0	0.0
refugiés	1.0	0.0	1.0	0.0	0.0

6. Apprentissage profond pour la classification du texte

6.1. Définition

L'apprentissage approfondi ou « Deep learning » est une technique d'apprentissage automatique (machine learning) qui vise à construire automatiquement des connaissances à partir de grandes quantités d'information, il est basé sur des réseaux neuronaux artificiels. Le processus d'apprentissage est qualifié de profond parce que la structure des réseaux neuronaux artificiels se compose de plusieurs couches d'entrée, de sortie et masquées. Chaque couche contient des unités qui transforment les données d'entrée en informations que la couche suivante peut utiliser une tâche prédictive spécifique. Grâce à cette structure, une machine est capable d'apprendre au travers de son propre traitement de données. [20]

Dans notre projet, nous avons appliqué deux modèles pour la classification des textes collectés : les réseaux de neurones à convolution et les réseaux de neurones récurrents LSTM.

6.2. CNN (Convolutional Neural Network)

Le réseau de neurones à convolution est un type de réseau neuronal artificiel utilisé dans la reconnaissance et le traitement de données. Un CNN utilise un système semblable à une perception multicouche qui a été conçu pour des besoins de traitement réduits.

Les couches d'un CNN se composent d'une couche d'entrée, d'une couche de sortie et d'une couche cachée qui comprend plusieurs couches convolutionnelles, des couches de regroupement, des couches entièrement connectées et des couches de normalisation.

La suppression des limitations et l'augmentation de l'efficacité pour le traitement des données aboutissent à un système beaucoup plus efficace, plus simple à former, et spécialisé pour le traitement du langage naturel.

Un réseau de neurone à convolution se distingue par les couches suivantes :

- **La couche de vectorisation** : où les textes à classifier seront vectorisées.
- **La couche de convolution** : son but est de repérer la présence d'un ensemble de caractéristiques dans les données reçues en entrée.
- **La couche de « Pooling »** : l'opération de 'Pooling' consiste à réduire la taille, des caractéristiques et conserver les caractéristiques les plus importants des donnés.
- **La couche « Flatten »** : cette couche sert à convertir la sortie de la partie convolution et *Pooling* en un vecteur de caractéristiques unidimensionnels.
- **La couche entièrement connectée** : le but de cette couche est de prendre le résultat final des trois couches précédentes pour réaliser la tâche de classification demandée.

6.3. LSTM (Long Short-Term Memory)

Un réseau neuronal récurrent LSTM est un type d'architecture RNN dont le but est de donner aux réseaux de neurones récurrents la capacité de maintenir un état sur une longue période de temps. Il possède une mémoire interne appelée cellule. Cette cellule permet de maintenir un état aussi longtemps que nécessaire, elle consiste en une valeur numérique que le réseau peut piloter en fonction des situations. [23]

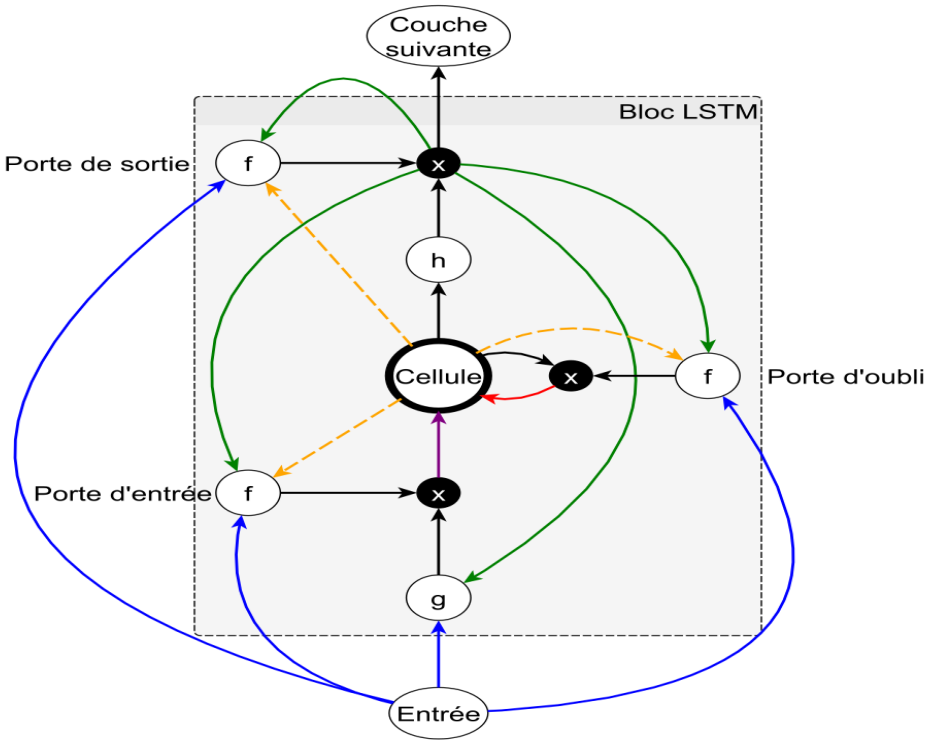


Figure 3.2. Structure d'un LSTM

Comme on peut constater dans le Figure 3.2, un LSTM se compose de trois portes de contrôle :

- **La porte d'entrée** : décide si l'entrée doit modifier le contenu de la cellule
- **La porte d'oubli** : décide s'il faut remettre à 0 le contenu de la cellule.
- **La porte de sortie** : décide si le contenu de la cellule doit influer sur la sortie du neurone.

Conclusion

Les compagnes de discours de haine contre les réfugiés africains en Algérie propagent rapidement et largement en parallèle avec l'augmentation du nombre des migrants illégaux. Vu la sensibilité de cette crise, nous avons proposé une étude de cas suivant l'approche de l'apprentissage automatique pour la classification et la détection de discours de haine dans ce sujet en utilisant les modèles de Deep Learning avec la vectorisation des textes en GloVe. Dans le chapitre suivant les résultats de ce cas d'étude.

Chapitre IV
Implémentation & Résultats

Introduction

Après la phase de conception, la phase la plus critique est la mise en œuvre. Le choix des outils de développement a un impact important sur la performance visée. Cette phase consiste à convertir le modèle conceptuel précédemment développé en code exécutable. Dans ce dernier chapitre, nous décrivons les étapes de réalisation de l'application de détection du discours de haine sur les médias sociaux.

1. Environnement et outils de travail

1.1. Environnement matériel

Le développement de l'environnement matériel est caractérisé par :

N°	Modèle PC	Processeur	Système d'exploitation	RAM
Poste de travail N°01	DELL	Intel(R) Core(TM) i5-6300U CPU @ 2.50ghz	Windows 10 Professional	8GB
Poste de travail N°02	LENOVO	Intel(R) Core(TM) i3-5100U CPU @ 2.30ghz	Windows 8.1	4GB

Tableau 4.1. Caractéristiques de l'environnement matériel

1.2. Langage de programmation

Nous avons choisi Python, version 3.7.4, pour implémenter notre système de détection du discours de haine. Python est un langage de programmation orienté objet de haut niveau qui est simple à comprendre et à écrire. Il est interactif et interprétable car ses phrases sont traduites dans un langage machine que l'ordinateur comprend lors de l'exécution du programme. Python est beaucoup plus facile à apprendre que les autres langages de programmation et permet aux utilisateurs de créer d'excellents programmes.

1.3. Editeur de code

Pour éditer le code de notre système, nous avons utilisé **Visual Studio Code.** C'est un environnement de développement intégré conçu par Microsoft pour Windows, Linux et Mac OS X. Les fonctionnalités incluent la prise en charge du débogage, la mise en évidence de la syntaxe, la complétion de code intelligente, les extraits de code, la refactorisation de code et Git intégré. Dans l'enquête auprès des développeurs Stack Overflow 2021, Visual Studio Code a été classé comme l'outil d'environnement de développement le plus populaire, avec 71,06 % des personnes interrogées déclarant l'utiliser.

Figure 4.1. Editeur de code et bibliothèques Python utilisés

1.4. Librairies et bibliothèques Python

Les principaux packages utilisés pour la réalisation de ce projet sont :

- **Pafy** : Pafy est une bibliothèque Python pour télécharger du contenu YouTube et récupérer des métadonnées telles que le nombre de vues, la durée, l'évaluation, l'auteur, la vignette, les mots-clés. Elle fonctionne avec Python 2.6+ et 3.3+.
- **Pandas** : pandas est une bibliothèque open source sous licence BSD fournissant des structures de données et des outils d'analyse de données hautes performances et faciles à utiliser pour le langage de programmation Python. [24]
- **Numpy**: est le package fondamental pour le calcul scientifique en Python. Il s'agit d'une bibliothèque Python qui fournit un objet tableau multidimensionnel, divers objets dérivés (tels que des tableaux et des matrices masqués) et un assortiment de routines pour des opérations rapides sur des tableaux, notamment mathématiques, logiques, manipulation de forme, tri, sélection, E/S, transformées de Fourier discrètes, algèbre linéaire de base, opérations statistiques de base, simulation aléatoire et bien plus encore. [25]

- **Matplotlib** : est une bibliothèque complète pour créer des visualisations statiques, animées et interactives en Python, il offre une alternative open source viable à MATLAB.
- **Tensorflow** : est une plate-forme open source de bout en bout pour l'apprentissage automatique. Il dispose d'un écosystème complet et flexible d'outils, de bibliothèques et de ressources communautaires qui permet aux chercheurs de pousser l'état de l'art en matière de Machine Learning et aux développeurs de créer et de déployer facilement des applications basées sur le Machine Learning. [26]
- **Keras** : est une bibliothèque de logiciels open source qui fournit une interface Python pour les réseaux de neurones artificiels. Keras agit comme une interface pour la bibliothèque TensorFlow. Conçu pour permettre une expérimentation rapide avec les réseaux de neurones profonds, il se concentre sur la convivialité, la modularité et l'extensibilité. [27]
- **Scikit-learn** : est probablement la bibliothèque la plus utile pour l'apprentissage automatique en Python. La bibliothèque Sklearn contient de nombreux outils efficaces pour l'apprentissage automatique et la modélisation statistique, notamment la classification, la régression, le clustering et la réduction de la dimensionnalité. [28]
- **Pyarabic** : Une bibliothèque de langue arabe spécifique pour Python, fournit des fonctions de base pour manipuler les lettres et le texte arabes, comme détecter les lettres arabes, les groupes et les caractéristiques de lettres arabes, supprimer les signes diacritiques, etc. [29]

2. Implémentation & Résultats

2.1. Informations générales

Une fois que nous avons décidé de l'objectif de notre projet et du type de données que nous devons collecter, l'étape suivante consiste à créer un script avec python pour extraire les données à l'aide de l'ID de la vidéo YouTube que nous avons sélectionnée. Le script utilise les informations d'identification de l'API pour accéder à l'**APIv3** de YouTube. Avec ce script YouTube, nous pouvons extraire les commentaires des utilisateurs, les données sur une chaîne YouTube et les légendes et d'autres informations plus encore. Les données extraites ont été finalement exportées à un DataFrame et sauvegardées sous format (.CSV).

Les caractéristiques de l'ensemble de données final sont décrites ci-après :

L'ensemble de données	DataFrame
Nombre de lignes	4730
Nombre de colonnes	15
Type de données	Objet
Noms des colonnes	(videoid , v.title , v.published , v.description , video_url , v.duration , v.dislikes , v.likes , v.rating , v.category , v.author , author , text , repcount , commentlike)
La taille du DataFrame	8,77 MO

Tableau 4.2. Caractéristiques de l'ensemble de données final

Le Tableau 4.3 montre la totalité des informations de DataFrame résultat :

Attributs	Description
videoid	L'identifiant unique de la vidéo
v.title	Le titre de la vidéo
v.published	La date et heure exacte de publication sur YouTube
v.description	La description de la vidéo sur YouTube
video_url	Le lien d'adresse de la vidéo sur internet
v.duration	La durée de la vidéo
v.dislikes	Le nombre des mentions « Je n'aime pas »
v.likes	Le nombre des mentions « J'aime »
v.rating	Le score d'évaluation de la vidéo
v.category	La catégorie de la vidéo
v.author	L'éditeur qui a mis en ligne sur la vidéo YouTube
author	L'éditeur du commentaire
text	Le commentaire
repcount	Le nombre de réponses au commentaire
commentlike	Le nombre de « J'aime » sur le commentaire

Tableau 4.3. Attributs de l'ensemble de données collecté

2.2. Distribution de classe

Les 4291 commentaires sont distribués sur cinq classes :

- 384 commentaires sont de type **Incitation (I)**
- 959 commentaires sont de type **Hate (H)**
- 776 commentaires sont de type **Refuse-NoHate (RNH)**
- 845 commentaires sont de type **Sympathetic (S)**

- 1327 commentaires sont de type **Comment (C)**

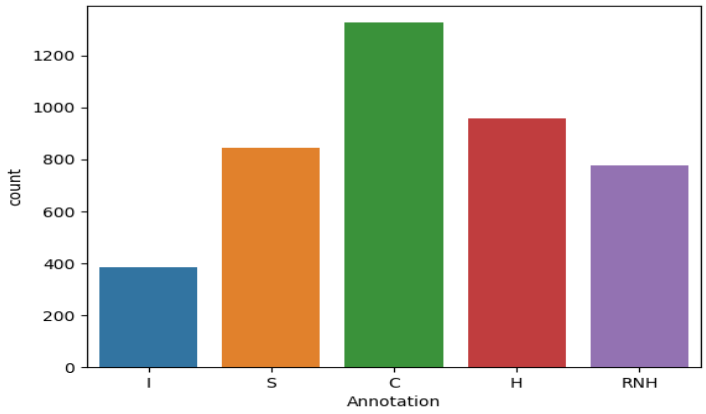


Figure 1.2 Distribution des classes et des types des données

2.3. Fréquences des termes (N-Grammes)

L'objectif ici est d'explorer le vaste ensemble de données non structurées et de découvrir tous les modèles, caractéristiques et points d'intérêt initiaux. On ne s'attend pas à ce qu'il révèle toutes les informations contenues dans l'ensemble de données à cette étape, mais plutôt à aider à créer une image globale des tendances importantes et des principaux points à étudier plus en détail.

La figure 4.3 montre les mots les plus fréquents dans notre corpus. Les principaux mots sont principalement des mots liés à l'Algérie, les africaines etc. Les figures 4.4 et 4.5 montrent les bigrammes et les trigrammes fréquents.

Figure 4.3 Les mots les plus fréquents dans l'ensemble de données

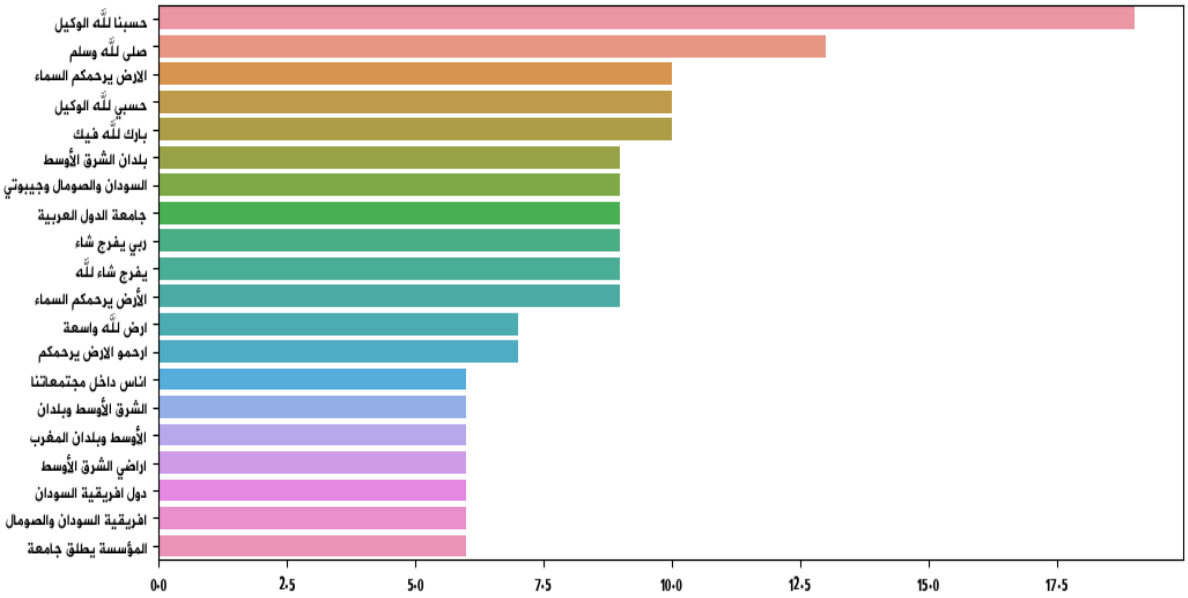


Figure 4.4 Les trigrammes les plus fréquents dans l'ensemble de données

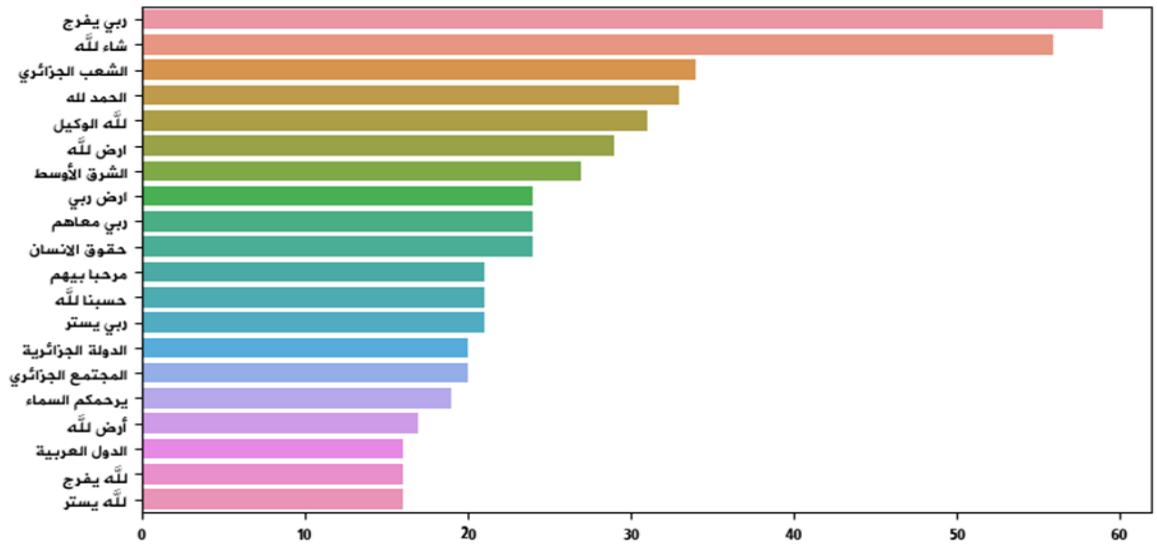


Figure 4.5 Les bigrammes les plus fréquents dans l'ensemble de données

2.4. Style linguistique des textes

2.4.1. Distribution de la longueur des textes

La longueur des échantillons dans l'ensemble de données est très importante, car elle peut affecter la façon dont on représente le texte en tant que fonctionnalité pour les modèles

Machine Learning. Par exemple, les réseaux LSTM sont bien meilleurs que les réseaux RNN sur les longs textes. Nous calculons le nombre de mots dans chaque commentaire et examinons la distribution des longueurs.

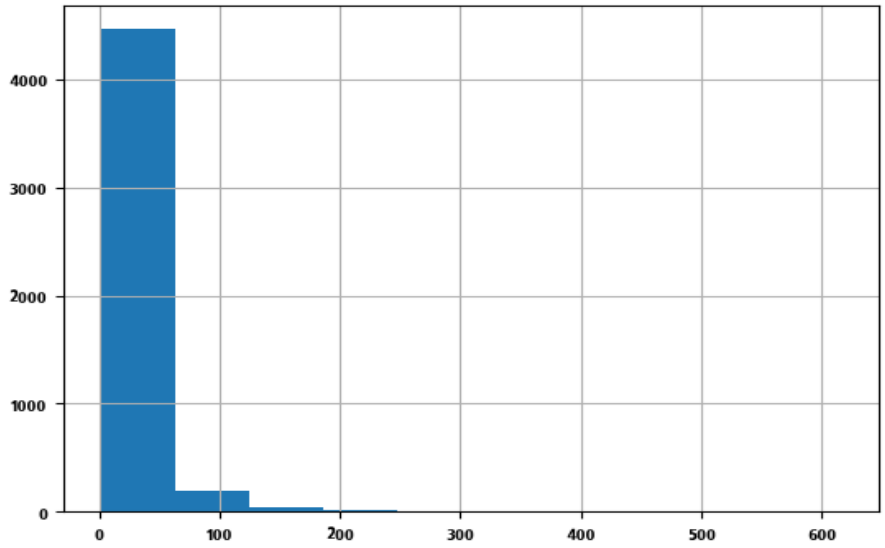


Figure 4.6 La distribution de la longueur des commentaires

2.4.2. Les mots vides les plus fréquents

Comme nous pouvons le voir, les données montrent qu'il existe de nombreux mots vides dans l'ordre « في » > « و » > « من » ...etc.

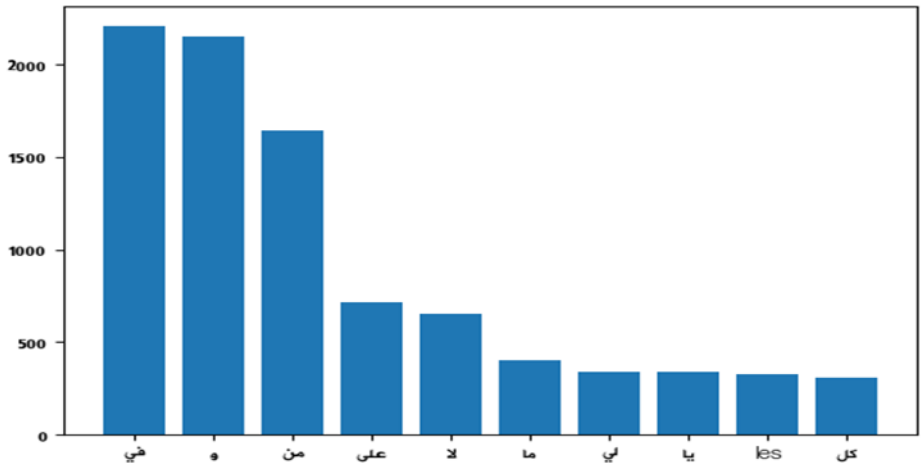


Figure 4.7 Top 10 mots vides

2.5. Perspective temporelle

L'analyse des séries chronologiques implique la compréhension de divers aspects du sujet étudié. On peut voir dans les figures que l'années 2017 et le mois 06 (juin) c'est les plus actives en débats pour les commentateurs algériens sur YouTube au sujet des migrants africains.

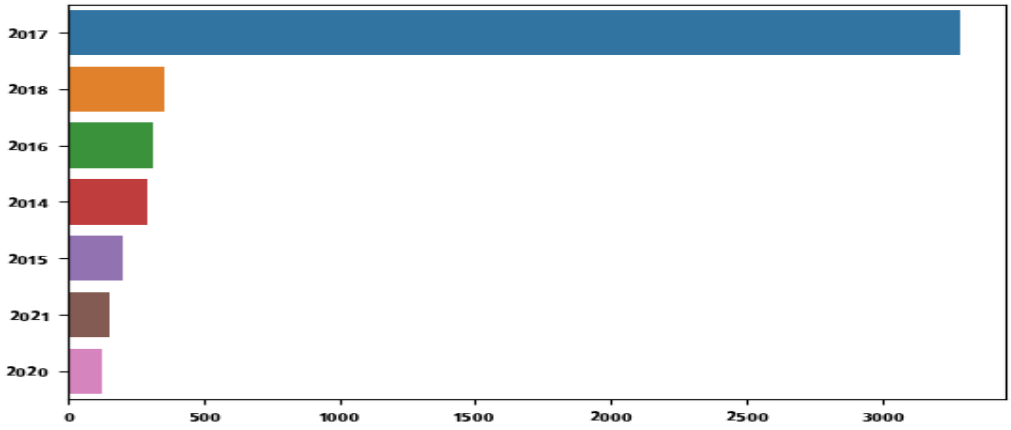


Figure 4.8 Classement des donnes par années

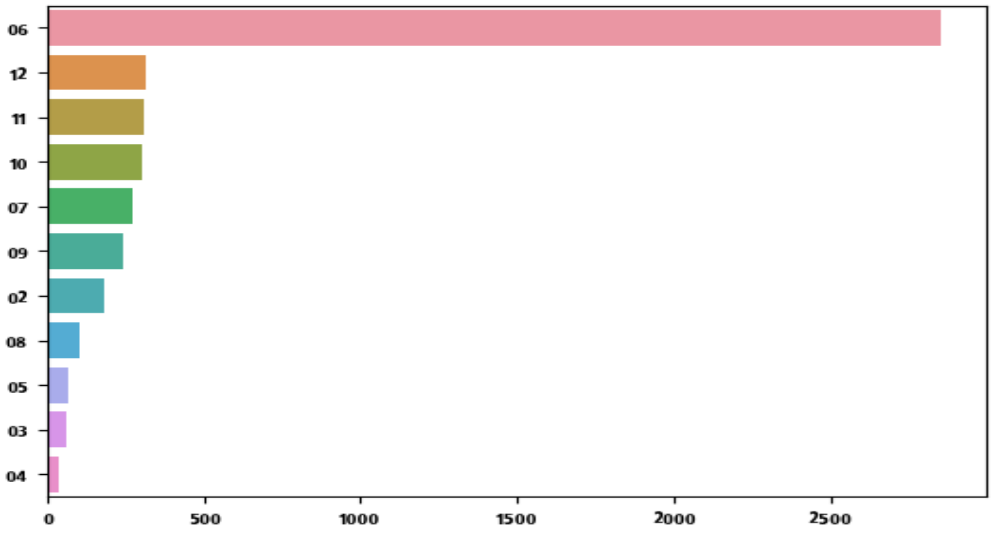


Figure 4.9 Classement des donnes par mois

3. Nettoyage & Prétraitement

Le prétraitement des données textuelles est une étape clé dans le développement d'un modèle de détection performant. Nous avons appliqué les fonctions de nettoyage et de prétraitement évoquées précédemment à notre corpus. L'exemple suivant montre les textes avant et après le prétraitement. :

Texte brut

خير من بلادو 😂😂😂😂😂 الجزائر دائما تحجز المهاجرين تقوم باعتقالهم وحبسهم في مكان مطوق بالعسكر وتقول عليهم لاجئين □□طبعا يهربون من الديكتاتورية إلى الدكتلتورية الجزائرية كيف يأكل هدا 😂

الحمد لله كي شفت التعليقات فرحت يا جزائريين أنتم شعب عظيم 💪

C\'est la meilleure cella\nIls veulent rester en Algerie

سبحان الله عايشين وحامدينها ربي يفرج على كل مومن ❤❤

علابيها عندي شحال و انا مشومر □

Texte final prétraiter

خير بلادو الجزاءر تحجز المهاجرين تقوم باعتقالهم وحبسهم مكان مطوق بالعسكر وتقول لاجءين طبعا يهربون الديكتاتورية الدكتلتورية الجزاءرية يءكل

الحمد لله شفت التعليقات فرحت جزاءرين شعب عظيم

cest meileur celan veulent rester algerie

سبحان لله عايشين وحامدينها ربي يفرج مومن

علابيها شحال مشومر

4. Vectorisation avec GloVe

Sur notre corpus, nous avons appliqué une vectorisation textuelle basée sur l'approche GloVe (6B tokens, 400K vocab, uncased, 50d vectors) décrite dans le chapitre 3. L'exemple suivant montre les vecteurs (*embeddingd*) d'un texte simple après la vectorisation :

تلقا': 17429, 'بالمءة': 17430, 'يسيحو': 17431, 'خمجوه': 17432, 'السياحي': 17433, '"رايحين": 17434, ' ,17428 'ينحو': 17435, 'القصدرية': 17436, 'جونا': 17437, 'الافرقياة': 17438, '"جزاءرية": 17439, 'واعلم': 17440, 'ظلما': 17441, 'خصرت': 17442, 'بالعامية': 17443, 'طوماطيشة': 17444, 'خاسرة': 17445, 'خنزت': 17446, 'صندوق': 17447, 'الحوادث': 17448, 'لشهدناها': 17449, 'خلانا': 17450, 'تقبلناهم': 17451, 'مامضى': 17452, 'نساندهم': 17453, 'بطيبتهم': 17454, 'كرمهم': 17455, 'نصرتهم': 17456, 'لمظلوم': 17457, 'موقفه': 17458, 'توضح': 17459, 'اتخذنا': 17460, 'لتصرفاتهم': 17461, 'فدينا': 17462, 'اخلاقنا': 17463, 'انسانيتنا': 17464, 'لبعضهم': 17465, 'فعلى': 17466, 'البحث': 17467, '"بارباس": 17468, 'يطبقو': 17469, "قراف"": 17470, 'طيبون': 17471, 'بءسلامهم': 17472, "بالءفضل"": 17473, 'البلاصة': 17474, 'يفسدوها': 17475, 'وزايد': 17476, 'زينة': 17477, 'وزادت': 17478,

5. Implémentation

Dans ce qui suit, nous montrons les couches que nous avons créées pour les deux types des réseaux de neurones utilisés. Les graphiques affichent certains paramètres (poids et biais) dans chaque couche ainsi que le total des paramètres du modèle (Figures 4.10 et 4.11).

5.1. Conception et paramétrage du modèle CNN

Nous adaptons maintenant notre modèle aux données. Ici, nous avons 10 EPOCHS et un BATCH SIZE de 64 motifs.

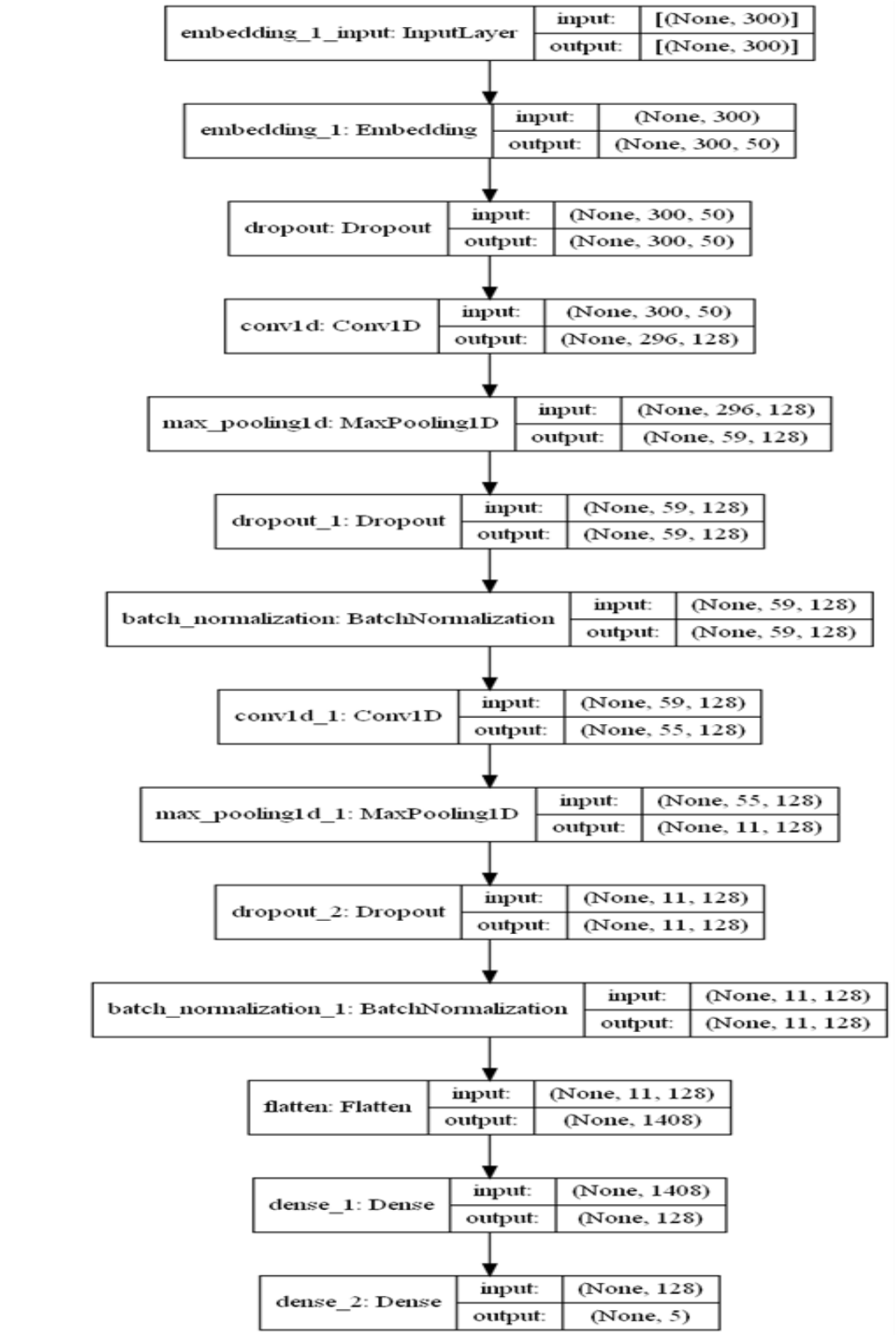


Figure 4.10 Modèle CNN

5.2. Conception et paramétrage du modèle LSTM

Et pour ce modèle, nous avons 10 EPOCHS et un BATCH SIZE de 16 motifs.

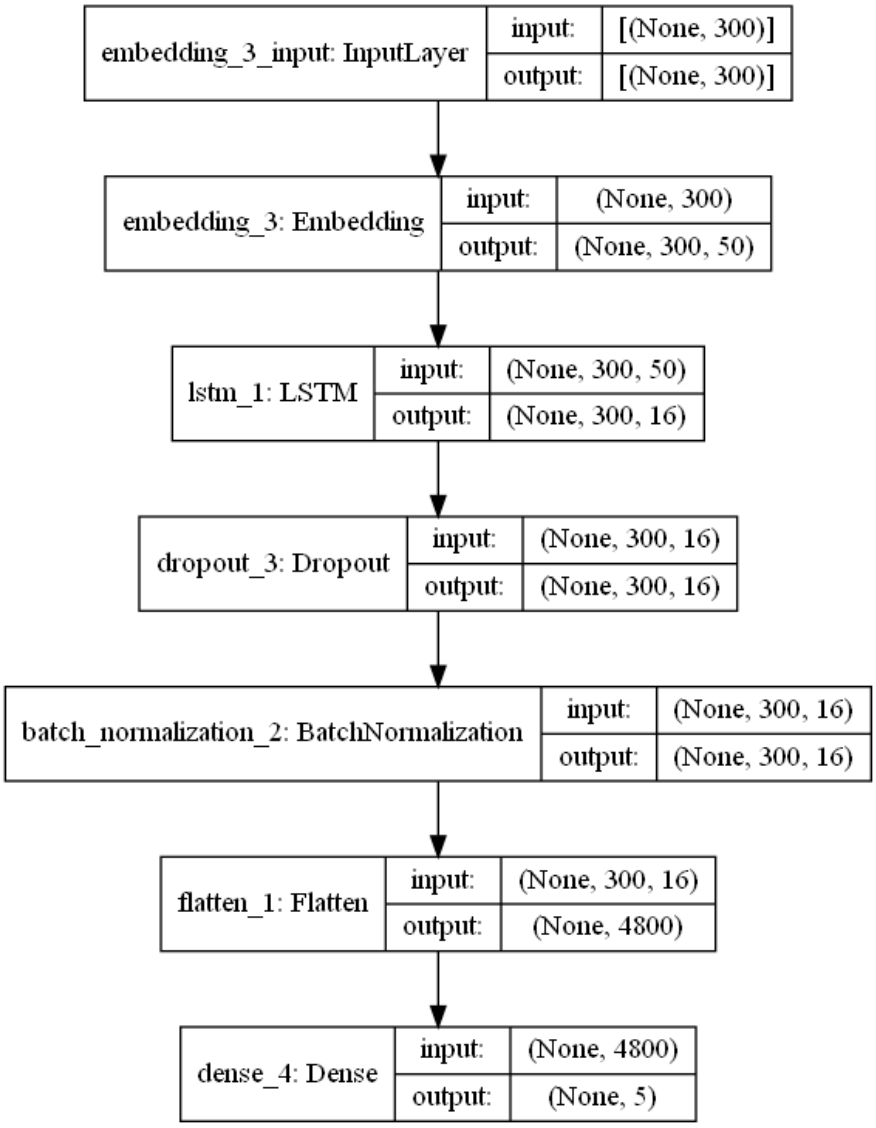


Figure 4.11 Modèle LSTM

Nous montrons aussi le développement de l'exactitude (Accuracy) du modèle et de la fonction de perte par rapport au nombre d'époques (Epochs) déterminé pour l'apprentissage.

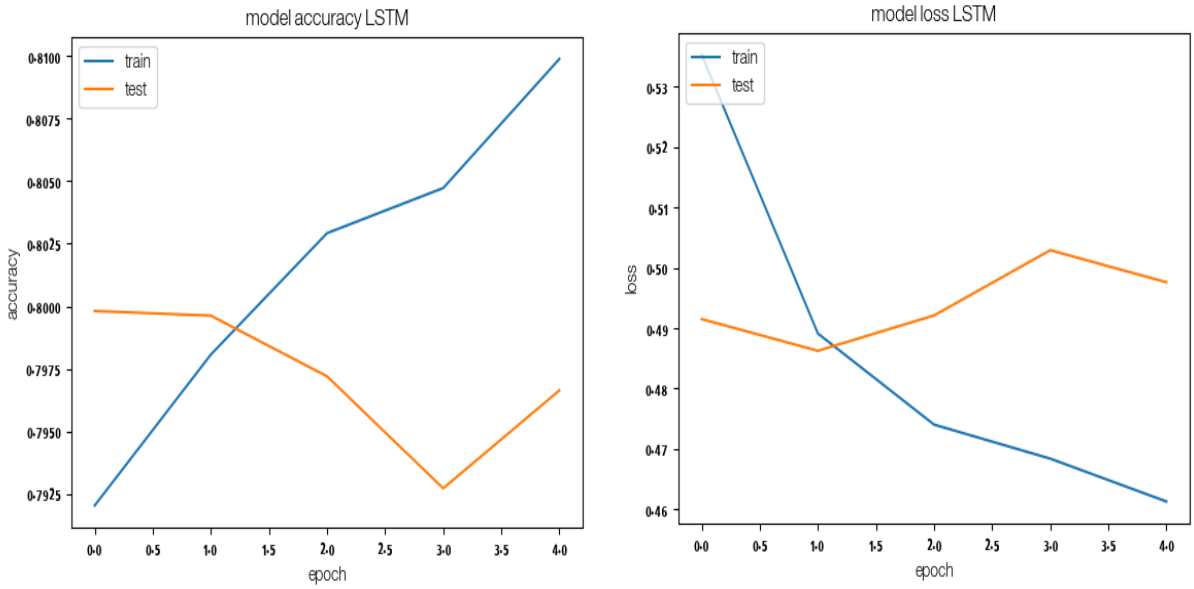


Figure 4.12 L'exactitude et la perte de la fonction d'apprentissage par rapport au nombre d'époques LSTM

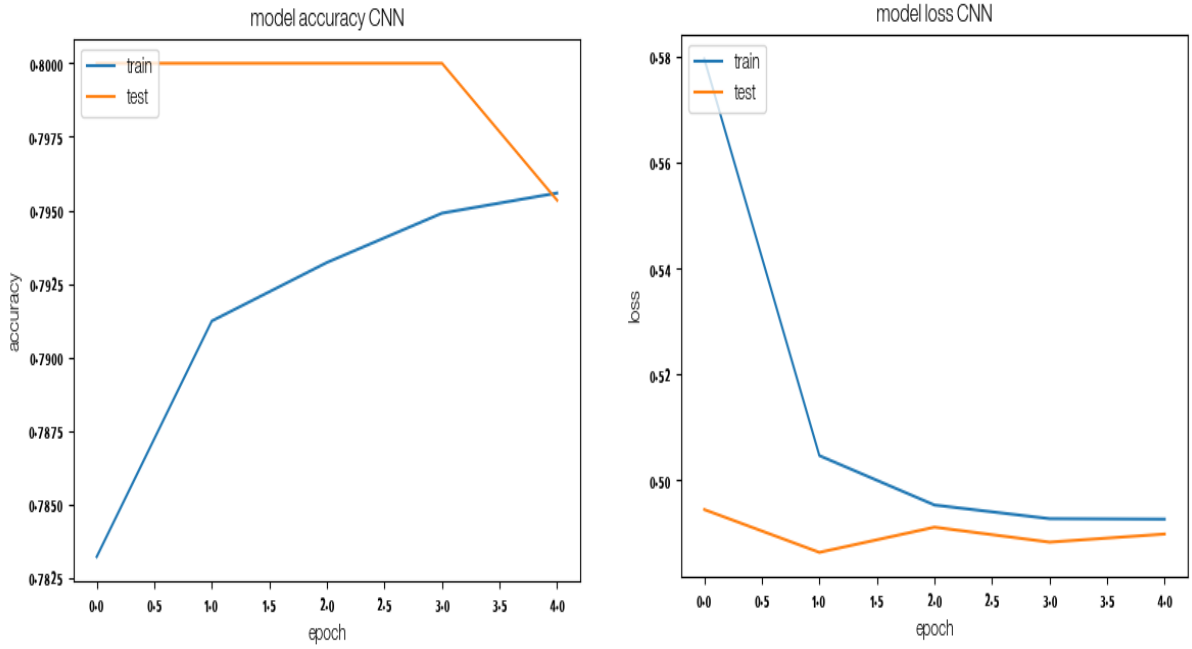


Figure 4.13 L'exactitude et la perte de la fonction d'apprentissage par rapport au nombre d'époques CNN

6. Résultats obtenus

6.1. Métriques d'évaluation

La performance des modèles de classification est généralement basée sur la façon dont ils prédisent les résultats pour les nouveaux ensembles de donnés. Cette performance est

mesurée par rapport à un ensemble de test. Plusieurs métriques déterminent les performances de prédiction d'un modèle, mais nous allons principalement se concentrer sur les métriques suivantes : [6]

- **La précision** (en anglais *precision*) : est définie comme le nombre de prédictions faites qui sont réellement correctes ou pertinent parmi toutes les prédictions basées sur la classe positive. Ceci est également connu comme **valeur prédictive** positive et peut être représentée par la formule :

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

- **Le rappel** (en anglais *recall*) : est défini comme le nombre d'instances de la classe positive qui étaient correctement prédit. Ceci est également connu sous le nom de couverture ou de sensibilité et peut être représenté par la formule :

$$Recall = \frac{TP}{\mathrm{TP} + \mathrm{FN}}$$

- **F1-Score** : est une autre mesure de précision qui est calculée en prenant la moyenne harmonique de la précision et du rappel et peut être représentée comme suit :

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

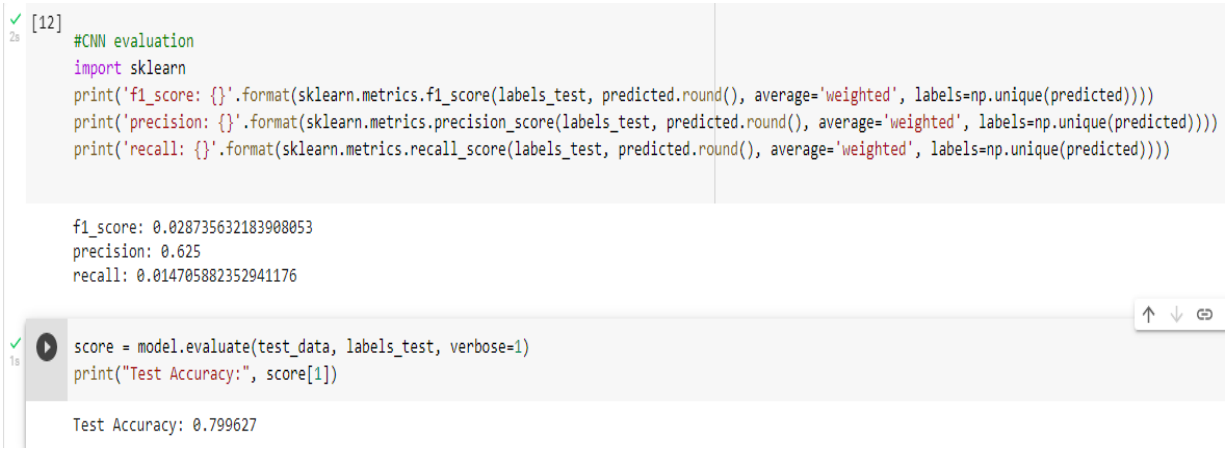
6.2. Résultats de classification

Pour classifier les textes de notre ensemble de données, nous avons appliqué deux modèles d'apprentissage profond : Réseaux de neurones à convolution CNN, Réseaux de neurones à mémoire long-proche terme LSTM.

Les résultats de classification (Tableau 4.4) sont évalués à l'aide des métriques : Précision (*Precision*), Rappel (*Recall*), F1 Score, et Taux de succès (*Accuracy*).

Modèles d'apprentissage	Accuracy	Precision	Recall	F1_Score
CNN	0.799	0,63	0,01	0,03
LSTM	0.785	0,38	0,15	0,09

Tableau 4.4. Résultats de détection automatique du discours de haine



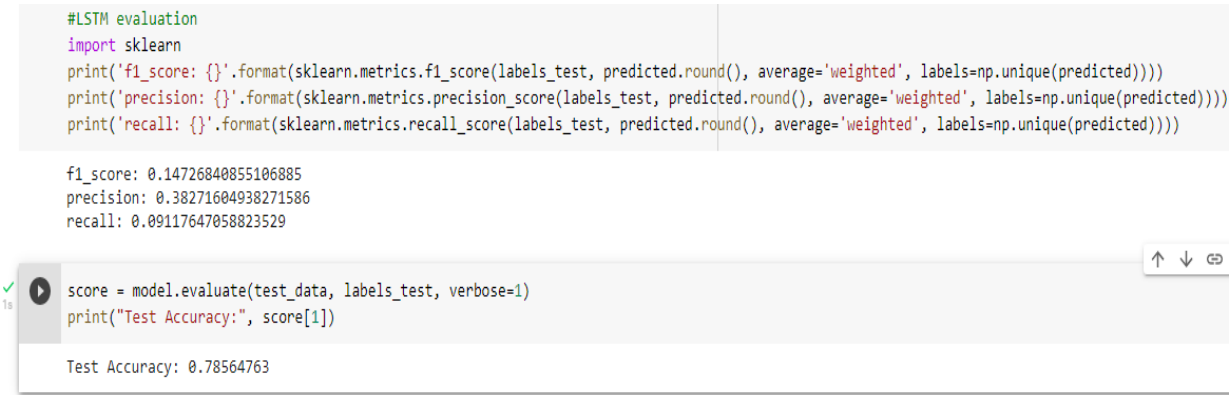
```
[12]
#CNN evaluation
import sklearn
print('f1_score: {}'.format(sklearn.metrics.f1_score(labels_test, predicted.round(), average='weighted', labels=np.unique(predicted))))
print('precision: {}'.format(sklearn.metrics.precision_score(labels_test, predicted.round(), average='weighted', labels=np.unique(predicted))))
print('recall: {}'.format(sklearn.metrics.recall_score(labels_test, predicted.round(), average='weighted', labels=np.unique(predicted))))

f1_score: 0.028735632183908053
precision: 0.625
recall: 0.014705882352941176

score = model.evaluate(test_data, labels_test, verbose=1)
print("Test Accuracy:", score[1])

Test Accuracy: 0.799627
```

Figure 4.14 Evaluation du modèle CNN



```
#LSTM evaluation
import sklearn
print('f1_score: {}'.format(sklearn.metrics.f1_score(labels_test, predicted.round(), average='weighted', labels=np.unique(predicted))))
print('precision: {}'.format(sklearn.metrics.precision_score(labels_test, predicted.round(), average='weighted', labels=np.unique(predicted))))
print('recall: {}'.format(sklearn.metrics.recall_score(labels_test, predicted.round(), average='weighted', labels=np.unique(predicted))))

f1_score: 0.14726840855106885
precision: 0.38271604938271586
recall: 0.09117647058823529

score = model.evaluate(test_data, labels_test, verbose=1)
print("Test Accuracy:", score[1])

Test Accuracy: 0.78564763
```

Figure 4.15 Evaluation du modèle LSTM

7. Discussion des résultats

Etant appliqué les deux méthodes d'apprentissage profond CNN et LSTM, nous pouvons comparer les performances de chaque modèle. Le score général calculé avec F1_Score suggère que le LSTM a obtenu de meilleurs résultats par rapport à CNN. Cela signifie que le modèle a moins d'erreurs dans la détection du discours de haine.

D'autre part, nous remarquons que le CNN a un léger avantage sur LSTM en termes d'exactitude et un remarquable résultat en termes de précision. Toutefois, sa performance se dégrade en termes de rappel et de F1Score où CNN surpasse LSTM avec une large différence.

Une justification possible est que les modèles LSTM sont plus adaptés aux données textuelles alors que les CNN sont plus adaptés aux données visuelles. C'est pour cela que CNN a obtenu des meilleurs scores en précision et en exactitude, mais un mauvais score en rappel.

Conclusion

Sur la base des données collectées au sujet des réfugiés et émigrés africains en Algérie, nous avons réalisé une étude qui décrit la mise en œuvre de la détection des discours de haine. Notre dernier chapitre sur la conception a abordé les différentes étapes d'exécution et les résultats obtenus par différents modèles. Le chapitre se termine par une interprétation des résultats basée sur les scores des métriques appropriées afin de bien comprendre les limites et les performances des modèles de classification utilisés.

Conclusion générale

La détection du discours de haine sur les réseaux sociaux à l'aide des techniques d'intelligence artificielle est très importante et efficace pour essayer de réduire ce phénomène qui affecte les gens dans une large mesure car il peut affecter la diffusion d'idées qui incitent à la haine, étant donné que la facilité de communication via les différents réseaux sociaux aide grandement les diffuseurs de haine à diffuser leurs idées librement et sans censure stricte sur certaines plateformes.

Malgré les efforts déployés par les plateformes de réseaux sociaux, il y en a qui n'ont pas été en mesure de lutter efficacement contre ce phénomène. Chaque jour, nous remarquons plusieurs publications, commentaires ou autres contenus qui appellent à la haine ou incitent des groupes de la société contre d'autres groupes, et cela contribue à créer une atmosphère propice à ces personnes pour répandre leurs croyances.

Les plateformes de médias sociaux ont créé une société parallèle à la société dans laquelle nous vivons, qui se caractérise par du négatif et du positif, et cela nous oblige à être au niveau de ce développement très rapide par l'utilisation des techniques d'apprentissage automatique et d'apprentissage approfondi pour trouver des solutions à tous les phénomènes tels que le discours de haine sur les réseaux sociaux.

À la suite de cette recherche, un certain nombre d'avancées ont été faites dans le domaine. Une analyse approfondie des données a d'abord été réalisée pour comprendre le caractère gravement déséquilibré et l'absence de caractéristiques discriminantes dans les ensembles de données typiques auxquels il faut faire face dans de telles situations. D'autre part, afin de collecter des caractéristiques implicites qui pourraient être bénéfiques pour la catégorisation. Une vaste collection d'ensembles de données pour le discours de haine a été utilisée pour tester nos approches, et ils se sont avérés extrêmement efficaces pour détecter et classer le contenu haineux (par opposition à non haineux), dont nous avons montré qu'il était plus difficile et sans doute plus important dans la pratique.

Nous espérons que ce travail sera à la hauteur et que l'application que nous avons développée sera efficace pour détecter les discours de haine, et nous espérons également que tous les lecteurs bénéficieront de la compréhension du sujet de la détection des discours de haine sur les réseaux sociaux à travers le travail modeste que nous avons fait.

Bibliographie

Articles

[1] Chakrabarty, T., Gupta, K., & Muresan, S. (2019, August). Pay "attention" to your context when classifying abusive language. In Proceedings of the Third Workshop on Abusive Language Online (pp. 70-79).

[2] Ring, C. E. (2013). Hate speech in social media: An exploration of the problem and its proposed solutions (Doctoral dissertation, University of Colorado at Boulder).

[5] Siegel, A. A. (2020). Online hate speech. Social Media and Democracy: The State of the Field, Prospects for Reform, 56-88.

[6] Chakrabarty, T., Gupta, K., & Muresan, S. (2019, August). Pay "attention" to your context when classifying abusive language. In Proceedings of the Third Workshop on Abusive Language Online (pp. 70-79).

[7] [8] Siegel, A. A. (2020). Online hate speech. Social Media and Democracy: The State of the Field, Prospects for Reform, 56-88.

[10] PRADHAN, Vidisha M., VALA, Jay, et BALANI, Prem. A survey on sentiment analysis algorithms for opinion mining. *International Journal of Computer Applications*, 2016, vol. 133, no 9, p. 7-11.

[11] Automatic Hate Speech Detection using Machine Learning: A Comparative Study ((IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 8, 2020)

[13] SARKAR, Dipanjan. Text analytics with Python: a practical real-world approach to gaining actionable insights from your data. New York: *Apress*; 2016.

[14] GOYAL, Palash, PANDEY, SUMIT, et JAIN, Karan. Deep learning for natural language processing. *Apress*, 2018, p. 138-143.

[15] GRANET, A., MORIN, E., MOUCHERE, H., QUINIOU, S., & VIARD-GAUDIN, C. Étude préliminaire CNN de reconnaissance d'écriture sur des documents historiques. *CORIA*, 2017.

[16] GOYAL, Palash, PANDEY, SUMIT, et JAIN, Karan. Deep learning for natural language processing. *Apress*, 2018, p. 138-143.

[17] Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes (Paula Fortuna)

[18] Arabic Offensive Language Detection Using Machine Learning and Ensemble Machine Learning Approaches Fatemah Husain Kuwait University, Department of Information Science, State of Kuwait

[19] Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive Language Detection on Arabic Social Media. Proceedings of the First Workshop on Abusive Language Online. Association for Computational Linguistics (ACL), 2017. pp. 52–56

[20] Arabic Offensive Language on Twitter: Analysis and Experiments Hamdy Mubarak1 , Ammar Rashed2 , Kareem Darwish1 , Younes Samih1 , Ahmed Abdelali1 1Qatar Computing Research Institute, 2Ozye ¨ gin University

[22] ETHOS: an Online Hate Speech Detection Dataset

[23] Akshay Kulkarni Adarsha Shivananda Natural Language Processing Recipes Unlocking Text Data with Machine Learning and Deep Learning using Python *Apress*, P 185-208

Sites web

[3] Social network users in leading markets 2025 | Statista (https://www.statista.com/statistics/278341/number-of-social-network-users-in-selected-countries/)

[4] Guidelines on reporting hate speech – practical tips for journalists / DW Akademie (https://www.dw.com/en/reporting-hate-speech-practical-tips-for-journalists/a-19152896)

[9] K PLUS PROCHES VOISINS KNN

ttps://www.xlstat.com/fr/solutions/fonctionnalites/k-nearest-neighbors-knn

[12] Initiez-vous au deep learning / découvrez le neurone formel. https://openclassrooms.com/fr/courses/5801891-initiez-vous-au-deep-learning/5801898-decouvrez-le-neurone-formel. (Site consulté en juin 2020).

[24] pandas documentation https://pandas.pydata.org/docs/

[21] Global Vectors for Word Representation https://nlp.stanford.edu/projects/glove/

[25] NumPy v1.21 Manual (https://numpy.org/doc/stable/)

[26] tensorflow Essential documentation https://www.tensorflow.org/guide

[27] keras Developer guides https://keras.io/guides/

[28] scikit-learn User Guide https://scikit-learn.org/stable/user_guide.html

[29] pyarabic https://pyarabic.sourceforge.io/