

République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique
Université 8 Mai 1945 – Guelma
Faculté des sciences et de la Technologie
Département d'Electronique et Télécommunications



Mémoire de fin d'étude

Pour l'obtention du diplôme de Master Académique

Domaine : **Sciences et Technologies**

Filière : **Télécommunications**

Spécialité : **Réseaux et Télécommunications**

Détection automatique des contenus offensifs en Bambara sur les réseaux sociaux

Présenté par :

DIALLO Abdoul Karim

Sous la direction du :

Dr. ABAINIA Kheireddine

Juin 2023

Remerciements

Tout d'abord, je tiens à remercier Dieu de m'avoir donné la force et la capacité de terminer ce projet.

Je suis profondément reconnaissant envers mes parents pour m'avoir facilité ma vie en me fournissant tout ce dont j'avais besoin pour réussir mes études.

Un grand merci à mon encadreur, Dr. ABAINIA Kheireddine, pour avoir accepté de me superviser, pour ses conseils avisés et sa disponibilité. Je lui suis extrêmement reconnaissant d'avoir partagé ses connaissances et sa méthode de travail avec moi.

J'adresse mes remerciements à l'ensemble du jury pour avoir accepté de participer à ma soutenance et pour avoir partagé leurs connaissances et leur expertise.

Je tiens à remercier chaleureusement tous mes oncles, tantes, camarades universitaires ainsi que mes amis spécialement : Abdramane Traoré, Moustapha Dembélé, Djibril Mounkoro, Mamoudou Sissoko, Hamdjata Diallo et Mariam Camara.

Je n'oublie pas toutes les personnes qui, de près ou de loin, ont contribué à l'élaboration de ce travail et que je n'ai pas pu citer.

Enfin, je tiens également à remercier tous ceux qui liront ce mémoire, que ce soit pour s'informer ou pour approfondir leur connaissance du sujet. Je vous souhaite une excellente continuation dans vos travaux futurs.

Merci à toutes et à tous !

Dédicace

À tous ceux qui ont été une source d'inspiration et de soutien tout au long de ce parcours.

Je souhaite dédier ce mémoire à chacun d'entre vous, car sans votre présence, votre soutien et vos encouragements, cette réalisation n'aurait pas été possible.

Vos contributions ont été inestimables et ont façonné ma vision et ma détermination.

À ma mère Fatoumata Bassi Traoré,

Celle qui a toujours cru en moi et m'as encouragé, l'oreille attentive, la main qui m'est toujours tendue, je n'ai même pas de qualificatif pour toi tellement que tu représentes tout pour moi, je te suis profondément reconnaissant ;

À feu mon père Abdoulaye Diallo,

Je souhaite rendre hommage à toi, mon modèle dont la présence me manque énormément. Ta persévérance et ta détermination m'ont toujours inspiré. Même si tu n'es plus physiquement présent, tu resteras à jamais dans ma vie.

À mon beau-père Daouda Dembélé,

Je tiens à t'exprimer ma profonde gratitude pour l'amour, le soutien et la bienveillance que vous m'apportez dans ma vie. Je vous remercie de tous les conseils que vous me donnez et je vous encourage à poursuivre, je vais m'en servir à bon escient.

À tous mes oncles et tantes : Jean Claude Coulibaly, Nèma Sagara, Adama Bassi Traoré, Joël Mounkoro, Biahana Diassana et Aliou Badara Sadia Tounkara,

Je voudrais vous exprimer mes sincères remerciements pour vos soutiens, vos encouragements et vos conseils. Je vous remercie pour tout ce que vous avez fait pour moi,

Pour finir je le dédie à tous mes frères et sœurs.

Abdou Karim Diallo

Résumés

Résumé :

Dans ce mémoire, nous abordons le problème de la détection automatique contenus offensifs en bambara sur les réseaux sociaux. Avec la grande popularité des réseaux sociaux, l'objectif est de protéger les utilisateurs contre les discours haineux, les insultes et les discriminations en ligne. Cependant, pour les langues moins documentées comme le Bambara, il n'existe pas de moyen pour pratiquer cette détection. Une approche basée sur l'apprentissage automatique est utilisée à l'aide de sept algorithmes et des techniques de prétraitement. Les résultats montrent une grande efficacité, notamment avec le modèle BiLSTM qui a produit des valeurs importantes de F1 (i.e. 0.94 et 0.88). Malgré le manque de données à notre possession, cette recherche contribue à la protection des utilisateurs et souligne l'importance de la recherche dans les langues moins documentées pour la détection des contenus offensifs sur les réseaux sociaux.

Mots-clés : Détection automatique, contenus offensifs, Bambara, réseaux sociaux, discours haineux, insultes, discriminations, langues moins documentées, apprentissage automatique, sécurité en ligne.

Abstract:

In this work, we deal with automatic detection of Bambara offensive content on social media. With the popularity of social media, our goal is to protect users from hate speech, insults and online discrimination. Nevertheless, for under-resourced languages like Bambara, it is tricky to achieve high performance in this task. To overcome this problem, we have used machine learning and deep learning approaches (seven classifiers) with a set of preprocessing tasks dedicated to Bambara. The evaluation results showed acceptable results produced by different classifiers, and in particular BiLSTM that produced the highest F1-score (i.e., 0.94 and 0.88). Despite the lack of data in our hand, this research in progress contributes to the protection of the users and underlines the importance of the research in under-resourced languages.

Keywords: Automatic detection, offensive content, Bambara, social networks, hate speech, Insults, discrimination, under-resourced languages, machine learning, online security.

Table des matières

REMERCIEMENTS	i
DEDICACE	ii
RESUMES	iii
SOMMAIRE	iv
LISTE DES FIGURES	vii
LISTE DES TABLEAUX	viii
LISTE DES ACRONYMES	ix
INTRODUCTION GENERALE	1
CHAPITRE I : ETAT DE L'ART SUR TRAITEMENT DU LANGAGE NATUREL	3
1. INTRODUCTION	4
2. ÉTUDE PRELIMINAIRE	4
2.1. Intelligence artificielle	4
2.2. Apprentissage automatique des machines	5
2.3. Linguistique calculée et traitement du langage naturel	6
3. HISTOIRE	7
4. PRINCIPES FONDAMENTAUX DE LA LINGUISTIQUE INFORMATIQUE	8
4.1 Linguistique	8
4.2. Informatique	8
4.3. Analyse du langage naturel	8
4.4. Modèles du langage	9
4.5. Corpus linguistiques	9
5. APPLICATIONS	9
5.1. Reconnaissance de la parole	9
5.2. Traduction automatique	10
5.3. Analyse des sentiments	10
5.4. Chatbots	11
5.5. Classification des textes	12
5.6. Extraction d'informations	12
6. LIMITES ET DEFIS	13
6.1. Sémantique	13
6.2. Ambiguïté	13
6.3. Humour et Poésie	14
6.4. Variabilité des langues	14

6.5. Données	14
7. CONCLUSION	15
CHAPITRE II : DECOUVERTE DE LA LANGUE BAMBARA.....	16
1. INTRODUCTION	17
2. GENERALITES SUR LA LANGUE ET LE DIALECTE	17
2.1. Langue	17
2.2.1. Dialecte	19
2.3.1. Différence entre langue et dialecte	19
3. LANGUE BAMBARA	20
3.1. Généralités	20
3.2. Caractéristiques de la langue	21
3.3. Ecriture en Bambara	21
3.1.4. Langue Bambara sur les réseaux sociaux	23
4. CONCLUSION	24
CHAPITRE III : PRESENTATION DU LANGAGE OFFENSIF.....	25
1. INTRODUCTION	26
2. DEFINITION.....	26
3. DIFFERENTS TYPES DU LANGAGE OFFENSIF	26
3.1. Violence verbale	26
3.2. Hate speech ou discours haineux	27
3.3. Langage abusif.....	27
3.4. Langage discriminatoire	28
4. CONSEQUENCES DU LANGAGE OFFENSIF.....	28
4.1. Effets psychologiques	28
4.2. Effets sociaux.....	29
4.3. Effets physiques.....	29
4.4. Effets sur la communauté	29
5. DETECTION AUTOMATIQUE DU LANGAGE OFFENSIF.....	30
5.1. Méthodes à base d'apprentissage ou Machine Learning	31
5.2. Apprentissage profond	34
5.3. Avantages et inconvénients des algorithmes d'apprentissage.....	38
6. CONCLUSION	39
CHAPITRE IV : METHODOLOGIE ET RESULTATS	40
1. INTRODUCTION	41
2. TRAVAUX CONNEXES	41

3. BASE DE DONNEES.....	42
3.1. Source d'informations	42
3.2. Collecte des données.....	43
3.3. Annotation	43
4. PRETRAITEMENT DES TEXTES BAMBARA.....	44
5. CLASSIFICATION	45
5.1. Méthodes d'évaluation des classifieurs	46
5.2. Paramètres des classifieurs	47
6. EVALUATION ET RESULTATS.....	50
6.1. Résultats.....	50
6.2. Discussions	54
7. CONCLUSION	55
CONCLUSION GÉNÉRALE.....	56
REFERENCES BIBLIOGRAPHIQUES.....	58

Liste des figures

FIGURE I.1: ILLUSTRATION D'APPLICATION DE LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE	10
FIGURE I.2: CAPTURE D'ECRAN SUR LE SERVICE GOOGLE TRANSLATE DE TRADUCTION AUTOMATIQUE.....	10
FIGURE I.3: EXEMPLE D'UN OUTIL D'ANALYSE DES SENTIMENTS	11
FIGURE I.4: EXEMPLE D'APPLICATION DE MESSENGER POUR UNE AGENCE DE LOCATION DE VOITURES	12
FIGURE II.1: ILLUSTRATION DE LA FAMILLE INDO-EUPEENNE	18
FIGURE II.2:ILLUSTRATION GRAPHIQUE DE QUELQUES LANGUES NIGERO-CONGOLAISE	19
FIGURE II.3: ALPHABET BAMBARA	22
FIGURE II.4: ALPHABET N'KO	23
FIGURE III.1: ILLUSTRATION DU SUPPORT VECTOR MACHINE (SVM) LINEAIRE ET NON LINEAIRE.....	31
FIGURE III.2: ILLUSTRATION DE L'ARBRE DE DECISION	32
FIGURE III.3: ILLUSTRATION D'UN EXEMPLE DE L'ALGORITHME DES K-VOISINS LES PLUS PROCHES	33
FIGURE III.4: ILLUSTRATION D'UN RESEAU DE NEURONES	34
FIGURE III.5: ARCHITECTURE CLASSIQUE D'UN RESEAU DE NEURONES CONVOLUTIF	35
FIGURE III.6: ILLUSTRATION D'UN RESEAU NEURONAL RECURRENT DEROULE	36
FIGURE III.7: ARCHITECTURE DE BASE D'UN LSTM	37
FIGURE III.8: ARCHITECTURE DE BASE D'UN BiLSTM.....	37
FIGURE IV.1: STATISTIQUES SUR LE CAS DE BALANCED CORPUS POUR 2ET 3 CLASSES.....	46
FIGURE IV.2: STATISTIQUES SUR LE CAS DE BALANCED CORPUS POUR 2ET 3 CLASSES.....	46

Liste des tableaux

Table 2.1 : Exemple de quelques textes Bambara sur les réseaux sociaux	24
Table 3.1 : Tableau récapitulatif des avantages et les inconvénients des algorithmes à base d'apprentissage	38
Table 4.1 : Nombre d'agrément des trois annotateurs	44
Table 4.2 : Paramètres utilisés dans le classifieur CNN	48
Table 4.3 : Paramètres utilisés dans le classifieur BiLSTM	49
Table 4.4: Paramètres utilisés pour entraîner l'outils FastText	49
Table 4.5 : Résultats du SVM avec deux et trois classes sur le corpus équilibré	50
Table 4.6 : Résultats du SVM avec deux et trois classes sur le corpus déséquilibré	50
Table 4.7 : Résultats de régression logistique avec deux et trois classes sur le corpus équilibré	50
Table 4.8 : Résultats de régression logistique avec deux et trois classes sur le corpus déséquilibré	51
Table 4.9 : Résultats de NB avec deux et trois classes sur le corpus équilibré	51
Table 4.10 : Résultats de NB avec deux et trois classes sur le corpus déséquilibré	51
Table 4.11 : Résultats de Random Forest avec deux et trois classes sur le corpus équilibré	51
Table 4.12 : Résultats de Random Forest avec deux et trois classes sur le corpus déséquilibré	51
Table 4.13 : Résultats du SGD avec deux et trois classes sur le corpus équilibré	51
Table 4.14 : Résultats du SGD avec deux et trois classes sur le corpus déséquilibré	52
Table 4.15 : Résultats du CNN avec deux et trois classes sur le corpus équilibré	52
Table 4.16 : Résultats du CNN avec deux et trois classes sur le corpus déséquilibré	52
Table 4.17 : Résultats du BiLSTM avec deux et trois classes sur le corpus équilibré	52
Table 4.18 : Résultats du BiLSTM avec deux et trois classes sur le corpus déséquilibré	52
Table 4.19 : Résultats du FastText avec deux et trois classes sur le corpus équilibré	52
Table 4.20 : Résultats du FastText avec deux et trois classes sur le corpus déséquilibré	53

Liste des acronymes

AD : Arbre de Décision

ANN : Artificial Neural Network (réseau neuronal artificiel)

BiLSTM : Bidirectional Long Short-Term Memory (mémoire à court terme bidirectionnelle)

CL : Computational Linguistics (Linguistique informatique)

CNN : Convolutional Neural Network (réseau neuronal convolutif)

DL : Deep Learning (apprentissage profond)

GloVe : Global Vectors (vecteurs globaux)

KNN : K-Nearest Neighbors (k-plus proches voisins)

LR : Logistic Regression (régression logistique)

ML : Machine Learning (apprentissage automatique)

MT : Machine Translation (traduction automatique)

NB: Naive Bayes (Bayes naïf)

NLP : Natural Language Processing (traitement du langage naturel)

NLU : Natural Language Understanding (compréhension du langage naturel)

ODA : Oracle Digital Assistant

OSV : Object-Subject-Verb (objet-sujet-verbe)

RNN : Recurrent Neural Network (réseau neuronal récurrent)

SVC : Support Vector Classification (classification par vecteurs de support)

SVM : Support Vector Machine (machine à vecteurs de support)

SVO : Subject-Verb-Object (sujet-verbe-objet)

TALN : Traitement Automatique du Langage Naturel

TF-IDF : Term Frequency-Inverse Document Frequency (fréquence de terme-inverse de la fréquence documentaire)

INTRODUCTION GÉNÉRALE

De nos jours, dans un monde de plus en plus connecté, les réseaux sociaux sont devenus incontournables dans notre société moderne, ils offrent aux individus un moyen de communication, de partage d'informations et surtout un moyen pour s'exprimer librement. En 2023, leur nombre d'utilisateurs atteignait approximativement 60 % de la population mondiale. Au Mali où nous trouvons la plus grande communauté bambarophone, comme ailleurs en Afrique, l'utilisation des réseaux sociaux pour la grande majorité de notre génération est devenue une partie indispensable de la vie quotidienne créant une plateforme pour la diffusion d'information, d'opinion et de divertissements, ainsi pour beaucoup ne pas avoir de présence sur les réseaux sociaux équivaut quasiment à ne pas exister.


Cependant, cette popularité des réseaux sociaux comporte plein de risques qui menacent la sécurité et le bien-être des utilisateurs, notamment la propagation des contenus offensifs, tels que des discours haineux, des insultes et des discriminations envers certaines personnes ou groupes. Pour beaucoup de langues ressourcées ou documentées (i.e. anglais ou français), on distingue déjà plusieurs techniques pour protéger les utilisateurs, ce qui n'est pas le cas pour les langues moins documentées comme le bambara.

C'est évidemment à cause de plusieurs de ces facteurs que nous avons choisis le bambara comme langue pour notre travail, car en dehors de l'aspect historique de la langue, c'est une langue utilisée comme moyen de communication entre des individus de langue maternelle distincte par la majorité des Maliens. De plus, il est parlé dans beaucoup d'autres pays frontaliers du Mali (même si c'est généralement sous différentes variantes), mais aussi pour la simple raison que depuis plusieurs années le Mali comme d'autres pays du sahel, font face à une crise multidimensionnelle dans laquelle nous pouvons retrouver entre autres : une vague de terroristes djihadistes, des conflits communautaires et une crise politique.

Les conséquences de ces conflits ont augmenté des tensions et des divisions au sein de la population, créant de plus en plus des pages à polémiques, qui se multiplient sur les réseaux sociaux et qui sont l'origine de plusieurs troubles dans la société. C'est dans ce contexte complexe que nous avons décidé de développer des méthodes efficaces pour détecter et atténuer les contenus offensifs pour le bambara, afin de protéger les utilisateurs du bambara sur les réseaux sociaux pour promouvoir la paix au sein de la population.

Dans ce mémoire, nous adopterons une approche basée sur l'apprentissage automatique pour la détection des contenus offensifs en bambara. Nous allons collecter un ensemble de données provenant de publications en bambara. Ensuite, nous appliquerons des techniques de prétraitement des textes pour normaliser et nettoyer les données, avant de développer un modèle d'apprentissage automatique capable de détecter les contenus offensants. Nous allons utiliser sept (7) algorithmes d'apprentissage automatique (ML) et d'apprentissage profond (DL) pour notre tâche de classification, qui sont (BiLSTM, CNN, FastText, LR, Multinomial NB, SVM et SGD Classifier). Les tests seront réalisés en deux (2) parties : une première partie avec trois catégories (normal, offensive et abusive) et une deuxième partie combinant les catégories offensive et abusive pour obtenir deux catégories distinctes. Nous utiliserons aussi deux bases de données différentes, une avec des catégories équilibrées et l'autre déséquilibrées pour visualiser l'effet de la quantité des données sur les résultats.

Ce mémoire est structuré en quatre chapitres principaux. Le premier chapitre présente une revue de la littérature sur le traitement du langage naturel. Le deuxième chapitre présente la langue Bambara. Le troisième chapitre décrit le langage offensif. Le quatrième chapitre décrit en détail notre méthodologie de collecte de données et de prétraitement présente dans les techniques d'apprentissage automatique et les algorithmes utilisés pour la détection des contenus offensifs en bambara ainsi qu'une illustration des résultats de nos expérimentations. Enfin, la dernière section conclut notre travail en soulignant nos remarques, les limites et les perspectives pour de futures recherches.



Chapitre I : Etat de l'art sur traitement du langage naturel

1. Introduction

La linguistique informatique est un domaine interdisciplinaire en constante évolution avec de nombreux noms correspondants à ses nombreuses facettes, telles que la linguistique computationnelle (en anglais computational linguistics ou CL), elle utilise des méthodes de l'informatique, de la linguistique et de la statistique pour étudier les langues naturelles et leur utilisation dans les systèmes informatiques.

Le but de ce nouveau domaine est de permettre aux ordinateurs d'effectuer des tâches utiles impliquant le langage humain, telles que faciliter la communication homme-machine, améliorer la communication entre humains ou simplement effectuer un traitement utile du texte ou de la parole [1].

Ce chapitre présente une introduction aux principes fondamentaux du traitement du langage naturel, où on commence par un aperçu sur l'intelligence artificielle, l'apprentissage automatique et la linguistique informatique. Par la suite, nous mettrons en évidence une brève biographie, son mode de fonctionnement, ses applications ainsi que les limites et défis. De plus, nous discutons également des technologies qui lui sont associées.

2. Étude préliminaire

2.1. Intelligence artificielle

L'intelligence artificielle comme on la connaît est un domaine qui chaque jour évolue un peu plus en s'appropriant des capacités cognitives humaines, en les développant et parfois même en surpassant ce que les meilleurs êtres humains dans leur domaine sont capables de réaliser [2].

On s'en serait douté, créer des agents intelligents n'est pas si simple. Pour cette raison, l'IA s'est divisée en de nombreuses sous-disciplines qui essaient chacune de traiter une partie d'un problème [3].

De nos jours, elle a diverses applications dans différents domaines pour aider à résoudre des problèmes complexes de manière efficace, où elle pourrait être utilisée comme dans les soins de santé, l'éducation, les médias sociaux, la robotique, la finance, etc. [4]

Pour de nombreux chercheurs du domaine, l'objectif scientifique à long terme de l'IA en tant que discipline est la « mécanisation » de l'intelligence humaine [5].

Certains spécialistes reconnus du domaine considèrent même que l'IA n'est pas si « intelligente » qu'on le dit, selon eux ce n'est pas l'intelligence qui caractérise les systèmes d'IA d'aujourd'hui, mais leur capacité de reconnaissance grâce à l'apprentissage machine sur lequel nous reviendrons [5].

2.2. Apprentissage automatique des machines

L'apprentissage automatique des machines (ou Machine Learning en anglais) est une branche de l'intelligence artificielle qui permet à des ordinateurs d'apprendre automatiquement à partir d'un ensemble de données, sans être explicitement programmés. En particulier, il utilise des algorithmes et des modèles statistiques pour analyser des données, trouver des relations entre les variables et effectuer des prévisions ou des classifications basées sur ces relations.

Les modèles d'apprentissage des machines sont entraînés à partir de données d'entraînement en utilisant diverses techniques, telles que : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement.

Pour de nombreux auteurs, les nouvelles perspectives d'automatisation ouvertes par l'IA nécessitent de repenser la division du travail entre humains et machines, car les progrès réalisés dans ce domaine font que les systèmes d'IA deviennent plus performants que les humains dans de nombreux cas comme le diagnostic ou l'analyse prédictive. Par exemple, on nous annonce régulièrement que des systèmes d'IA obtiennent de meilleures performances que les médecins dans la détection ou l'estimation de probabilités d'apparition de maladies [5].

Voici une liste des avancées et des exploits majeurs dans le domaine du Machine Learning au fil de ces dernières années :

- 2014 : Google développe un système de reconnaissance d'images avec une précision proche de celle des humains [6] ;
- 2015 : AlphaGo de Google DeepMind bat le champion du monde de Go, Lee Sedol [7] ;
- 2016 : AlphaGo Zéro de Google DeepMind apprend à jouer à Go sans aucune connaissance préalable des humains [8] ;
- 2018 : OpenAI crée une IA capable de rédiger des textes presque indiscernables de ceux écrits par des humains [9] ;

- 2019 : GPT-2 de OpenAI génère des textes de qualité presque humaine, mais suscite des préoccupations quant à l'utilisation abusive potentielle [9] ;
- 2020 : GPT-3 de OpenAI bat des records de performance en matière de génération de texte et de compréhension du langage naturel [9] ;
- 2021 : AlphaFold de Google DeepMind résout le problème de prédiction de la structure des protéines, qui était considéré comme l'un des problèmes les plus importants en biologie structurale [10] ;

Il est important de noter que de nombreuses avancées ont été rendues possibles grâce à des années de recherche et de développement, et qu'elles ont souvent été le résultat de collaborations entre plusieurs entreprises, équipes de recherche et individus.

2.3. Linguistique calculée et traitement du langage naturel

La linguistique calculée (computational linguistics ou CL) et traitement du langage naturel (Natural Language processing ou NLP) sont deux domaines qui se concentrent sur l'utilisation du calcul pour comprendre et générer le langage naturel. NLP se concentre principalement sur les tâches de traitement du langage naturel telles que la reconnaissance vocale, l'analyse syntaxique et sémantique et la génération de textes. CL, d'autre part, est un domaine interdisciplinaire plus large qui combine la linguistique et l'informatique pour étudier le langage naturel et son utilisation dans les systèmes informatiques. Bien que le NLP soit un sous-ensemble du CL, les deux domaines sont importants pour la compréhension informatique et la génération du langage naturel, avec des implications pour de nombreux domaines [11].

Par exemple, le NLP est la technologie de base des assistants virtuels, tels que l'Oracle Digital Assistant (ODA), Siri, Cortana ou Alexa. Lorsque nous posons des questions à ses assistants virtuels, le NLP leur permet non seulement de comprendre la demande de l'utilisateur, mais aussi de répondre en langage naturel. Le NLP s'applique à la fois au texte écrit et à la parole et peut être appliqué à toutes les langues humaines [12].

3. Histoire

La linguistique informatique est l'étude scientifique du langage d'un point de vue informatique [13]. La linguistique informatique est apparue aux États-Unis au début de la seconde moitié du XXe siècle et a pour objectif de faire en sorte qu'un ordinateur puisse traduire automatiquement des textes (certains affirment que la finalité était de pouvoir déchiffrer les codes de communication du bloc communiste). À cette époque, les scientifiques ont commencé à explorer comment les ordinateurs pouvaient être utilisés pour traiter et analyser le langage naturel [14].

En 1950, Alan Turing a proposé un test appelé test de Turing, qui est devenu une référence pour évaluer la capacité d'une machine à démontrer un comportement intelligent équivalent ou indiscernable de celui d'un être humain. Le test consiste à évaluer si une machine est capable de converser de manière suffisamment naturelle pour tromper un humain.[15]

Au cours des années 1960, le développement de l'intelligence artificielle et la théorie de la grammaire transformationnelle de Noam Chomsky ont conduit à l'émergence de la linguistique informatique moderne. Des chercheurs pionniers tels que Zellig Harris, Yehoshua Bar-Hillel et Noam Chomsky ont combiné la théorie du langage et l'informatique pour jeter les bases de la linguistique informatique [15].

Les années 1970 et 1980 ont vu des avancées significatives dans la conception de systèmes de traitement du langage naturel utilisant des algorithmes statistiques et des techniques d'apprentissage automatique. Des systèmes tels que la traduction automatique, la reconnaissance vocale et le traitement de texte ont été développés et mis en pratique.

Dans les années 1990 et 2000, l'avènement d'Internet et des grandes quantités de données textuelles ont permis le développement de nouvelles techniques de traitement du langage naturel telles que l'analyse des sentiments et la reconnaissance des entités nommées. Aujourd'hui, grâce aux progrès de l'intelligence artificielle et de l'apprentissage automatique, la linguistique informatique évolue rapidement et est devenue un domaine important pour de nombreuses applications telles que : la recherche d'informations, compréhension du langage naturel, la génération de texte et la communication homme-machine [16].

4. Principes fondamentaux de la linguistique informatique

En utilisant les concepts linguistiques et les techniques informatiques, la linguistique informatique cherche à comprendre et à traiter le langage naturel de manière automatisée, ces domaines constituent les bases théoriques et méthodologiques de cette discipline interdisciplinaire, qui est basée sur les fondamentaux suivants :

4.1 Linguistique

La linguistique est la discipline qui s'occupe de l'étude scientifique du langage humain. [17]. Elle s'intéresse notamment à l'étude des aspects formels des langues tels que la phonétique, la phonologie, la morphologie, la syntaxe, la sémantique et la pragmatique. Appliquée à une langue donnée, l'étude de ces aspects formels constitue l'étude de la grammaire de cette langue. Une comparaison des grammaires des langues permet de mettre au jour les caractéristiques des langues humaines [18]. La linguistique informatique utilise les concepts et les outils de la linguistique pour comprendre et traiter le langage naturel à l'aide de l'informatique.

4.2. Informatique

De nos jours ce n'est plus une discipline à présenter alors disons juste pour faire simple que c'est la science du traitement automatique et rationnel de l'information.

La linguistique informatique utilise des algorithmes informatiques pour traiter le langage naturel, tels que l'apprentissage automatique et les réseaux de neurones.

4.3. Analyse du langage naturel

L'analyse du langage naturel est le processus d'extraction de la signification du langage naturel. Un de ces sous-domaines, c'est la compréhension du langage naturel (Natural Language Understanding en anglais ou NLU). Cette dernière permet de comprendre ce que signifie exactement un corpus de texte, pour classer, archiver et analyser le texte. En outre, ce domaine atteint un niveau supérieur permettant la prise de décision en fonction de la signification [19].

4.4. Modèles du langage

Un modèle de langage est une fonction ou un algorithme d'apprentissage d'une telle fonction, qui capture les principales caractéristiques statistiques de la distribution de séquences de mots dans une langue naturelle [20]. Ce modèle permet généralement de faire des prédictions probabilistes du mot suivant étant donné les précédents [20]. C'est aussi une représentation mathématique des règles linguistiques.

4.5. Corpus linguistiques

Quand la linguistique du corpus est définie, elle constitue pour certains une discipline qui s'est développée à partir des années 1980, mais qui a vraiment pris son essor avec l'arrivée sur le marché d'ordinateurs personnels [21]. C'est une discipline très liée à l'utilisation de l'informatique, où les corpus linguistiques sont des ensembles de données textuelles qui sont utilisés pour entraîner des modèles du langage [21]. Ils sont aussi utilisés pour comprendre les variations régionales ou culturelles du langage naturel.

5. Applications

Dans cette section, nous citons quelques applications les plus fréquentes et les plus connues de la linguistique informatique.

5.1. Reconnaissance de la parole

Elle est appelée également la reconnaissance automatique de la parole ou la reconnaissance de la parole par ordinateur, où elle utilise le traitement du langage naturel pour transcrire la parole humaine dans un format écrit. Il est bon à savoir qu'elle est parfois confondue avec la reconnaissance vocale, qui cherche uniquement à identifier la voix d'un utilisateur individuel, tandis que la reconnaissance de la parole se concentre sur la traduction d'une parole de la forme verbale en un texte [22].

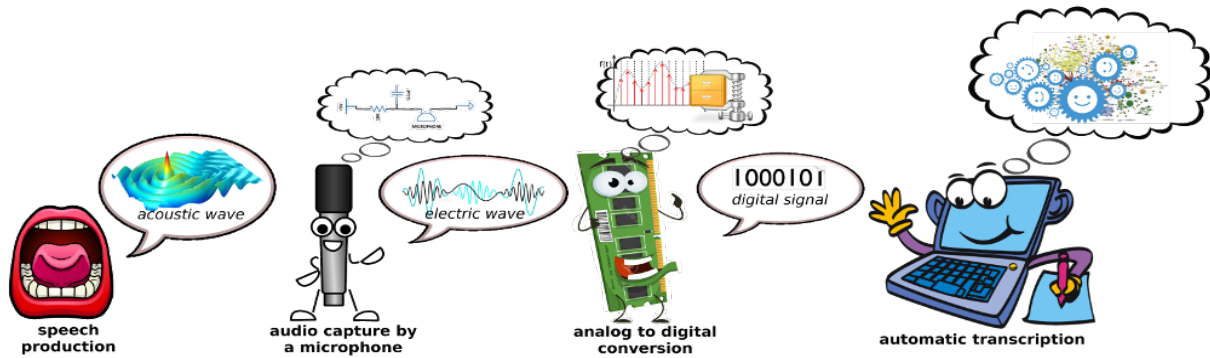


Figure I.1: Illustration d'application de la reconnaissance automatique de la parole [36]

5.2. Traduction automatique

La traduction automatique (machine translation en anglais ou MT) est le processus de traduction automatique d'un texte d'une langue à une autre. La linguistique informatique joue un rôle important dans la traduction automatique, car elle implique l'analyse de la structure des langues source et cible, et le développement d'algorithmes pour traduire avec précision les textes. Elle est considérée comme l'un des domaines les plus difficiles de la linguistique informatique, car elle nécessite une connaissance approfondie dans différents domaines, à savoir l'informatique, la linguistique, la cognition du langage, la traduction et les méthodes de description du langage [4].

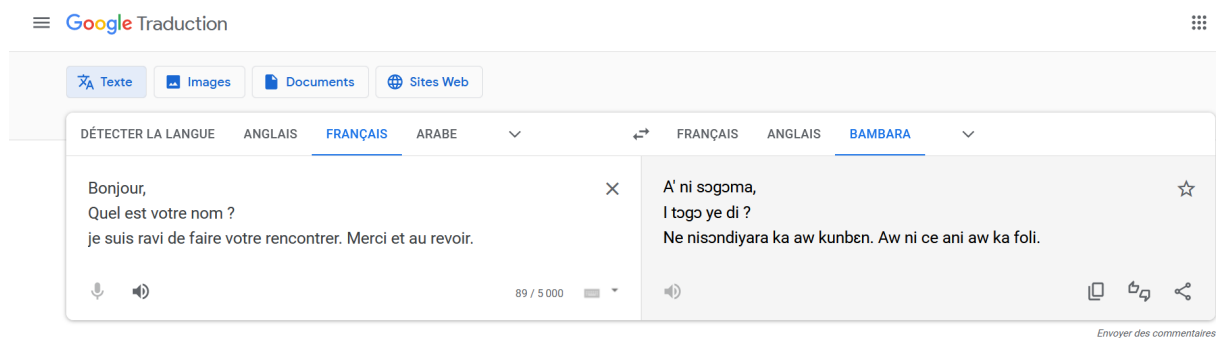


Figure I.2: Capture d'écran sur le service Google Translate de traduction automatique

5.3. Analyse des sentiments

D'une manière générale, l'analyse des sentiments est connue également sous plusieurs notations comme sentiment analysis ou opinion mining. C'est une procédure permettant de suivre les opinions et les émotions exprimées par les clients (ou auteurs) autour d'un objet ou

d'un sujet spécifique [23]. Elle comprend la création d'un cadre pour recueillir et examiner les opinions sur l'élément dans les entrées de blogs, les remarques, les audits ou les tweets [23].

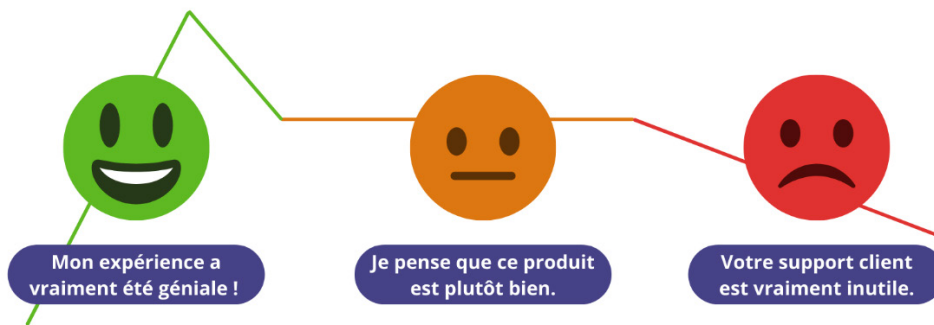


Figure I.3:Exemple d'un outil d'analyse des sentiments [29]

5.4. Chatbots

Un Chatbot, aussi appelé « agent conversationnel », est un programme informatique capable de simuler une conversation avec un ou plusieurs humains par échange vocal ou textuel [24].

La linguistique informatique est utilisée pour développer des algorithmes qui permettent à ces programmes de comprendre l'entrée en langage naturel et de générer une sortie en langage naturel. Cet agent conversationnel est très utilisé dans le domaine du marketing, où les entreprises mettent en place des chatbots dans le but de faciliter les relations avec les internautes. Les utilisateurs bénéficient en effet d'un service client amélioré. Les réponses avec cette machine de chat sont instantanées, claires et très utiles. De plus, le robot intelligent est souvent utilisé en tant que service client. Il est disponible à tout moment pour répondre à toutes les questions. Il s'adapte pour fournir toujours des réponses satisfaisantes. Il peut même arriver qu'il transfère les informations collectées au département concerné au sein de la société [25].



Figure I.4: Exemple d'application de Messenger pour une agence de location de voitures [30]

5.5. Classification des textes

La classification des textes est une application importante de la linguistique informatique qui consiste à catégoriser des documents textuels en fonction de leur contenu. Cette catégorisation peut être réalisée à l'aide de différentes techniques d'analyse de texte telles que l'analyse lexical, l'analyse sémantique et l'analyse syntaxique.

Elle connaît ces derniers temps un fort regain d'intérêt. Cela est dû essentiellement à la forte croissance des documents numériques disponibles, et à la nécessité de les organiser de façon rapide [26].

5.6. Extraction d'informations

Un système de langage informatique destiné à collecter et extraire rapidement des informations utiles qui peuvent être utilisées pour développer et améliorer diverses entreprises. L'extraction d'informations utilise des puissants algorithmes et applications logicielles pour transformer les données non structurées des conversations sur les réseaux sociaux, des e-mails et des interactions avec les agents du service client en données accessibles, qui aident les entreprises à prendre de meilleures décisions [4]. Les techniques de la linguistique informatique sont utilisées pour identifier les entités nommées dans les textes, ainsi que les relations entre ces entités.

6. Limites et Défis

Plusieurs difficultés subsistent en NLP, dont on citera les plus fréquentes.

6.1. Sémantique

La sémantique, qui étudie le sens des mots et des phrases, est une zone de recherche importante en linguistique informatique, mais elle reste un défi important.

6.2. Ambiguïté

Par nature, le langage est ambigu. En raison du manque d'indices suffisamment significatifs et de connaissances contextuelles, la machine ne peut pas faire une interprétation correcte. Il s'agit d'un problème de longue date, malgré les progrès de l'apprentissage automatique et de grandes quantités de données [27]. On distingue 3 types d'ambiguïté qui sont :

- **Ambiguïté phonétique** : L'ambiguïté phonologique se produit lorsque la prononciation, l'intonation ou l'accentuation d'une phrase ou d'un mot peut conduire à des interprétations différentes. Pour le traitement du langage naturel, cela peut être difficile, car les modèles TALN doivent être capables de comprendre les inflexions de la parole afin d'interpréter correctement les phrases ou les mots. La technologie de reconnaissance vocale et les modèles formés aux données qui tiennent compte de ces changements peuvent aider à résoudre cette ambiguïté [27].
- **Ambiguïté syntaxique** : L'ambiguïté syntaxique est un problème pour le traitement automatique du langage naturel, car une même phrase peut être interprétée de différentes manières. Par exemple, la phrase "J'ai vu le chat avec des jumelles" peut être interprétée de deux manières différentes : soit le locuteur a vu le chat en utilisant des jumelles, soit le locuteur a vu quelqu'un d'autre qui avait des jumelles avec le chat. Les modèles de TALN utilisent l'analyse syntaxique et sémantique pour déterminer le sens le plus probable d'une phrase en fonction du contexte, mais l'ambiguïté syntaxique reste un défi important pour la TALN, car il y a souvent plusieurs interprétations possibles. Les modèles de TALN doivent donc être capables de prendre en compte le contexte et d'utiliser des informations supplémentaires pour déterminer le sens le plus probable d'une phrase [27].
- **Ambiguïté lexicale** : L'ambiguïté lexicale est un problème pour le traitement automatique du langage naturel, car une même phrase peut être interprétée de différentes manières. Par exemple, la phrase "J'ai vu le chat avec des jumelles" peut

être interprétée de deux manières différentes : soit le locuteur a vu le chat en utilisant des jumelles, soit le locuteur a vu quelqu'un d'autre qui avait des jumelles avec le chat. Les modèles de TALN utilisent l'analyse syntaxique et sémantique pour déterminer le sens le plus probable d'une phrase en fonction du contexte, mais l'ambiguïté syntaxique reste un défi important pour la TALN, car il y a souvent plusieurs interprétations possibles. Les modèles de TALN doivent donc être capables de prendre en compte le contexte et d'utiliser des informations supplémentaires pour déterminer le sens le plus probable d'une phrase [27].

6.3. Humour et Poésie

Les défis de l'humour et la poésie posés à la linguistique informatique sont liés à leur nature subjective, émotionnelle et nuancée. Les modèles de traitement du langage naturel peuvent avoir du mal à capturer les différentes couches de sens et les subtilités sous-jacentes à l'humour et à la poésie, ce qui les rend difficiles à modéliser automatiquement. De plus, les évaluations de ces modèles sont souvent subjectives et basées sur des réponses émotionnelles plutôt que sur des critères quantifiables. Pour relever ces défis, les chercheurs doivent développer des méthodes créatives et innovantes pour saisir la richesse et la complexité de cet aspect du langage [27].

6.4. Variabilité des langues

Les langues naturelles ont une grande variabilité, avec des différences régionales, culturelles et historiques. Les outils de linguistique informatique doivent être adaptés à cette variabilité pour être efficaces.

6.5. Données


Les modèles de linguistique informatique nécessitent de grandes quantités de données pour fonctionner correctement, d'où l'obtention de ces données peut être difficile et coûteuse. En outre, la source des données est très cruciale, où les textes bruités et informels dégradent les performances des modèles entraînés.

7. Conclusion

Ce chapitre présente un historique de l'intelligence artificielle et une introduction à la linguistique informatique, ainsi que quelques défis auxquels est confrontée la linguistique informatique.

En conclusion, le traitement du langage naturel est un domaine en pleine expansion qui continue de susciter l'intérêt des chercheurs et des entreprises. Les défis actuels dans le traitement du langage naturel comprennent la compréhension de la sémantique, la résolution de l'ambiguïté et la nécessité d'une grande quantité de données pour l'apprentissage automatique.

Cependant, grâce aux progrès rapides dans les domaines de l'IA et du TALN, ces défis seront probablement surmontés dans un proche avenir. En revanche, l'humour, les énoncés poétiques ou étranges, risquent de rester pour longtemps d'un accès difficile aux intelligences artificielles.



Chapitre II : Découverte de la langue bambara

1. Introduction

La langue bambara est l'une des langues africaines les plus importantes, vu le nombre de locuteurs et le nombre de pays auxquels elle est parlée. Dans ce chapitre, nous présentons la langue bambara et ses caractéristiques, ainsi nous parlons brièvement sur son utilisation dans les réseaux sociaux.

2. Généralités sur la langue et le dialecte

Le thème de la langue et du dialecte est un sujet complexe et souvent débattu en linguistique. Nous éclaircissons ces deux concepts dans cette section.

2.1. Langue

Une langue est un système de signes permettant l'intercompréhension à l'intérieur d'un groupe humain, ou encore un système de signes vocaux ou graphiques conventionnels, utilisé par un groupe d'individus pour l'expression du mental et la communication. Elle passe par la parole, le geste ou par d'autres types de signaux, et implique les mêmes processus pour tous les humains [28]. Toutefois, les langues parlées actuellement dans le monde sont très nombreuses, et il existe de grandes différences entre elles sur le plan des systèmes phoniques comme celui des structures grammaticales. On distingue plusieurs familles de langues telle que :

2.1.1. Langues indo-européennes

C'est un ensemble de langues d'Europe et d'Asie dont la parenté structurale s'est révélée remarquable. Ce qui signifie que la plupart des langues d'Europe et une grande partie des langues de l'Iran, de l'Afghanistan, du Pakistan, du Bangladesh et de l'Inde appartiennent à la famille indo-européenne [29]. La famille indo-européenne est divisée en plusieurs branches ou sous-familles. Elle comprend de nombreuses langues :

- Les langues d'origine latine dites romaines (comme le français, l'espagnol, le catalan ou l'italien) ;
- Les langues germaniques (comme l'allemand, l'anglais ou le suédois) ;
- Les langues slaves (comme le russe, le polonais ou le serbe) ;
- Les langues indo-iraniennes ;
- Le grec ;
- Les langues baltes ;
- Les langues celtiques.

Cette famille est celle qui a été la plus étudiée, c'est aussi celle dont on possède le plus de documents anciens, et celle dont on peut établir des liens génétiques absolument sûrs [28].

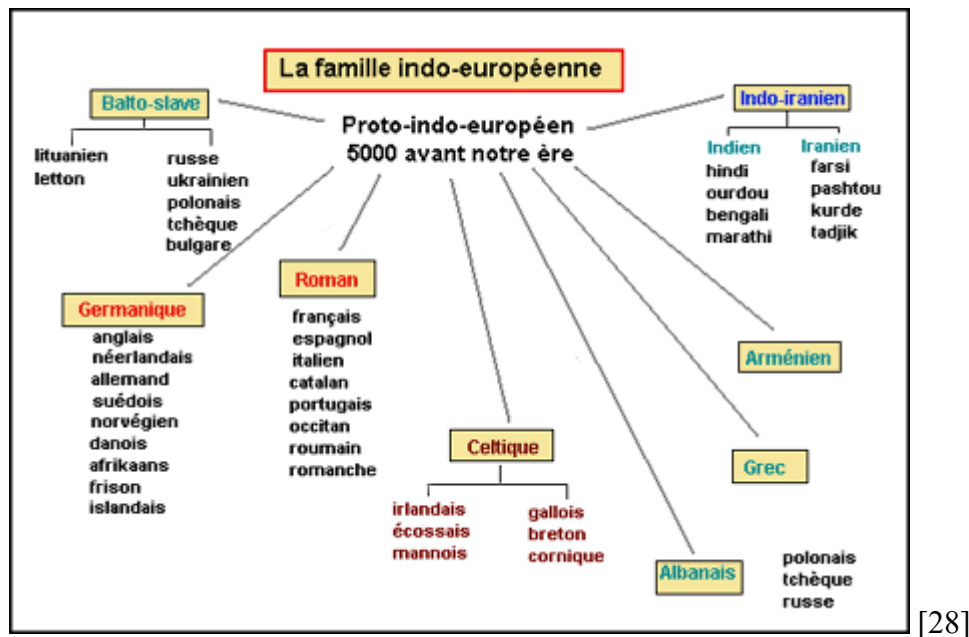


Figure II.1: Illustration de la famille Indo-Européenne

2.1.2.2 Langues mandées

Elles sont une branche de la famille des langues nigéro-congolaises, et a commencé à se développer comme l'une des principales langues de l'Afrique occidentale à l'époque de la création de l'empire du Mali (vers le XIII^e siècle) [30]. De nos jours, on peut estimer le nombre de ses locuteurs (natifs et non-natifs) à un chiffre situé entre 30 et 40 millions. Le mandingue est un continuum de dialectes recouvrant de vastes territoires au Mali, Guinée, Burkina-Faso, Côte d'Ivoire, Sénégal, Gambie, Guinée-Bissau, Liberia et en Sierra-Leone [30]. Les Mandingues constituent un ensemble culturel qui parlent 27 variantes dialectales classées en deux groupes :

- Les Mandingues occidentaux : Soninké, Mandingue du Sénégal oriental (Mandingo), Badibunke de Gambie, Pakawunke du Sénégal, Woyinké du Kaabu (Guinée-Bissau), etc.
- Les Mandingues orientaux : Bambaras et Khassonkés du Mali, Maninka ou Malinkés de Guinée et de Côte d'Ivoire.

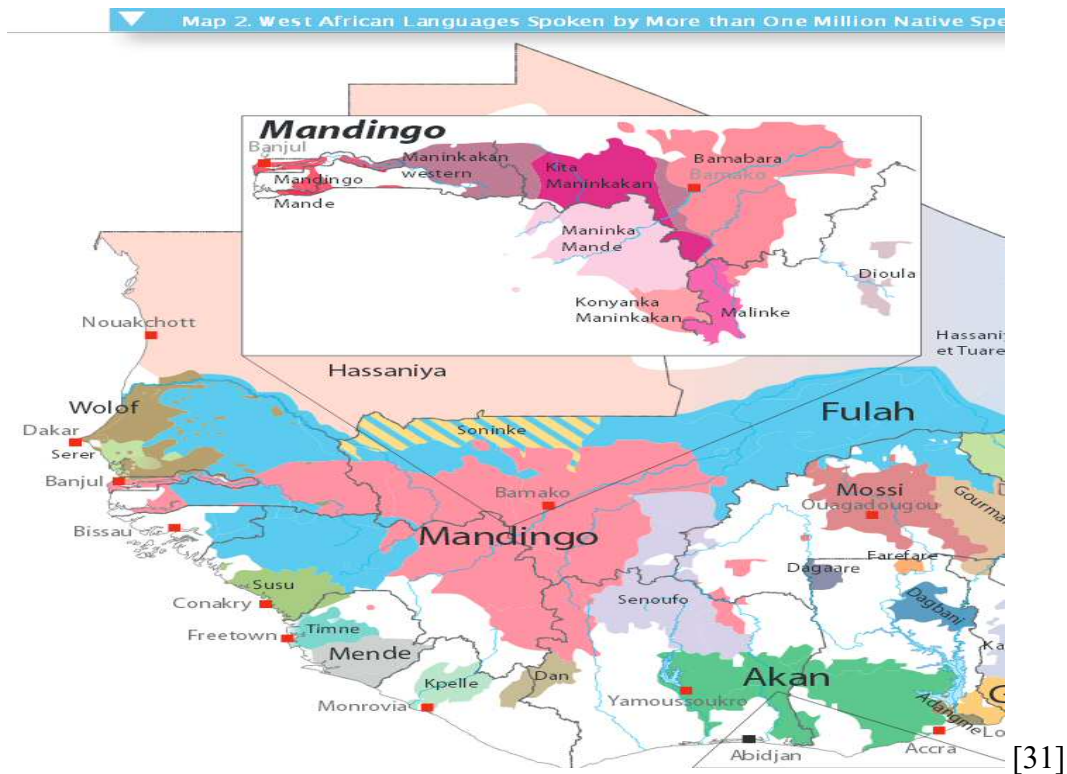


Figure II.2: Illustration graphique de quelques langues nigéro-congolaise

2.2.1. Dialecte

Un dialecte est le proche parent d'une langue dominante ou officielle, mais qui s'en distingue et qui, avec cette dernière, étaient autrefois variétés régionales l'une de l'autre. En d'autres termes, c'est la forme par laquelle une langue, parlée et surtout écrite se distingue d'une langue ancienne. Par exemple on peut prendre exemple sur le bambara, qui aujourd'hui est une langue, mais avant au temps des grands empires n'était qu'une variété (i.e. dialecte).

2.3.1. Différence entre langue et dialecte

Il est important de noter que, la différence entre une langue et un dialecte est souvent subjective et dépend de facteurs politiques et culturels [28]. Ainsi, certaines langues peuvent être considérées comme des dialectes à cause de la domination culturelle d'une autre langue, ou de la discrimination linguistique. De plus, il est possible que certains dialectes soient considérés comme des langues distinctes en raison de leur importance culturelle et historique [28]. Finalement, la distinction entre une langue et un dialecte est souvent floue et repose sur de nombreux facteurs. Cependant, il est important de comprendre que les langues et les dialectes sont tous deux des formes de communication qui évoluent et se transforment au fil du temps.

3. Langue Bambara

Dans cette section, nous présentons brièvement la langue bambara et ses caractéristiques, ainsi nous passons par sa morphologie et ses systèmes d'écriture.

3.1. Généralités

Les Bambaras aussi connus sous les appellations : bamanan, banmana, banmara ou bamananw pour le pluriel, sont un peuple mandingue de l'Afrique de l'Ouest sahélienne, établi principalement au Mali. Ils formaient le « Royaume bambara de Ségou » entre le XVIIIe et XIXe siècles [32]. Étant originaires du Mandé, ils auraient quitté cette région pour échapper à la domination des Malinkés¹, à l'époque de l'empire du Mali, d'où leur nom qui signifie « ceux qui ont refusé de se soumettre » de **ban** qui signifie « refus » et **mara** qui signifie « maître ». D'autre traduisent cette étymologie par « ceux qui ont refusé de se soumettre à l'islam ». Ils iront vivre au Ouassoulou, entre Sikasso et la Côte d'Ivoire, au Do et Kri (actuel Ségou), au Kaniaga (actuel : Bélé Dougou), au Kaarta (Sandaré, Oussoubidjadjan, etc.). Les Bambaras sont un sous-groupe des Malinkés, tout comme beaucoup d'autres groupes Mandingues [32].

Le bambara aussi appelé bamanankan est la variante normative la plus importante des langues mandingues, le terme « mandingue » est une déformation du mot **mandenka**, c'est-à-dire « habitant du Manden », le foyer historique du célèbre empire médiéval du Mali [30]. La langue mandingue représente un ensemble de dialectes ou langues prochement apparentés, parlées dans les pays de l'Afrique de l'Ouest. Les variétés mandingues que sont le bambara, le dioula (très proche du bambara et peut même être considéré comme une variante territoriale), le maninka (aussi plus ou moins proche du bambara) etc. Ils ont été vus comme des dialectes d'une seule langue, mais avec les indépendances, les variantes nationales du mandingue sont de plus souvent considérés comme langues.

Le bambara est la langue la plus importante du Mali avec près de 16 à 17 millions de locuteurs soit 82% de la population dont 5 à 6 millions sont des locuteurs natifs, elle est souvent considérée par les habitants comme « la langue malienne » par excellence. Le bambara existe sous différentes formes. Il s'agit d'abord du « bambara standard » dont la base est la variante utilisée à Bamako (cette forme se répand de plus en plus dans l'ensemble du Mali). Puis, ce sont les dialectes locaux, parfois fort différents du « bambara standard » [30].

¹ Un peuple d'Afrique de l'ouest

Le bambara est parlé principalement au Mali, mais également en Guinée, Burkina Faso, Côte d'Ivoire et dans d'autres pays. Il est considéré comme une langue véhiculaire importante dans la région, utilisée dans les affaires, l'éducation, la musique et les arts, ainsi que pour la communication entre différents groupes ethniques et linguistiques. La langue a également été influencée par l'arabe et le français en raison de la colonisation française en Afrique de l'ouest [30]. Bien que l'origine exacte de la langue bambara ne soit pas connue, elle a une longue histoire et une riche culture qui continuent d'être transmises à travers la langue.

3.2. Caractéristiques de la langue

Le Bambara a sept voyelles orales et sept voyelles nasales brèves, il est une langue tonale (le ton haut est marqué par un accent aigu, et le ton bas par un accent grave), l'accent tonique est important, car il permet de différencier les significations des mots qui ont des graphies similaires [33].

Le Bambara est une langue isolante, où les propriétés grammaticales sont exprimées surtout par des mots fonctionnels, plutôt que par des modifications de la forme des mots.

En ce qui concerne la structure de la phrase, la langue Bambara suit un ordre SVO (sujet-verbe-objet) dans les phrases déclaratives. Cependant, elle peut également utiliser un ordre OSV (objet-sujet-verbe) dans certaines constructions, notamment pour mettre en évidence l'objet de la phrase. Les verbes en Bambara sont conjugués pour indiquer le temps, l'aspect et le mode. Il existe également des particules verbales qui modifient le sens du verbe, telles que les particules de négation et d'affirmation [33].

3.3. Écriture en Bambara

Le système d'écriture pour le bambara utilise l'alphabet Latin avec quelques modifications pour représenter les sons uniques de la langue bambara depuis 1967. La liste de l'alphabet bambara ci-dessous sur la figure II.3 [34], avec ses 27 lettres.

A a	B b	C c	D d	E e	Ɛ ɛ	F f	G g
a	be	ce	de	e	ɛ	ef	ge
[a]	[b]	[tɕ]	[d]	[e]	[ɛ]	[f]	[g]
H h	I i	J j	K k	kh	L l	M m	N n
ha	i	je	ka		el	em	en
[h]	[i]	[dʒ]	[k]	[x]	[l]	[m]	[n]
-n	Ŋ ŋ	Ɔ ɔ	O o	Ɔ ɔ	P p	R r	S s
	ne	ɛŋ	o	ɔ	pe	er	es
[~n]	[ŋ]	[ɛŋ]	[o]	[ɔ]	[p]	[r]	[s]
sh	T t	U u	W w	Y y	Z z		
she	te	u	wa	je	ze		
[ɕ]	[t]	[u]	[w]	[j]	[z]		

Figure II.3: Alphabet Bambara

On souligne que la lettre “e” est considérée comme “é” comme dans le mot “époque”, Ɛ est considéré comme “è” comme le mot “prospère”, ŋ est considéré comme “gn” comme le mot “gagner”, ɛŋ est considéré comme “ng” comme le mot “mango” et enfin ɔ est considéré comme “o” comme le mot “porte”.

La langue bambara peut aussi être écrite avec un autre système d’écriture qui est le N’ko². C’est un système d’écriture utilisé pour écrire plusieurs langues africaines, dont le bambara. Il a été créé par Solomana Kanté, un érudit Guinée dans les années 1940, comme moyen de préserver les langues et cultures africaines [34]. Il comprend un alphabet de 20 lettres, dont 19 consonnes, 7 voyelles et une neutre, et est écrite de droite à gauche comme le cas de la langue arabe.

Bien que l’utilisation du bambara ne soit pas encore très répandue dans les médias ou dans la vie quotidienne, mais il est considéré comme un élément important de la préservation et de la promotion de la culture des langues africaines. Le N’ko est la traduction de l’expression « je dis » en bambara. De nos jours, plus de 500 mots ont été traduits et 183 livres réalisés en bambara [35].

² Script alphabétique pour écrire les langues Mandingues

(𞤀𞤁𞤂𞤃) Vowels						
𞤀	𞤁	𞤂	𞤃	𞤄	𞤅	𞤆
ɔ	o	u	ɛ	i	e	a
[ɔ]	[o]	[u]	[ɛ]	[i]	[e]	[a]
(𞤇𞤈𞤉𞤊) Consonants						
𞤇	𞤈	𞤉	𞤊	𞤋	𞤌	𞤍
ra	da	ca	ja	ta	pa	ba
[r]	[d]	[c]	[ʃ]	[t]	[p]	[b]
𞤎	𞤏	𞤐	𞤑	𞤒	𞤓	𞤔
ma	la	ka	fa	ga	sa	ra
[m]	[l]	[k]	[f]	[g/R/ʎ]	[s]	[r]
	𞤕	𞤖	𞤗	𞤘	𞤙	𞤚
	nasalizes	ya	wa	ha	na	nya
	vowels	[j]	[w]	[h]	[n]	[ŋ]
(𞤛𞤜𞤝) Tones						
	𞤛	𞤜	𞤝	𞤞		
	long	long	short	short		
	low	high	low	high		

Figure II.4 : Alphabet N’ko

3.1.4. Langue Bambara sur les réseaux sociaux

La langue Bambara est utilisée sur les réseaux sociaux, notamment sur Facebook, TikTok, Instagram et WhatsApp, principalement par des locuteurs du Mali et de la Côte d'Ivoire. Les utilisateurs de la langue Bambara sur les réseaux sociaux utilisent souvent des abréviations et des raccourcis pour écrire plus rapidement et la majorité des utilisateurs l'écrivent de leur manière, ce qui peut rendre la classification des textes plus difficile. En général, il est important de noter que les langues africaines ont souvent été sous-représentées sur les réseaux sociaux et sur Internet en général, en raison de la barrière linguistique et des ressources limitées pour la numérisation des langues africaines. Cependant, des initiatives sont en cours pour remédier à cette situation, et de plus en plus des contenus en langues africaines sont disponibles en ligne.

Par exemple, nous pouvons citer l'exemple de l'application Lenali³ du docteur en informatique, Mamadou Gouro Sidibé, qui a créé une solution pour permettre aux

³ <http://lenali.ml>

analphabètes d'utiliser les réseaux sociaux. L'application est conçue pour être utilisée vocalement et est accessible dans la langue locale.

Table 2.1: Exemple de quelques textes Bambara sur les réseaux sociaux

Réseaux sociaux	Bambara Standard	Traduction
A kougolo ma son daii	A kunkolo te son o ma	Il n'est pas d'accord
Niye bara kai ka gna	Ne ye o baara ke ka ne	J'ai bien fait le travail
Ne yie san bikonoton ni kononton den do yie	Ne ye san bi kononton ni kononton den ye	Je suis un enfant de l'année quatre-vingt-dix-neuf
A fo n ye où se trouve le djourou	A fo n ye juru be yoro min na	Dites-moi où se trouve la corde

Mamadou Gouro Sidibé veut combler la fracture numérique et les faibles taux d'alphabétisation dans les pays africains. Il a créé une application simple pour Ada, une femme de ménage malienne qui ne sait ni lire ni écrire, afin qu'elle puisse enregistrer sa voix et utiliser les fonctionnalités de Lenali, telles que les émissions de messagerie vocale, les appels gratuits, les notifications de service local, etc.

L'application Lenali est une solution de réseau social inclusive et accessible qui permet aux personnes analphabètes d'utiliser les réseaux sociaux pour des besoins pratiques et une utilité sociale. Nous montrons ci-dessous quelques exemples de textes en Bambara sur les réseaux sociaux.

4. Conclusion

Ce chapitre présente un aperçu entre la langue et le dialecte, ainsi qu'une brève présentation de la langue bambara.

Pour conclure, nous pouvons affirmer que le bambara est une langue importante en Afrique de l'ouest, qui a une histoire riche et une forte présence culturelle. En raison de l'utilisation croissante des médias sociaux dans la région, la langue bambara est de plus en plus utilisée sur ces plateformes.

Cependant, en comprenant les caractéristiques de la langue et de son utilisation sur les réseaux sociaux, il est possible de développer des outils de détection plus efficaces pour prévenir les futurs problèmes auxquels seront confrontés les utilisateurs en ligne.



Chapitre III : Présentation du langage offensif

1. Introduction

Ces dernières années, l'utilisation des réseaux sociaux a connu une croissance exponentielle partout dans le monde mettant en danger la sécurité et le bien-être des utilisateurs, avec de plus en plus de personnes qui utilisent des plateformes (e.g. Facebook, TikTok, Instagram et WhatsApp) pour communiquer avec leurs amis, leur famille et leur communauté. Le contenu offensant aussi est devenu de plus en plus populaire. Le langage offensif est un problème croissant sur les réseaux sociaux dans le monde, notamment dans les régions où le Bambara est utilisé pour harceler et critiquer les gens.

Dans ce chapitre, nous allons d'abord définir le langage offensif, puis examiner les différents types que nous pouvons rencontrer et leur impact sur les utilisateurs des réseaux sociaux. Ainsi, nous détaillons les techniques de détection de ce type de messages afin de les contrer.

2. Définition

Le langage offensif se réfère à tout type de discours qui peut être considéré comme insultant, discriminatoire, ou offensant envers un individu ou un groupe de personnes [36].

Le langage offensif sur les réseaux sociaux est particulièrement préoccupant, car il est souvent anonyme et peut causer de graves conséquences émotionnelles, psychologiques et même physiques. Les propos offensants sur les réseaux sociaux peuvent être particulièrement destructeurs. Comme pour le Bambara, les mots et les expressions ont souvent des significations plus profondes et symboliques que dans d'autres langues, ce qui peut donner lieu à des malentendus ou des interprétations erronées. Depuis l'avènement des réseaux sociaux, le langage a évolué, avec l'arrivée des emojis, des gifs et des memes dans nos échanges quotidiens. De plus, compte tenu du caractère informel de la communication en ligne, les utilisateurs peuvent être plus susceptibles d'utiliser un langage offensant que dans une conversation en face à face [37].

3. Différents types du langage offensif

Le langage offensif peut prendre diverses formes et types, dont on citera :

3.1. Violence verbale

L'abus verbal ou violence verbale est l'utilisation des mots pour causer du tort à la personne dont on parle, d'où il est difficile à définir et peut prendre de nombreuses formes. De même,

le préjudice causé est souvent difficile à mesurer (forme fréquente est l'injure). La violence verbale peut consister, entre autres formes de communication, à crier, insulter, intimider, menacer, humilier ou utiliser un langage désobligeant. Si la violence verbale est principalement perçue comme un blasphème oral envers une autre personne, elle peut aussi exister sous forme écrite. Sachant que le blasphème semble être le déclencheur le plus courant de la violence verbale [38].

3.2. Hate speech ou discours haineux

Par « discours haineux », on entend habituellement des propos discriminatoires à l'encontre de personnes ou de groupes pour des motifs comme l'appartenance ethnique ou culturelle, l'origine, la nationalité, la religion, le sexe, l'orientation sexuelle ou le handicap. Cependant, le discours de haine englobe aussi des expressions non verbales, comme celles véhiculées par des images, des vidéos ou toute forme de communication en ligne et hors ligne [39].

Selon ANTONIO GUTERRES, le secrétaire général des nations unies, 2021 « *Les médias sociaux servent de porte-voix pour diffuser la haine mondialement* ». Dans leur stratégie et plan d'action pour la lutte contre les discours haineux, il est mentionné que la multiplication des contenus haineux en ligne est un défi énorme pour nos sociétés, car ces contenus peuvent être produits et partagés facilement, atteignant un grand nombre de public dans le monde entier. Les gouvernements ont du mal à faire respecter les lois nationales dans le monde virtuel, et les entreprises en ligne ont généralement des limites technologiques dans la surveillance automatisée des contenus. Les algorithmes des entreprises ont contribué à l'instrumentalisation grandissante des médias sociaux pour répandre des récits haineux et conflictuels, stigmatisant les communautés vulnérables et menaçant la démocratie [40].

3.3. Langage abusif

C'est un type de langage offensif qui vise à blesser, humilier ou dégrader une personne ou un groupe de personnes. Il peut être utilisé pour intimider, contrôler, manipuler ou dominer une autre personne. Le langage abusif peut prendre différentes formes, telles que [41] :

- Les insultes : Les propos offensants peuvent comprendre l'utilisation d'insultes, de blasphèmes ou de noms désobligeants pour dénigrer une personne ou un groupe de personnes [41] ;

- La critique constante : Il peut s'agir de critiquer constamment le comportement, l'apparence ou les décisions d'une personne, de la faire se sentir inadéquate ou sans valeur [41] ;
- La menace : Il peut s'agir de menaces verbales ou non verbales d'intimidation ou de contrôle d'une autre personne [41] ;
- La mise en doute de la valeur personnelle : Cela peut inclure le fait de dénigrer la valeur ou l'estime de soi d'une personne en la faisant se sentir indigne ou en sous-estimant son importance [41] ;
- Cyber-harcèlement : Cela peut inclure la diffusion de rumeurs ou de fausses informations dans le but de discréditer ou harceler une personne en ligne [41].

3.4. Langage discriminatoire

Cela peut englober la stigmatisation d'une personne ou d'un groupe de personnes en raison de sa race, de son orientation sexuelle, de sa religion ou de tout autre aspect de son identité. Il peut causer des dommages émotionnels, psychologiques et physiques aux personnes qui en sont victimes. Il peut être appelé de plusieurs manières différentes. Voici quelques termes couramment utilisés pour décrire ce type de langage [42]:

- Langage raciste : se réfère à des propos ou comportements qui discriminent une personne ou un groupe en raison de leur race ou de leur origine ethnique.
- Langage sexiste : désigne les propos ou comportements qui discriminent une personne ou un groupe en raison de leur sexe, leur genre ou leur orientation sexuelle.
- Langage xénophobe : se réfère à des propos ou comportements qui discriminent une personne ou un groupe en raison de leur nationalité, de leur culture ou de leur religion.

4. Conséquences du langage offensif

Le langage offensif peut avoir des conséquences graves sur les individus, les groupes et la société dans son ensemble. Voici quelques exemples de conséquences du langage offensif :

4.1. Effets psychologiques

La dépression, l'anxiété, le stress, le manque de confiance en soi, la colère et la frustration. Les personnes victimes du langage offensif peuvent également ressentir de la honte, de la culpabilité et de l'isolement social.

- La dépression : c'est une maladie mentale qui peut toucher tout le monde. En 2019, 280 millions de personnes présentaient une dépression, dont 23 millions d'enfants et d'adolescents [43]. Au cours de la première année de la pandémie de COVID-19, la prévalence mondiale de l'anxiété et de la dépression a augmenté massivement, de 25 %, selon une note scientifique publiée par l'Organisation mondiale de la Santé (OMS). La dépression diffère des sautes d'humeur habituelles et des réactions émotionnelles passagères face aux problèmes du quotidien [43]. Lors d'un épisode dépressif, l'individu présente une humeur morose (sentiment de tristesse, d'irritabilité, de vide) ou une perte de plaisir ou d'intérêt pour les activités, pendant la majeure partie de la journée, presque tous les jours, pendant au moins deux semaines [43]
- L'anxiété : C'est un trouble émotionnel qui se manifeste par un sentiment d'insécurité. L'anxiété est caractérisée par un sentiment d'appréhension, de tension, de malaise, de terreur face à un péril de nature indéterminée. Elle est souvent exprimée par le patient sous les termes de nervosité ou de soucis. L'anxiété doit être distinguée de la peur qui répond à une situation menaçante réelle [44]. On différencie :
 - Le trouble panique correspondant à des crises d'angoisse aiguës survenant de manière périodique ;
 - L'anxiété permanente, l'anxiété généralisée ;
 - La névrose d'angoisse affectant les sujets de manière durable [44].

4.2. Effets sociaux

Le langage offensif peut avoir des effets sociaux négatifs tels que la stigmatisation, la discrimination, la marginalisation et l'exclusion. Les victimes peuvent se sentir ostracisées par la société et subir des préjudices économiques et éducatifs [42].

4.3. Effets physiques

Dans certains cas, le langage offensif peut conduire à des violences physiques et à des comportements agressifs. Les menaces et les insultes peuvent encourager la violence et les comportements haineux envers les victimes [42].

4.4. Effets sur la communauté

Le langage offensif peut également avoir des effets négatifs sur la communauté dans son ensemble. Il peut renforcer les préjugés, les stéréotypes et les discriminations et conduire à la

polarisation et à la division sociale [41]. Voici quelques cas réels de violence qui ont été attisés et repoussés par des propos haineux à l'endroit de groupes et de minorités ciblés ;

- L'attaque contre la mosquée de Québec en 2017, où six personnes ont été tuées et de nombreuses autres blessées par un tireur qui avait publié des commentaires haineux contre les musulmans sur les réseaux sociaux.
- Les violences contre les personnes d'origine asiatique aux États-Unis pendant la pandémie de COVID-19, où des langages offensifs utilisés pour lier la pandémie à la Chine ont alimenté des actes de violence et de discrimination contre les Asiatiques-Américains.
- Les violences contre les migrants et les réfugiés, où des langages offensifs et déshumanisants ont été utilisés pour justifier la discrimination, la détention et même la torture et la mort.
- Les actes de terrorisme perpétrés par des groupes extrémistes, où des discours haineux sont utilisés pour diaboliser des groupes entiers et justifier des violences.

5. Détection automatique du langage offensif

La détection des contenus offensants est une technique visant à identifier et signaler le contenu textuel contenant des mots offensants, abusifs, discriminatoires ou violents.

La détection du texte offensant peut être effectuée manuellement par des modérateurs humains ou automatiquement par des algorithmes de traitement du langage naturel (TALN). Les modérateurs humains lisent les messages ou les commentaires publiés par les utilisateurs et décident s'ils contiennent du matériel répréhensible ou inapproprié. Cette méthode est généralement utilisée dans les contenus sensibles ou les petites et moyennes plateformes.

La détection automatique du langage offensant est un domaine de recherche en plein croissance visant à développer des outils capables de détecter et d'analyser des textes contenant un langage abusif, haineux ou discriminatoire. Pour détecter automatiquement un langage offensant dans un texte, plusieurs méthodes peuvent être utilisées pour résoudre la tâche de détection du langage offensant. L'apprentissage automatique et l'apprentissage profond sont les méthodes les plus investies par la communauté de recherche dans ce domaine. Une approche hybride peut également être utilisée, laquelle combine les deux techniques précédentes pour améliorer la précision de détection [41].

5.1. Méthodes à base d'apprentissage ou Machine Learning

Pour la tâche de détection du langage offensif, l'apprentissage automatique est utilisé en entraînant un modèle de classification avec des données déjà classifiées. Dans la suite, nous citons quelques algorithmes d'apprentissage automatique.

5.1.1. Machine à vecteurs de support

Une machine à vecteurs de support ou Support Vector Machine (SVM) est un type d'algorithmes d'apprentissage automatique utilisé pour la classification et l'analyse de régression. Les SVM fonctionnent en trouvant l'hyperplan qui sépare le mieux les données en différentes classes (ou catégories). Les SVM sont de plus en plus populaires dans la Machine Learning et bien d'autres domaines. Il existe différents types de SVM qui sont : Linéaire SVM ou non-Linéaire SVM [41].

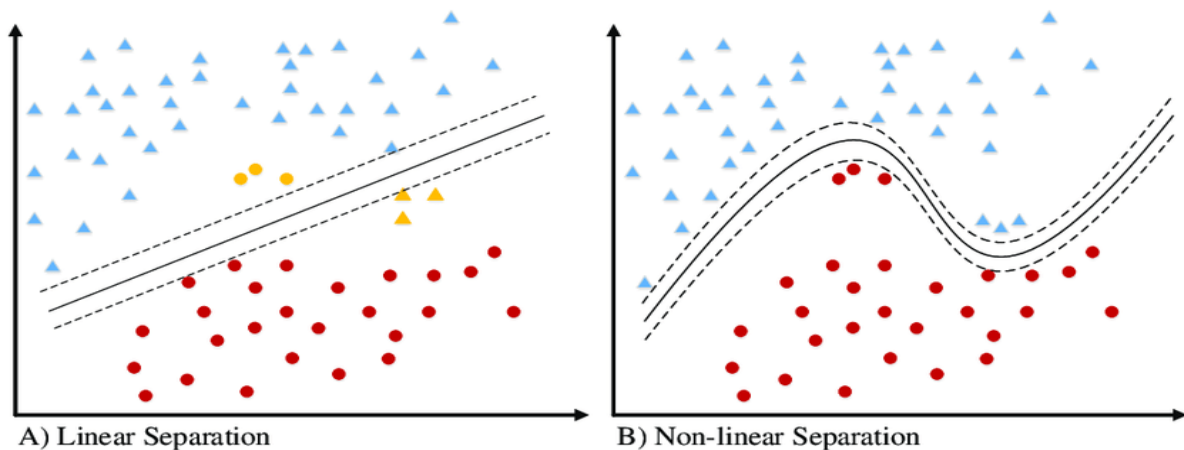


Figure III.1 : Illustration du Support Vector Machine (SVM) linéaire et non linéaire [45]

5.1.2. Arbre de décision

Les arbres de décision (AD) sont une catégorie d'arbres utilisée dans l'exploration de données et en informatique décisionnelle. Ils emploient une représentation hiérarchique de la structure des données sous forme des séquences de décisions (tests) en vue de la prédiction d'un résultat ou d'une classe. Chaque individu (ou observation), qui doit être attribué(e) à une classe, est décrit(e) par un ensemble de variables qui sont testées dans les nœuds de l'arbre. Les tests s'effectuent dans les nœuds internes et les décisions sont prise dans les nœuds feuille [46].

Le classificateur de l'arbre décisionnel est l'une des méthodes les plus connues pour classer les données. Sa caractéristique la plus importante est sa capacité de transformer les problèmes de prise de décision compliqués en processus simples sous la forme graphique d'un arbre. Les

différentes décisions possibles sont situées aux extrémités des branches (les feuilles de l'arbre), et sont atteintes en fonction de décisions prises dans chaque étape, trouvant ainsi une solution compréhensible et plus facile à interpréter. L'arbre de décision est un algorithme d'aide à la décision représentant un ensemble de choix [46].

Exemple de problème adapté à un approche par arbres de décision : comment répartir une population d'individus (e.g. clients, produits, utilisateurs, etc.) en groupes homogènes selon un ensemble de variables descriptives (e.g. âge, temps passé sur un site Web, etc.) et en fonction d'un objectif fixé (variable de sortie ; par exemple : chiffre d'affaires, probabilité de cliquer sur une publicité, etc.) [46].

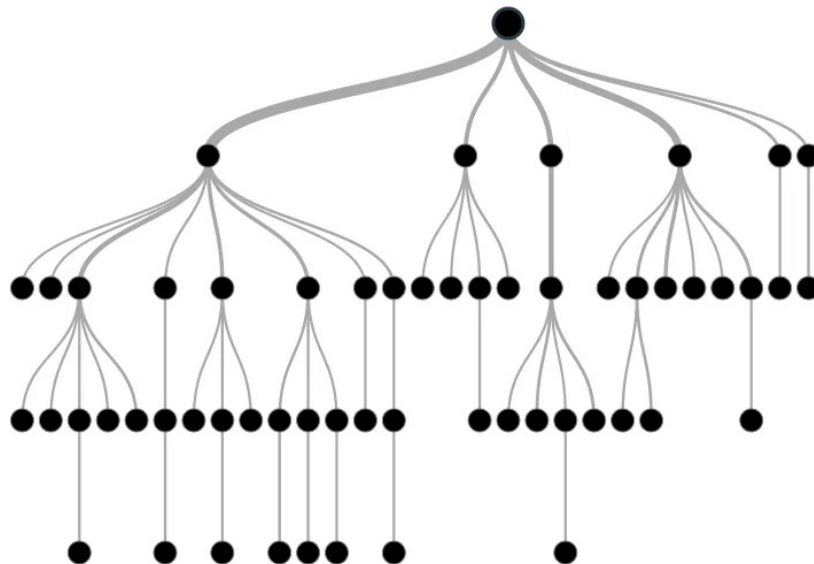


Figure III.2 : Illustration de l'arbre de décision

5.1.3. Naive Bayes

Le classificateur Naive Bayes (ou réseaux bayésien naïf) est un algorithme d'apprentissage automatique qui simplifie considérablement l'apprentissage en supposant que les fonctionnalités appartiennent à la classe donnée. Le classifieur bayésien est un type de classification probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses, appartenant à la famille des classifieurs linéaires. Il suppose que l'existence d'une caractéristique pour une classe est indépendante de l'existence d'autres caractéristiques. L'avantage du classifieur bayésien naïf est qu'il requiert relativement peu de données d'entraînement pour estimer les paramètres nécessaires à la classification, à savoir moyennes et variances des différentes variables [47].

5.1.4. K-voisins les plus proches

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 (i.e. classe connue a priori). Il est basé sur le concept dont les points de données d'une classe particulière sont voisins les uns des autres [48]. Pour un échantillon de test inconnu, trouver la classe dont elle est voisine est basé sur la mesure des distances euclidiennes de la donnée inconnue de tous les points de données de toutes les classes dans le corpus d'apprentissage. Ensuite, sur les k distances les plus proches, où k est un nombre supérieur ou égal à 1, la classe à laquelle les données inconnues sont le plus proche le plus souvent est la classe attribuée aux données de test ou aux données inconnues. [49]

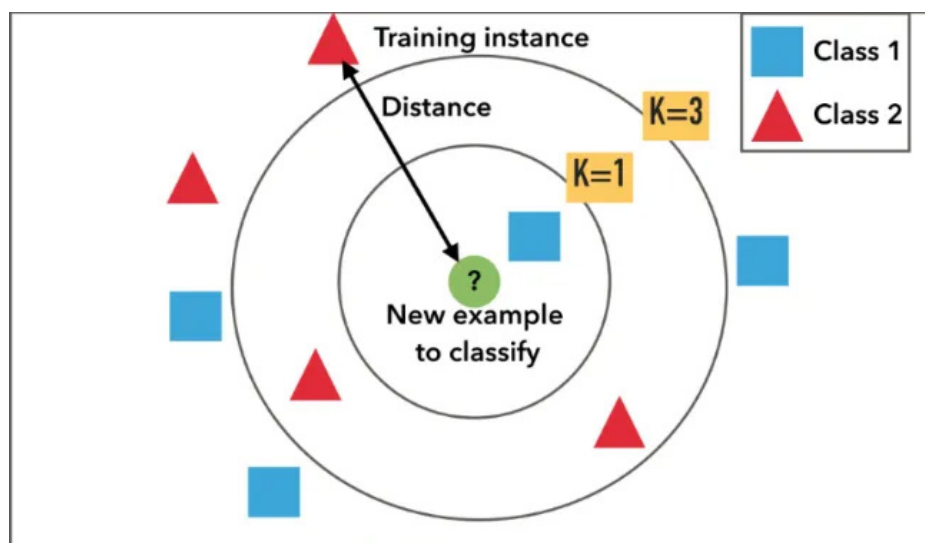


Figure III.3 : Illustration d'un exemple de l'algorithme des k-voisins les plus proches (KNN)

5.1.5. Réseaux de neurones

Un réseau de neurones ou Artificial Neural Network en anglais (abrév. ANN) est un modèle de calcul dont la conception est très schématiquement inspirée du fonctionnement de vrais neurones (humains ou non). Généralement ils sont composés de couches de nœuds interconnectés qui traitent et transmettent les informations. 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, tandis que le dernier renvoie la classe (ou catégorie) d'appartenance. L'algorithme ANN accepte uniquement les données numériques et structurées [50].

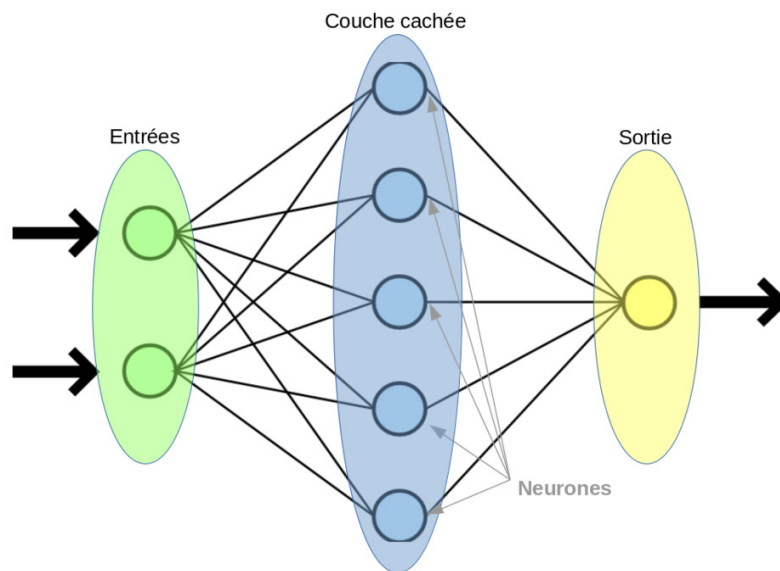


Figure III.4 : Illustration d'un réseau de neurones

5.2. Apprentissage profond

Le 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. Ci-dessous nous citons quelques algorithmes d'apprentissage profond :

5.2.1. Réseaux de neurones convolutifs

Un réseau de neurone convolutif en anglais Convolutional Neural Networks (CNN) est un réseau de neurone à plusieurs couches qui tire son inspiration du cortex visuel animal. Ce type d'architecture s'avère particulièrement utile dans le traitement d'images. Le premier CNN a été

créé par LeCun avec une orientation spécifique à la reconnaissance de caractères manuscrits, tels que l'interprétation des codes postaux. En tant que réseau profond, les premières couches de Yann LeCun reconnaissent les caractéristiques de base, comme les bords, tandis que les couches ultérieures combinent ces caractéristiques pour obtenir des attributs de plus haut niveau à partir de l'entrée.

L'architecture du CNN comprend plusieurs couches qui permettent l'extraction des caractéristiques et la classification (voir l'image suivante). L'image est divisée en zones appelées "champs réceptifs", qui sont ensuite utilisées par une couche convolutive pour extraire les caractéristiques de l'image d'entrée. Ensuite, vient l'étape de mise en commun, qui réduit la dimensionnalité des caractéristiques extraites en utilisant un échantillonnage réduit tout en préservant les informations les plus importantes, généralement à l'aide d'une mise en commun maximale. Cette opération de convolution et de mise en commun est répétée plusieurs fois, alimentant finalement un perceptron multicouche entièrement connecté. La couche de sortie finale de ce réseau est composée de plusieurs nœuds qui identifient les caractéristiques de l'image, par exemple, un nœud par numéro identifié. L'entraînement du réseau se fait en utilisant la rétropropagation, une méthode d'optimisation couramment employée [51].

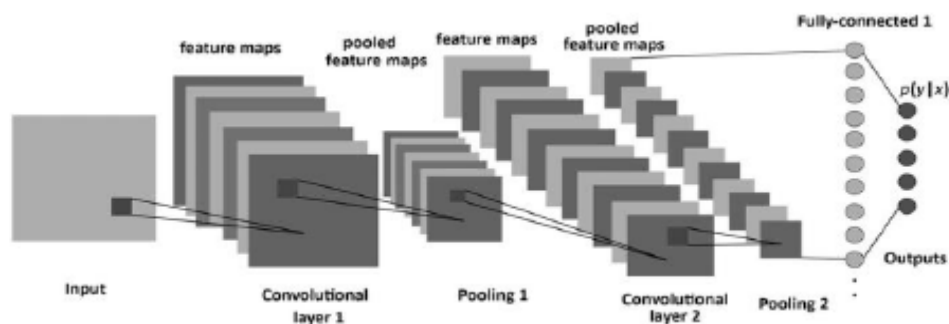


Figure 5: Architecture classique d'un réseau de neurones convolutif [51].

5.2.2. Réseaux de neurones récurrents

Les Réseaux neuronaux récurrents ou RNN (Recurrent Neural Networks) sont des architectures fondamentales sur lesquelles sont basées les autres architectures d'apprentissage profond. La principale différence entre un réseau multicouche classique et un réseau récurrent réside dans le fait qu'un réseau récurrent possède des connexions qui rétroagissent vers les couches précédentes (ou vers la même couche), au lieu d'avoir uniquement des connexions dirigées vers l'avant. Cette rétroaction permet aux RNN de conserver en mémoire les entrées passées et de modéliser les problèmes dans le temps. Un réseau neuronal récurrent est un type

de réseau de neurones artificiels qui présente des connexions récurrentes. Les RNN sont spécifiquement conçus pour traiter des données séquentielles, telles que la parole ou le texte. Ils sont dotés d'une boucle de rétroaction qui permet de transmettre les informations d'une étape de la séquence à la suivante, ce qui permet au réseau de conserver une mémoire des entrées précédentes [51].

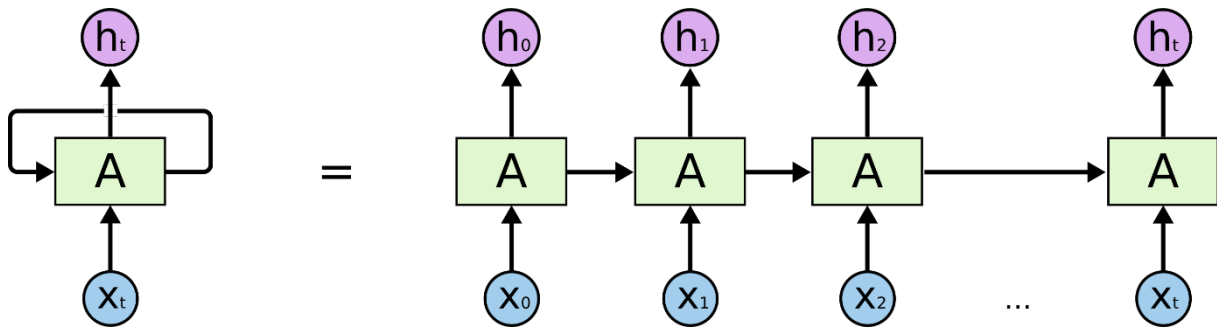


Figure III.6: Illustration d'un réseau neuronal récurrent déroulé [52]

5.2.3. Réseaux de neurones Long Short Terme Memory

Le réseau mémoire à long et court terme généralement appelé LSTM (Long Short-Term Memory) est un type particulier de RNN, capable d'apprendre les dépendances à long terme. Il a été créé en 1997 par Hochreiter et Schmidhuber, mais il a gagné en popularité ces dernières années en tant qu'architecture RNN pour diverses applications. Le LSTM s'est écarté des architectures de réseau neuronal typiques basées sur les neurones et a introduit le concept de cellule mémoire. La cellule mémoire peut conserver sa valeur pendant une période plus ou moins longue en fonction de ses entrées, ce qui permet à la cellule de se souvenir de ce qui est important et pas seulement de sa dernière valeur calculée. La cellule mémoire LSTM contient trois portes qui contrôlent la manière dont les informations entrent ou sortent de la cellule :

- La porte d'entrée contrôle quand une nouvelle information peut entrer dans la mémoire.
- La porte d'oubli contrôle l'oubli d'un élément d'information existant, permettant à la cellule de se souvenir de nouvelles données.
- Enfin, la porte de sortie contrôle quand l'information contenue dans la cellule est utilisée dans la sortie de la cellule [51].

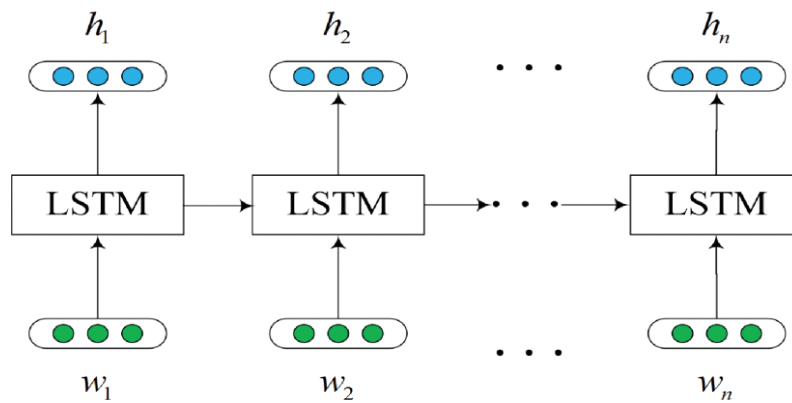


Figure III.7: Architecture de base d'un LSTM [53]

5.2.4. Réseaux de neurones bidirectionnels

Un réseau de neurones bidirectionnel est un type de réseau qui peut traiter les données de manière séquentielle dans les deux sens. Cela permet au réseau de capturer des informations à partir d'entrées passées et futures, ce qui contribue à rendre utiles des tâches telles que la reconnaissance de la parole et du langage naturel. Les réseaux de neurones bidirectionnels ont été utilisés dans diverses applications telles que le traitement de la parole. En BiLSTM, les neurones comportent plusieurs mécanismes internes tels qu'une cellule mémoire, une porte d'entrée, une porte de sortie et une porte d'oubli. Cela permet 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 » [4].

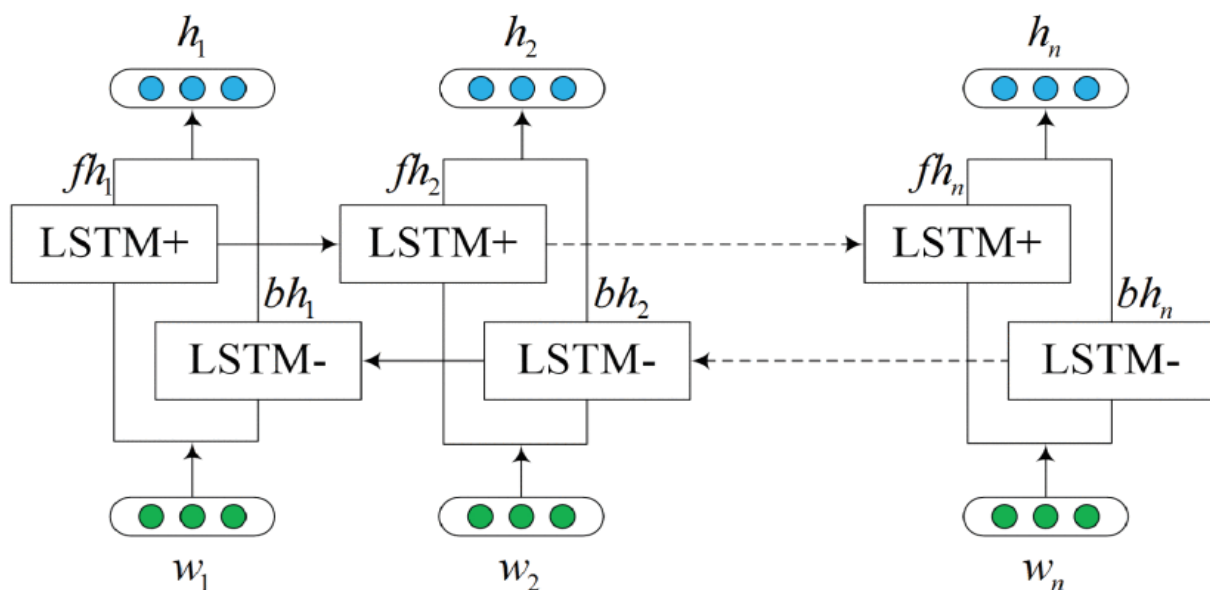


Figure III.8 : Architecture de base d'un BiLSTM [53]

5.3. Avantages et inconvénients des algorithmes d'apprentissage

Nous résumons les avantages et les inconvénients liés aux différents modèles par le tableau récapitulatif suivant.

Table 2.1: Tableau récapitulatif résumant les avantages et les inconvénients des algorithmes à base d'apprentissage

TECHNIQUES	AVANTAGES	INCONVÉNIENTS
SVM	<ul style="list-style-type: none"> - Grande précision de prédiction - Fonctionne bien sur de petits corpus 	<ul style="list-style-type: none"> - Ne convient pas à des jeux de données plus volumineux - Moins efficace sur les jeux de données contenant du bruit
Arbre de décision	<ul style="list-style-type: none"> - Facile à comprendre et à visualiser - Bonne performance sur des ensembles de données de petite à moyenne taille 	<ul style="list-style-type: none"> - Peut être sensible à l'overfitting (Surapprentissage) - Peuvent devenir très complexes et difficiles à interpréter
RESEAUX DE NEURONES	<ul style="list-style-type: none"> - Peuvent être utilisés pour une variété de problèmes de classification et de régression - Bonne performance sur des ensembles de données de grande taille - Peut apprendre des représentations de caractéristiques complexes 	<ul style="list-style-type: none"> - Besoin de beaucoup de données pour l'entraînement - Peut être sensible à la sélection de l'architecture et des hyperparamètres - Les temps d'entraînement peuvent être très longs sur des ensembles de données très larges
KNN	<ul style="list-style-type: none"> - Facile à comprendre et à mettre en œuvre - Bonne performance sur des ensembles de données de petite à moyenne taille 	<ul style="list-style-type: none"> - Temps d'exécution élevé - Peut être sensible à la sélection de la valeur K
CNN	<ul style="list-style-type: none"> - Bonne performance sur les données d'image et de vision par ordinateur - Peut apprendre des caractéristiques à différents niveaux d'abstraction - Peut gérer des images de grande taille 	<ul style="list-style-type: none"> - Besoin de beaucoup de données pour l'entraînement - Temps d'entraînement peuvent être longs sur des ensembles de données très larges
RNN	<ul style="list-style-type: none"> - Convient aux données séquentielles - Peut prendre en compte les relations temporelles entre les données - Peut gérer des entrées de longue séquence 	<ul style="list-style-type: none"> - Les temps d'entraînement peuvent être longs - Peut souffrir de problèmes de gradient qui rendent l'entraînement difficile
LSTM	<ul style="list-style-type: none"> - Peut prendre en compte les relations temporelles - Peut gérer des entrées de longue séquence 	<ul style="list-style-type: none"> - Les temps d'entraînement peuvent être longs - Peut souffrir de problèmes de gradient

	- Peut gérer des dépendances à long terme	qui rendent l'entraînement difficile
BILSTM	- Peut prendre en compte les relations temporelles - Peut gérer des entrées de longue séquence - Peut gérer des dépendances à long terme entre les données	- Les temps d'entraînement peuvent être longs - Peut souffrir de problèmes de gradient qui rendent l'entraînement difficile - Besoin de beaucoup de données pour l'entraînement
NB	- Relativement simple à comprendre et à construire. - Facile à former	- Implique que chaque fonctionnalité soit indépendante

6. Conclusion

Dans ce chapitre, nous avons abordé les différents types de langage offensif, ainsi que les conséquences qu'ils peuvent avoir sur les individus et la société en général. Nous avons également discuté l'importance de la détection automatique des contenus offensifs, ainsi que les différentes méthodes d'apprentissage automatique qui peuvent être utilisées à cette fin.

Les techniques d'apprentissage automatique peuvent être supervisées ou non supervisées, et chaque approche présente ses avantages et ses inconvénients. Parmi les algorithmes d'apprentissage automatique que nous avons présentés, nous avons remarqué que l'algorithme Bayésien est simple et rapide à implémenter, mais peut être limité par sa faible précision. Les algorithmes de classification non linéaire, tels que les machines à vecteurs de support (SVM), sont plus précis, mais peuvent être plus complexes à mettre en place.

En résumé, détecter le contenu offensif sur les réseaux sociaux est un sujet clé pour garantir la sécurité des utilisateurs et leur bien-être. Les techniques d'apprentissage automatique peuvent détecter efficacement ces contenus, mais leur précision et leur efficacité dépendent des caractéristiques de l'algorithme utilisé. Par conséquent, il est important de choisir la méthode la plus appropriée en fonction des besoins du projet.

Enfin, la détection du contenu offensif sur les réseaux sociaux est un thème complexe et important, d'où des approches multidisciplinaires sont nécessaires.



Chapitre IV : Méthodologie et Résultats

1. Introduction

Dans ce chapitre, nous présentons en détail la méthodologie que nous avons utilisé pour réaliser ce travail. Pour mieux comprendre notre méthodologie, nous avons commencé par la collecte des données, suivie d'une étape de prétraitement pour préparer les données. Ensuite, nous avons essayé des techniques de vectorisation, ainsi que l'utilisation des différentes approches d'apprentissage automatique. Pour terminer nous évaluons les performances de nos modèles de détection automatique en utilisant des mesures telles que la précision, le rappel et le score F1.

2. Travaux connexes

Dans cette section, nous exposons quelques travaux de recherches effectués dans la langue Bambara, notamment ceux concernant le traitement automatique de la langue.

Diallo et al. (2021), dans leur étude présentaient le premier ensemble de données dialectales bambara basé sur *Common Crawl*⁴, qui est dédié à l'analyse des sentiments. Les auteurs ont présenté deux versions de l'ensemble de données bambara (Bambara V1 et Bambara V2). Ils ont utilisé des modèles d'apprentissage automatique tels que SVM et LSTM pour analyser les sentiments dans ces phrases. Les résultats ont montré que le SVM a donné de meilleurs résultats que les modèles de deep learning utilisés. Les auteurs ont partagé leur base de données pour des recherches supplémentaires et ont prévu d'augmenter la taille de la base de données plus tard.

Une étude sur la traduction par apprentissage automatique neurale pour les langues africaines à ressources extrêmement limitées a été réalisée, avec une étude de cas sur le bambara (Tapo et al., 2020). Les résultats ont montré que les langues à ressources limitées présentent des défis uniques pour la traduction par apprentissage automatique. Les ressources semblables disponibles pour la traduction du bambara vers le français ou l'anglais se sont avérées peu nombreuses et donc peu pratiques à utiliser comme sources de données d'entraînement. Le problème résidait dans le manque de conformité à l'orthographe bambara standardisée, en raison de la prédominance de la langue orale.

Majumder et al. (2021), dans leur article "Cross-Lingual Offensive Language Identification in Low-Resource Settings" traitent le défi de la détection des contenus offensifs

⁴ Méthode de collecte et annotation

dans les langues à faibles ressources en utilisant des techniques de transfert de connaissances. Ils proposent une approche qui exploite les modèles pré-entraînés dans des langues mieux représentées pour améliorer les performances dans les langues à faibles ressources.

Fayssal Zerargui et Omar Benzaoui (2021), dans leur mémoire intitulé « Le discours de haine sur le web et les médias sociaux » abordent 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. Ils ont regroupé les commentaires en cinq catégories distinctes. Le taux F1 montre que LSTM a obtenu de meilleurs résultats par rapport au CNN.

3. Base de données

Nous commençons par présenter la source de nos informations que nous avons utilisées pour constituer notre base de données. Nous détaillons également les différentes manières utilisées pour collecter les données, et par la suite finir par l'annotation de celles-ci.

3.1. Source d'informations

Pour cette tâche, nous disposons d'une base de données de contenu bambara issues des réseaux sociaux et d'une ancienne base de données pour l'analyse des sentiments en bambara que nous avons adaptés pour notre travail.

Nous avons pris les réseaux sociaux comme source pour collecter nos données, car les données collectées là-bas offrent largement des contenus qui représentent le bambara utilisé dans les conversations en ligne, ce qui est mieux que ce qu'on pourra trouver dans les livres, car très peu utilisés sur les réseaux sociaux.

Les données collectées sur les médias sociaux sont issues des profils ou des pages sur *Facebook*⁵ et *TikTok*⁶. Nous avons ciblé spécialement ces deux réseaux sociaux, car ce sont les plus utilisés par la communauté bambaraphones. Les commentaires collectés sont issus de plusieurs personnes de différents genres, âges, religions ou ethnies de partout dans le monde comprenant le bambara.

⁵ <https://www.facebook.com>

⁶ <https://www.tiktok.com>

3.2. Collecte des données

La collecte peut se faire de deux manières soit manuellement ou automatiquement. Nous avons utilisé les deux manières pour essayer d'avoir le maximum de données utilisables. Dans un premier temps (automatiquement), il existe plusieurs techniques possibles comme : l'utilisation des d'APIs, l'utilisation de scripts (le web scraping) et de robots (le web crawling) etc. L'extraction de données web ou scraping, est la méthode qui consiste à extraire des informations spécifiques à partir des sites web de manière automatisée en utilisant des programmes ou des outils informatique qui parcourent les pages Web, analysent leur contenu désiré et les extrait.

Dans notre cas, on a utilisé un script pour nous permettre de collecter facilement les commentaires des publications en bambara qui nous intéressait. Cela nous a permis de recueillir un ensemble diversifié de contenus provenant de sources variées.

Pour la collecte manuelle la technique est très simple, il suffit de collecter les données par ses propres moyens, ce qui peut être très difficile et fatigant. En ajout de la collecte automatique, nous avons effectué une collecte manuelle de données supplémentaires. Nous avons ciblé des publications spécifiques sur les réseaux sociaux ainsi que des forums susceptibles de contenir des contenus offensants ou abusifs, mais nous avons aussi ciblé des grins (groupes de jeunes qui se réunisse généralement autour d'un thé) pour mieux accroitre notre chance de capturer un maximum de données offensantes. Cette approche nous a permis d'inclure des expressions et des insultes couramment utilisées en bambara, qui pourraient échapper à la collecte automatique.

3.3. Annotation

Après notre collecte de données, nous avons procédé à l'annotation de celles-ci en trois différentes catégories comme mentionné ci-dessous. Pour cela, nous avons formé une équipe d'annotateurs natifs en bambara, qui ont suivi les consignes précises pour étiqueter les textes comme étant normal par (0), offensant par (1) ou abusif par (2). Les annotateurs étaient aux nombres de trois (3) tous des hommes, âgés de 20 à 25 ans à qui nous avons donné une liste de 1,026 textes de commentaires à annoter chacun.

Table 4.1. Nombre d'agrément des trois annotateurs

	# textes
Offensif	429
Abusif	280
Total	709

Cette étape d'annotation est décisive pour former nos classificateurs après. Cette annotation manuelle nous a permis de constituer un jeu de données étiqueté pour l'entraînement de nos classificateurs.

4. Prétraitement des textes Bambara

Les données collectées sont généralement surchargées d'autres données qui ne nous intéressent pas souvent, donc il faut décharger la base de données pour prendre ce qui nous intéresse. L'étape du prétraitement s'occupe généralement des tâches comme :

- Supprimer les données bruits (par exemple : commentaires vides, URLs, etc.)
- Supprimer les mots fonctionnels (Stopwords en anglais).
- Supprimer les caractères spéciaux.
- Transformer les chiffres en lettres.
- Supprimer la ponctuation.
- Convertir les lettres du texte en minuscules.
- Suppression des hashtags, des mentions (@), et des émojis.

En résumé, le prétraitement des textes peut être réalisé manuellement aussi en suivant des étapes spécifiques adaptées au cas d'utilisation.

Après le prétraitement, généralement on effectue une vectorisation qui est une partie essentielle, où elle consiste à représenter les textes sous forme de vecteurs numériques, afin de les rendre exploitables par les algorithmes d'apprentissage automatique. Il existe différentes techniques de vectorisation, telles que la représentation en sac de mots (bag-of-words ou BOW), la représentation TF-IDF (Term Frequency-Inverse Document Frequency), et l'utilisation de modèles de langage pré-entraînés comme *Word2Vec* ou *GloVe*. Chaque technique a ses propres avantages et peut être adaptée en fonction du problème spécifique et des données disponibles. Pour notre travail nous avons utilisés deux techniques de vectorisation, i.e. TF-IDF et BOW.

5. Classification

Dans cette partie nous discuterons les méthodes de validation, ainsi que des paramètres que nous avons utilisés. Au cours de notre travail nous avons effectué quatre expérimentations pour mieux voir si le modèle s'adaptait en ayant plus ou moins de données, alors pour cela nous avons expérimenté :

- Avec 2 classes (offensives et normales), où les classes abusives et offensives étaient fusionnées sous la même classe (offensive) avec le corpus équilibré (nombre de textes normaux équivalant au nombre de textes offensifs) ;
- Avec 3 classes (offensive, abusive et normale) avec le corpus équilibré (nombre de textes normaux équivalant au nombre de textes offensifs + abusifs) ;
- Avec 2 classes (offensives et normales), où les classes abusives et offensives étaient fusionnées sous la même classe (offensive) avec le corpus déséquilibré (nombre de textes différents) ;
- Avec 3 classes (offensive, abusive et normale) avec le corpus déséquilibré (nombre de textes différents) ;

La base de données est divisée en 90% pour l'entraînement dont 10% destiné à la validation, et 10% pour le test. Pour le choix des classificateurs nous nous sommes un peu basés sur le guide de scikit-learn⁷.

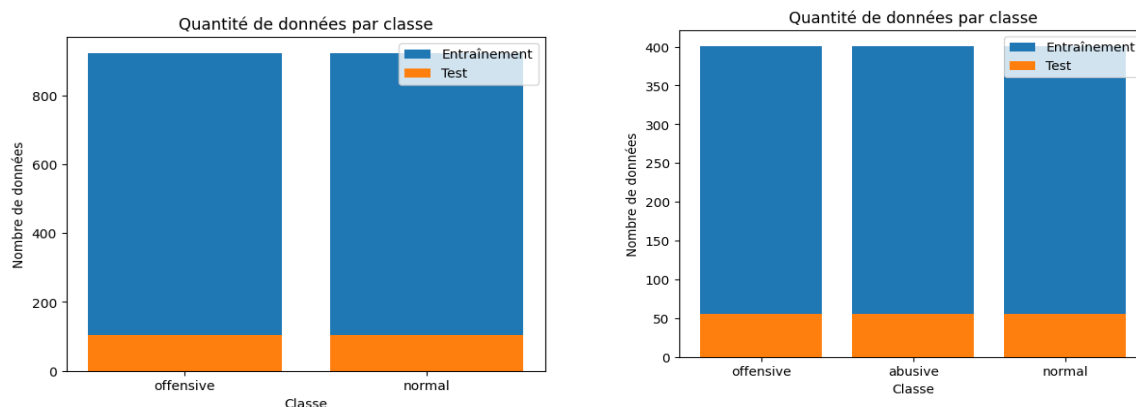
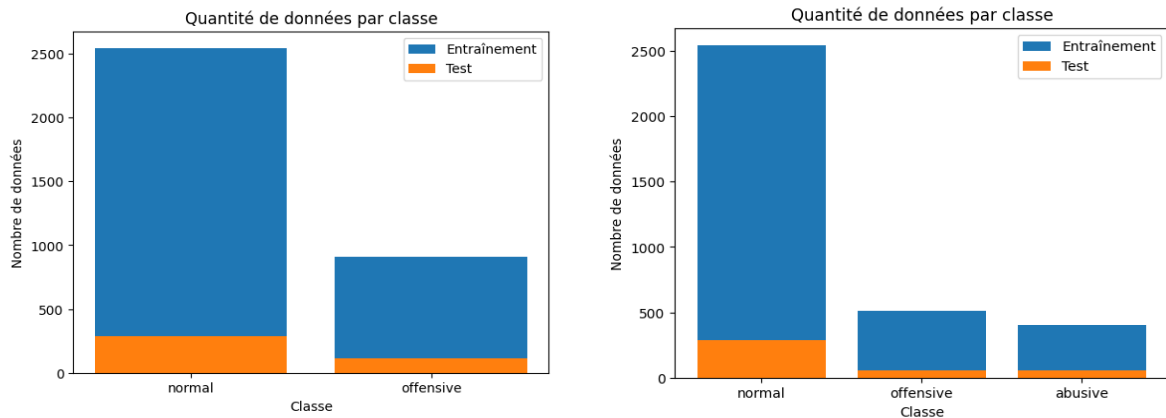


Figure IV.1 : Statistiques sur le cas du corpus équilibré pour 2 et 3 classes

⁷ <https://scikit-learn.org/>

Figure IV.2 : Statistiques sur le cas du corpus déséquilibré pour 2 et 3 classes.

5.1. Méthodes d'évaluation des classifieurs

Dans cette partie, nous discutons les différentes méthodes utilisées pour évaluer la performance des classifieurs dans notre étude.

- **Rappel** : permet de savoir le pourcentage de positifs bien prédit par le modèle. En d'autres termes c'est le nombre de positifs bien prédit « *vrai positif* » (en anglais *true positive* ou *TF*) divisé par l'ensemble des positifs « *vrai positif + faux négatif* ». Quand le rappel est haut, cela veut dire qu'il ne ratera aucun positif. Sous forme mathématique, on a :

$$\text{rappel} = \frac{\text{Vrai Positif}}{\text{Vrai Positif} + \text{Faux Négatif}} \quad (1)$$

- **Précision** : est presque similaire au rappel, sauf elle permet de connaître le nombre de prédictions positives bien effectuées. En particulier, c'est le nombre de positifs bien prédit (*vrai positif*) divisé par l'ensemble des positifs prédit (*vrai positif + faux positif*). Plus elle est élevée, plus le modèle minimise le nombre de *faux positif*. Sous forme mathématique on a :

$$\text{précision} = \frac{\text{Vrai Positif}}{\text{Vrai Positif} + \text{Faux positif}} \quad (2)$$

- **F1-Score** : est une métrique permettant de combiner la précision et le rappel. Elle permet d'effectuer une bonne évaluation de la performance des modèles, car bien que le rappel et la précision soient utiles, ils ne permettent pas d'évaluer entièrement un modèle, séparément ces deux métriques sont inutiles. Si le modèle prédit tout le temps « *positif* », le rappel sera élevé, et contrairement si le modèle ne prédit jamais « *positif* », la précision sera élevée. On aura donc des métriques qui nous indique que

notre modèle est efficace, alors qu'il sera au contraire plus naïf qu'intelligent. Plus le F1-score est élevé, plus le modèle est performant, et est calculé comme suit :

$$F1 - Score = 2 \times \frac{recall \times précision}{recall + précision} \quad (3)$$

5.2. Paramètres des classifieurs

Nous détaillons dans cette section les différents paramètres des classifieurs utilisés dans cette série d'expérimentations.

- **Régression Logistique** : le paramètre C contrôle l'inverse de la force de régularisation dans la régression logistique. Les valeurs les plus petites de C indiquent une régularisation plus forte, ce qui peut aider à éviter le surajustement du modèle. Nous avons testé différentes valeurs telles que 0.1, 1.0, 1.02, 1.05, 1.2, 2.0 et 10.0.
- **Multinomial NB** : le paramètre $alpha$ est un hyperparamètre de l'algorithme Naïve Bayes multinomial qui contrôle le lissage de *Laplace*. Il est utilisé pour régulariser les estimations de probabilité, où nous avons testé les valeurs suivantes : 0.1, 1.0, 1.0, 1.02, 1.05, 1.2, 2.0 et 10.0.
- **SVM** : le paramètre C contrôle la marge d'erreur du modèle SVM. Des valeurs plus petites de C permettent une marge d'erreur plus large, tandis que des valeurs plus grandes la réduisent. On a essayé les mêmes valeurs des classifieurs précédents (i.e. 0.1, 1.0, 1.0, 1.02, 1.05, 1.2, 2.0 et 10.0.). Le paramètre *kernel* spécifie le noyau à utiliser dans le modèle SVM, où *linear* indique un noyau linéaire, tandis que *RBF* (ou *radial basis function*) indique un noyau gaussien. On a testé les deux noyaux pour savoir le meilleur dans notre application, malgré le dernier a prouvé ses performances dans différentes applications.
- **Random Forest** : le paramètre $nb_estimators$ spécifie le nombre d'arbres dans la forêt aléatoire. Une valeur plus élevée peut améliorer les performances du modèle, mais cela nécessite également plus de temps de calcul. Dans notre expérimentation, nous avons utilisé 50, 100, 200 et 300.

Il faut souligner que tous ces paramètres sont utilisés pour tester chaque modèle avec différentes combinaisons d'hyperparamètres en utilisant *GridSearchCV*⁸. Pour trouver les meilleures combinaisons d'hyperparamètres pour chaque modèle.

- **CNN** : dans ce classifieur, on a fixé la dimension du vecteur d'entrée à la taille du vocabulaire du corpus d'apprentissage, la dimension de sortie à 100, ainsi la taille du texte à la taille maximale du plus long texte. Si un texte ne satisfait pas cette condition, la différence est remplie avec indicateurs spéciaux. Le corpus d'apprentissage est divisé en 32 parties (taille du *batch*), ainsi on a entraîné le modèle en utilisant seulement 10 itérations vu la capacité matériel dont on dispose. En outre, concernant les autres paramètres, on les résume dans le tableau 4.1.

Table 4.2. Paramètres utilisés dans le classifieur CNN

Paramètres	Valeurs
Couche de convolution	filters = 128, kernel_size = 5, activation = RELU
Couche dense	units = 64, activation = RELU
Sortie de la couche dense	units = 1, activation = sigmoid
Fonction de perte	binary crossentropy
Optimiseur	Adam
Mesures	accuracy

- **BiLSTM** : Contrairement au CNN, le corpus d'apprentissage est divisé en 4 parties (taille du *batch*), ainsi on a entraîné le modèle en utilisant seulement 5 itérations, car BiLSTM est trop long vu qu'il travaille dans les deux sens. Ces deux facteurs sont relatifs, où plus on augmente la taille du premier plus on diminue le nombre d'itérations et vice-versa. On a fixé la dimension du vecteur d'entrée à la taille du vocabulaire du corpus d'apprentissage, la dimension de sortie à 32, ainsi la taille du texte à la taille maximale du plus long texte. Le reste des paramètres est résumé dans le tableau 4.2.

⁸ Une méthode de recherche de meilleurs paramètres automatiquement on essayant différentes combinaison

Table 4.3. Paramètres utilisés dans le classifieur BiLSTM

Paramètres	Valeurs
Couche LSTM 1	32
Couche LSTM 1	16
Couche dense	1
Fonction d'activation	sigmoid
Fonction de perte	binary crossentropy
Optimiseur	Adam

- **FastText**⁹: cet outil est très puissant et ultra rapide pour créer un modèle de langue efficace, où il a été développé par Facebook pour la classification des langues. Nous l'utilisons dans notre étude afin de répondre à notre problématique, où nous lançons l'entraînement sur notre corpus avec les paramètres résumés dans Table 4.3.

Table 4.4. Paramètres utilisés pour entraîner l'outils FastText

Paramètre	Valeur
Epochs	10
Taille du batch	4
Patience (pour EarlyStopping)	2
Dimension de sortie (embedding)	100
Taille maximale du texte	texte le plus long
Units (dense layer)	1
Fonction d'activation	sigmoid
Fonction de perte	binary crossentropy
Optimiseur	Adam

⁹ <https://fasttext.cc/>

6. Evaluation et résultats

Dans cette section, nous présentons les différents résultats obtenus à travers la série d'expérimentations effectuées, ainsi nous discutons les performances de chacun des classifieurs.

6.1. Résultats

Nous avons remarqué que la meilleure valeur du paramètre C du classifieur SVM est 2.0 (cherchée par *GridSearch*), tandis que le noyau se diffère d'une expérimentation à une autre. En particulier, pour le cas de l'expérimentation avec deux classes sur le corpus équilibré, le meilleur noyau est *RBF*, par contre le noyau *linéaire* est le meilleur choix pour les autres expérimentations.

Table 4.5. Résultats du SVM avec deux et trois classes sur le corpus équilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.78	0.92	0.84		0.81	0.80	0.82	
Offensif	0.92	0.76	0.83		0.49	0.48	0.49	
Abusif	-	-	-		0.51	0.54	0.52	
Avg	0.85	0.84	0.85	0.84	0.61	0.61	0.61	0.61

Table 4.6. Résultats du SVM avec deux et trois classes sur le corpus déséquilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.90	0.95	0.92		0.89	0.95	0.92	
Offensif	0.85	0.73	0.78		0.45	0.34	0.39	
Abusif	-	-	-		0.45	0.39	0.42	
Avg	0.88	0.88	0.88	0.88	0.76	0.79	0.77	0.79

Contrairement à SVM, le classifieur de régression logistique a gardé la même valeur du paramètre C à travers tout l'ensemble d'expérimentations (valeur égale à 10). Similairement, le classifieur naïf de Bayes a gardé la même valeur de α (i.e. 0.1). Dans le cas du classifieur Random Forest, on a obtenu différentes valeurs du paramètre $nb_estimators$, dont 300, 200 et 50.

Table 4.7. Résultats de régression logistique avec deux et trois classes sur le corpus équilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.85	0.83	0.84		0.82	0.80	0.81	
Offensif	0.85	0.87	0.86		0.48	0.50	0.49	
Abusif	-	-	-		0.53	0.52	0.52	
Avg	0.85	0.85	0.85	0.85	0.61	0.61	0.61	0.61

Table 4.8. Résultats de régression logistique avec deux et trois classes sur le corpus déséquilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.90	0.96	0.93		0.88	0.97	0.92	
Offensif	0.87	0.73	0.79		0.53	0.34	0.42	
Abusif	-	-	-		0.44	0.36	0.40	
Avg	0.89	0.89	0.89	0.89	0.76	0.79	0.77	0.79

Table 4.9. Résultats de NB avec deux et trois classes sur le corpus équilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.80	0.91	0.85		0.83	0.80	0.82	
Offensif	0.91	0.80	0.84		0.55	0.46	0.50	
Abusif	-	-	-		0.48	0.57	0.50	
Avg	0.86	0.86	0.85	0.85	0.61	0.61	0.61	0.61

Table 4.10. Résultats de NB avec deux et trois classes sur le corpus déséquilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.91	0.95	0.93		0.88	0.98	0.93	
Offensif	0.86	0.78	0.82		0.47	0.38	0.42	
Abusif	-	-	-		0.39	0.27	0.32	
Avg	0.90	0.90	0.90	0.90	0.76	0.79	0.77	0.79

Table 4.11. Résultats de Random Forest avec deux et trois classes sur le corpus équilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.71	0.91	0.80		0.65	0.90	0.76	
Offensif	0.89	0.67	0.76		0.54	0.43	0.48	
Abusif	-	-	-		0.54	0.27	0.36	
Avg	0.80	0.79	0.78	0.78	0.59	0.57	0.57	0.57

Table 4.12. Résultats de Random Forest avec deux et trois classes sur le corpus déséquilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.81	0.96	0.88		0.82	0.97	0.89	
Offensif	0.82	0.44	0.57		0.50	0.29	0.37	
Abusif	-	-	-		0.56	0.27	0.36	
Avg	0.82	0.81	0.79	0.81	0.72	0.77	0.73	0.77

Table 4.13. Résultats du SGD avec deux et trois classes sur le corpus équilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.84	0.84	0.84		0.80	0.77	0.78	
Offensif	0.86	0.85	0.85		0.49	0.52	0.50	
Abusif	-	-	-		0.47	0.46	0.47	
Avg	0.86	0.85	0.86	0.86	0.76	0.78	0.77	0.58

Table 4.14. Résultats du SGD avec deux et trois classes sur le corpus déséquilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.93	0.86	0.89		0.82	0.89	0.85	
Offensif	0.70	0.84	0.76		0.47	0.30	0.37	
Abusif	-	-	-		0.46	0.59	0.52	
Avg	0.86	0.85	0.86	0.85	0.59	0.60	0.58	0.78

Table 4.15. Résultats du CNN avec deux et trois classes sur le corpus équilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.84	0.82	0.83		0.83	0.81	0.82	
Offensif	0.84	0.86	0.85		0.47	0.60	0.53	
Abusif	-	-	-		0.51	0.39	0.44	

Avg	0.84	0.84	0.84	0.84	0.59	0.60	0.58	0.60
------------	------	------	------	------	------	------	------	------

Table 4.16. Résultats du CNN avec deux et trois classes sur le corpus déséquilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.94	0.93	0.94		0.99	0.81	0.89	
Offensif	0.84	0.90	0.85		0.45	0.43	0.44	
Abusif	-	-	-		0.25	0.50	0.34	
Avg	0.87	0.89	0.88	0.90	0.56	0.58	0.56	0.71

Table 4.17. Résultats du BiLSTM avec deux et trois classes sur le corpus équilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.87	0.87	0.87		0.91	0.73	0.81	
Offensif	0.89	0.89	0.89		0.51	0.46	0.49	
Abusif	-	-	-		0.46	0.59	0.52	
Avg	0.88	0.88	0.88	0.88	0.63	0.60	0.60	0.60

Table 4.18. Résultats du BiLSTM avec deux et trois classes sur le corpus déséquilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.95	0.98	0.97		0.91	0.97	0.94	
Offensif	0.94	0.88	0.91		0.59	0.59	0.59	
Abusif	-	-	-		0.43	0.29	0.34	
Avg	0.95	0.93	0.94	0.95	0.64	0.61	0.62	0.82

Table 4.19. Résultats du FastText avec deux et trois classes sur le corpus équilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.91	0.77	0.83		0.75	0.84	0.79	
Offensif	0.82	0.93	0.87		0.48	0.55	0.51	
Abusif	-	-	-		0.50	0.36	0.42	
Avg	0.86	0.85	0.85	0.85	0.57	0.58	0.57	0.58

Table 4.20. Résultats du FastText avec deux et trois classes sur le corpus déséquilibré

	Deux classes				Trois classes			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Normal	0.95	0.94	0.95		0.96	0.89	0.92	
Offensif	0.86	0.89	0.91		0.40	0.72	0.52	
Abusif	-	-	-		0.52	0.27	0.35	
Avg	0.91	0.91	0.91	0.93	0.63	0.63	0.60	0.78

6.2. Discussions

Après évaluation des performances de tous ces classificateurs sur les différentes configurations de la base de données, nous remarquons que SVM a produit 0.84 de F1-score dans le cas de deux classes sur le corpus équilibré, ce qui représente un bon score. En revanche, dans le test de trois classes le score a chuté jusqu'à 0.62, ce qui est logique avec l'augmentation du nombre de classes. Pour le cas du corpus déséquilibré, le même modèle a produit des scores F1 de 0.85 dans deux classes et de 0.77 dans trois classes. On peut constater que le modèle SVM a du mal à gérer les classes minoritaires.

Le modèle de régression logistique a montré presque les mêmes performances que le SVM. Les scores F1 obtenus sont de 0.85 pour le corpus équilibré à deux classes et de 0.62 pour trois classes. Contrairement, dans le cas du corpus déséquilibré, le modèle de régression logistique a obtenu des scores F1 très supérieur (0.89 pour deux classes et 0.77 pour trois classes), ce qui veut dire qu'il s'adapte bien aux classes minoritaires.

Le NB a des performances pas mal, avec des scores F1 allant de 0.85 à 0.60 pour les corpus équilibrés. Cependant, le modèle a bien géré avec le corpus déséquilibré pour les deux classes, les scores F1 sont de 0.87 pour les deux classes et de 0.77 pour les trois classes. Quant au Random Forest il a obtenu des résultats moyennes avec le corpus équilibré à deux classes mais a su gérer le corpus déséquilibré, les scores F1 obtenus variaient de 0.58 à 0.78 pour les corpus équilibrés, et de 0.73 à 0.74. Le SGD a des performances comparables à celles de la régression logistique et du SVM dans le corpus équilibré. Cependant, contrairement aux deux autres, il a bien géré le corpus déséquilibré.

CNN a obtenu des scores F1 de 0.83 et 0.60 dans le corpus équilibré de 2 et 3 classes. Et pour le corpus déséquilibré des super scores pour chaque classe avec des F1 de 0.91 et 0.76, très bonne performance comparée au corpus équilibré. Dans le corpus équilibré, BiLSTM a le record de mesure F1 score avec 0.87 pour deux classes et 0.64 pour trois classes. De très bon

scores pour le corpus déséquilibré avec 0.95 pour 2 classes et 0.80 pour trois. FastText produit de très bonne performance comme pour le BiLSTM avec presque les mêmes scores pour le corpus déséquilibré et un score de F1 très faible pour 3 trois classes avec le corpus équilibré.

Enfin pour finir, nos résultats ont montré que les performances des modèles se changent en fonction du corpus utilisé (équilibré ou déséquilibré). Les résultats indiquent que les modèles ont mieux fonctionné lorsque le nombre de classes était réduit, tandis qu'ils sont diminués avec l'augmentation du nombre de classes et surtout à la faible quantité de nos données, car les données de la classe normale qui sont plus nombreuses sont toujours bien prédites partout. Cela montre l'importance de prendre un maximum de données de même taille pour des résultats corrects.

7. Conclusion

Ce chapitre présente en détail tout notre processus de la création et l'exécution de notre modèle en passant de la collecte des données à leur traitement, ainsi qu'à leur annotation par des annotateurs en 3 catégories pour notre classification. L'évaluation nous a données des F1 scores remarquables dans l'ensemble.

Pour conclure, nous pouvons dire sans crainte à quel point il est important d'avoir de très grandes bases de données pour la détection des contenus offensifs. Car avoir plein de données variées joue un rôle essentiel dans la capacité du modèle à généraliser et à identifier efficacement les contenus offensifs.

Cependant notre méthodologie a prouvé son efficacité dans la création et l'exécution de modèles de détection des contenus offensifs, les résultats obtenus le démontre. Les futures avancées nous permettront d'atteindre des performances encore meilleures, contribuant ainsi à créer un environnement en ligne plus sûr et plus sain.

CONCLUSION GÉNÉRALE

Pour conclure, notre travail avait pour objectif principal de développer un système permettant de détecter les contenus offensifs automatiquement en bambara sur les réseaux sociaux. L'objectif était de contribuer à la lutte contre les discours haineux et les contenus offensifs en ligne. La détection des contenus offensifs en bambara est une grande importance dans un monde de plus en plus connecté. Nous avons pu constater que les réseaux sociaux sont devenus des lieux incontournables où les individus se connectent et interagissent les uns avec les autres et que cette popularité avait également engendré des risques pour la sécurité et le bien-être des utilisateurs du bambara, ce à quoi les techniques existantes pour les protéger n'existent pas.

C'est dans ce contexte que notre travail a cherché à combler ce problème en proposant une approche basée sur l'apprentissage automatique pour détecter et atténuer les contenus offensifs en bambara. Nous avons collecté un ensemble de données en bambara, puis avons appliqué des techniques de prétraitement des textes pour normaliser et nettoyer ces données. En développant des modèles d'apprentissage automatique, nous avons obtenu des résultats prometteurs, en particulier avec le modèle BiLSTM, FastText, LR et NB.

Toutefois, bien que les résultats de notre travail soient prometteurs, il y a quelques limites à celles-ci. Tout d'abord, le manque de disponibilité des données en bambara a constitué un défi majeur, car la langue est de loin très sous documentée comparé à pleins d'autres langues, à cause de sa prédominance orale. Nous avons dû nous appuyer sur un ensemble restreint de données, ce qui pourrait modifier les résultats. De plus, comme les réseaux sociaux évoluent rapidement, de nouvelles formes de contenu offensif peuvent émerger, nécessitant une mise à jour régulière de nos modèles de détection.

Malgré ces limitations, notre recherche ouvre des perspectives intéressantes pour la protection des utilisateurs du bambara sur les réseaux sociaux. Nos résultats préliminaires démontrent l'efficacité de l'apprentissage automatique dans la détection des contenus offensifs en bambara, ce qui pourrait servir de base pour le développement de futurs systèmes de filtrage et de modération des réseaux sociaux.

En conclusion, notre travail contribue à la promotion d'un environnement en ligne plus sûr et respectueux pour les utilisateurs du bambara, en fournissant un outil

efficace pour identifier et traiter ces contenus. Il souligne l'importance de la recherche et du développement de solutions adaptées aux langues moins documentées dans le domaine de la détection des contenus offensifs. Nous espérons que notre travail encouragera davantage de recherches dans ce domaine et apportera des avancées significatives pour la protection des utilisateurs sur les réseaux sociaux.

REFERENCES BIBLIOGRAPHIQUES

- [1] D. Jurafsky et J. H. Martin, « Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition Second Edition ».
- [2] Adrian, « L'intelligence artificielle : une solution à la prise de décision quotidienne ? », Mémoire de recherche appliquée, École Supérieure de Génie Informatique de Paris, 2019.
- [3] Meghyn Bienvenu, « Introduction à l'intelligence artificielle ». Consulté le: 18 mars 2023. [En ligne]. Disponible sur: <https://www.labri.fr/perso/meghyn/papers/>
- [4] Boucherit Oussama, « Automatic Algerian offensive language detection in social media networks », Mémoire de Master Académique, Université 8 Mai 1945 de Guelma, 2021.
- [5] Zouinar et Moustafa, « Évolutions de l'Intelligence Artificielle : quels enjeux pour l'activité humaine et la relation Humain-Machine au travail ? », <http://journals.openedition.org/activites>, n° 17-1, avr. 2020, doi: 10.4000/ACTIVITES.4941.
- [6] Christian Szegedy, « Building a deeper understanding of images », 2014. <https://ai.googleblog.com/2014/09/building-deeper-understanding-of-images.html> (consulté le 18 mars 2023).
- [7] D. Silver *et al.*, « A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play », *Science (1979)*, vol. 362, n° 6419, p. 1140-1144, déc. 2018, doi: 10.1126/SCIENCE.AAR6404.
- [8] David Silver et Demis Hassabis, « AlphaGo Zero: Starting from scratch », 2017. <https://www.deepmind.com/blog/alphago-zero-starting-from-scratch> (consulté le 18 mars 2023).
- [9] Alec Radford *et al.*, « Better language models and their implications ». <https://openai.com/research/better-language-models> (consulté le 18 mars 2023).
- [10] « AlphaFold: a solution to a 50-year-old grand challenge in biology ». <https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology> (consulté le 18 mars 2023).
- [11] « Overview of Natural Language Processing - NetBase Quid ». <https://netbasequid.com/blog/overview-natural-language-processing/> (consulté le 18 mars 2023).
- [12] Oracle Suisse, « Qu'est-ce que le traitement du langage naturel (NLP) ? » <https://www.oracle.com/ch-fr/artificial-intelligence/what-is-natural-language-processing/> (consulté le 18 mars 2023).
- [13] Dragomir Radev, « The Association for Computational Linguistics Member Portal », 2023. <https://www.aclweb.org/portal/> (consulté le 12 mars 2023).

- [14] SMG Languages, « Linguistique informatique », 2017.
<https://www.smglanguages.com/linguistique-informatique/?lang=fr> (consulté le 15 mars 2023).
- [15] P. Goutefangea et P. Goutefangea Alan, « Alan Turing et l'intelligence artificielle : le " jeu de l'imitation " et " l'IA forte " ». [En ligne]. Disponible sur: <https://hal.science/hal-01330278v2>
- [16] Stanford Encyclopedia of Philosophy, « Computational Linguistics », 2014.
<https://plato.stanford.edu/entries/computational-linguistics/> (consulté le 12 mars 2023).
- [17] Université de Genève, « Présentation de la discipline Linguistique », 2006.
- [18] Département de linguistique - UQAM, « Qu'est-ce que la linguistique ? »
<https://linguistique.uqam.ca/futurs-etudiants/quest-ce-que-la-linguistique/> (consulté le 16 mars 2023).
- [19] « Qu'est-ce que le traitement du langage naturel ? | Google Cloud | Google Cloud ». <https://cloud.google.com/learn/what-is-natural-language-processing?hl=fr> (consulté le 18 mars 2023).
- [20] Y. Bengio, « Neural net language models », *Scholarpedia*, vol. 3, n° 1, p. 3881, 2008, doi: 10.4249/SCHOLARPEDIA.3881.
- [21] Marcel Cori et Sophie David, « Les corpus fondent-ils une nouvelle linguistique ? », 2008. <https://www.cairn.info/revue-langages-2008-3-page-111.htm#no66> (consulté le 18 mars 2023).
- [22] « Qu'est-ce que la reconnaissance vocale ? | IBM ». <https://www.ibm.com/ca-fr/topics/speech-recognition> (consulté le 16 mars 2023).
- [23] Nedioui Med Abdelhamid, « Techniques d'apprentissage automatique pour l'analyse et la fouille des sentiments dans les réseaux sociaux », thèse de Doctorat, Université Mohamed Kheider-Biskra, 2021.
- [24] Futura Tech, « Définition | Chatbot - Agent conversationnel ». <https://www.futura-sciences.com/tech/definitions/internet-chatbot-15778/> (consulté le 16 mars 2023).
- [25] Botnation, « Botnation, le chatbot français facilement adaptable à vos besoins ». https://botnation.ai/site/fr/chatbot-francais/#Le_chatbot_quest-ce_que_cest_concretement%E2%80%89 (consulté le 16 mars 2023).
- [26] MATALLAH Hociine, « Classification automatique de textes approches orientée agent », Mémoire de Magister, Université Aboubekr Belkaid-Tlemcen, 2011.
- [27] « Introduction au NLP - Partie 5/6 - Limites de l'interaction homme-machine ». <https://blog.clevy.io/fr/introduction-au-nlp-5eme-partie/> (consulté le 17 mars 2023).
- [28] Géoconfluences, « Langues et dialectes », 2023. <http://geoconfluences.ens-lyon.fr/glossaire/langues> (consulté le 28 avril 2023).
- [29] « La famille indo-européenne | Compendium of Language Management in Canada (CLMC) | Université d'Ottawa ». <https://www.uottawa.ca/calcul/famille-indo-europeenne> (consulté le 31 mars 2023).

- [30] inalco, « formation_mandingue_licence_llcer », 2021.
- [31] MALI PENSE, « akbk1-cours 01 ». <https://www.mali-pense.net/akbk1-cours-01.html> (consulté le 28 avril 2023).
- [32] Agence de Presse Régionale, « Les bambaras : Qui sont-ils ? D'où viennent-ils? », 2018. <https://apr-news.fr/fr/actualites/les-bambaras-qui-sont-ils-dou-viennent-ils> (consulté le 25 mars 2023).
- [33] Konta Mahamadou, AMALAN, et Vydrin Valentin, « Propositions pour l'orthographe du bamanankan ».
- [34] Simon Ager, « Bambara alphabet, pronunciation and language ». <https://omniglot.com/writing/bambara.htm> (consulté le 1 mai 2023).
- [35] « UN PINCEAU SUR LE TABLEAU MALI: Ecriture : LE N'KO S'EXPORTE en France ». <http://tableaumali.blogspot.com/2009/10/ecriture-le-nko-sexporte-en-france.html> (consulté le 31 mai 2023).
- [36] Danièle Dubois, « Psycholinguistique et psychologie du langage - Persée », 1972. https://www.persee.fr/doc/psy_0003-5033_1972_num_72_2_27960 (consulté le 17 avril 2023).
- [37] Enrique Moreira, « Comment le langage évolue sur les réseaux sociaux ? », 2017. <https://www.lesechos.fr/2017/05/comment-le-langage-evolue-sur-les-reseaux-sociaux-167558> (consulté le 26 mai 2023).
- [38] Hela SAYADI, « Violence verbale dans les espaces publics et sur les réseaux sociaux : Les insultes fusent de partout », 2020. <https://lapresse.tn/82774/violence-verbale-dans-les-espaces-publics-et-sur-les-reseaux-sociaux-les-insultes-fusent-de-partout/> (consulté le 17 avril 2023).
- [39] A. de Latour *et al.*, *Le discours de haine*, n° 2017. Conseil de l'Europe, 2019. Consulté le: 6 avril 2023. [En ligne]. Disponible sur: <https://www.cairn.info/alternatives--9789287184603-page-30.htm?contenu=article>
- [40] World Health Organization, « Qu'entend-on par « discours de haine » ? | Nations Unies », 2023. <https://www.un.org/fr/hate-speech/understanding-hate-speech/what-is-hate-speech> (consulté le 17 avril 2023).
- [41] Fayssal Zerargui et Omar Benzaoui, « Le discours de haine sur le web et les médias sociaux », Mémoire de Master Académique, Université de Bordj Bou Arreridj Mohammed El Bachir El Ibrahimi, 2021.
- [42] NASSIROU DAOUDA Aminou, « FILTRAGE DU TRAFIC OFFENSIF SUR LES RESEAUX SOCIAUX », Mémoire de Master Académique, Université 8 Mai 1945 Guelma, 2022.
- [43] World Health Organization, « Troubles mentaux », 2022. <https://www.who.int/fr/news-room/fact-sheets/detail/mental-disorders> (consulté le 17 avril 2023).

- [44] Doctissimo, « L'anxiété - Définition, symptômes et traitement », 2022. https://www.doctissimo.fr/html/sante/encyclopedie/sa_781_anxiete.htm (consulté le 17 avril 2023).
- [45] Kowsari *et al.*, « Text Classification Algorithms: A Survey », 2019. https://www.researchgate.net/figure/This-figure-shows-the-linear-and-non-linear-SVM-for-2D-dataset-for-text-data-we-have_fig11_332494043 (consulté le 12 mai 2023).
- [46] Cnam – UE RCP209, « Cours - Arbres de décision ». <https://cedric.cnam.fr/vertigo/cours/ml2/coursArbresDecision.html> (consulté le 12 mai 2023).
- [47] MANSOUR Abd Elfettah et TRAD Aissa, « La détection automatique du discours abusif, offensant et obscène dans le dialecte algérien », Mémoire de Master Académique, UNIVERSITE ECHAHID HAMMA LAKHDAR - EL OUED, 2018.
- [48] « KNN : Découvrez cet algorithme de Machine Learning ». <https://datascientest.com/knn> (consulté le 12 mai 2023).
- [49] Adi Bronshtein, « A Quick Introduction to K-Nearest Neighbors Algorithm », 2017. <https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7> (consulté le 26 mai 2023).
- [50] Gourav Singh, « Introduction to Artificial Neural Networks », *Analytics Vidhya*. <https://www.analyticsvidhya.com/blog/2021/09/introduction-to-artificial-neural-networks/> (consulté le 1 juin 2023).
- [51] FOFANA Mohamed Laffi, « Identification biométrique par descripteur de Texture multi-échelle: application à l'oreille », Mémoire de Master Académique, Université 8 Mai 1945 de Guelma, 2022.
- [52] C. Olah, « Understanding LSTM Networks », 2015. <https://colah.github.io/posts/2015-08-Understanding-LSTMs/> (consulté le 18 mai 2023).
- [53] J. Xie, B. Chen, X. Gu, F. Liang, et X. Xu, « Self-Attention-Based BiLSTM Model for Short Text Fine-Grained Sentiment Classification », *IEEE Access*, vol. 7, p. 180558-180570, 2019, doi: 10.1109/ACCESS.2019.2957510.