# DEDICACES

# REMERCIEMENTS

# GLOSSAIRE

# LISTE DES FIGURES

# LISTES DES TABLEAUX

# RESUME

# ABSTRACT

SOMMAIRE

[DEDICACES 1](#_Toc187678680)

[REMERCIEMENTS 2](#_Toc187678681)

[GLOSSAIRE 3](#_Toc187678682)

[LISTE DES FIGURES 4](#_Toc187678683)

[LISTES DES TABLEAUX 5](#_Toc187678684)

[RESUME 6](#_Toc187678685)

[ABSTRACT 7](#_Toc187678686)

[INTRODUCTION 10](#_Toc187678687)

[CHAPITRE I : ETAT DE L’ART 11](#_Toc187678688)

[I. ETAT DE L’ART SUR LE DOMAINE 12](#_Toc187678689)

[1. CONTEXTE GENERALE 12](#_Toc187678690)

[2. Application au domaine de violences 12](#_Toc187678691)

[3. Problématiques associées 13](#_Toc187678692)

[II. ETAT DE L’ART SUR LES TRAVAUX DE RECHERCHE LIES AU SUJET 15](#_Toc187678693)

[1. Méthode de détection de violence 15](#_Toc187678694)

[1.1. Apprentissage automatique (Machine Learning) 15](#_Toc187678695)

[1.2. Apprentissage Profond (Deep Learning) 15](#_Toc187678696)

[2. Applications Pratiques 16](#_Toc187678697)

[2.1. Détection de discours haineux (Hate Speech) 16](#_Toc187678698)

[2.2. Identification des Zones de Recherche Manquantes liées au Sujet 17](#_Toc187678699)

[3. Pertinence du Sujet 18](#_Toc187678700)

[CHAPITRE II : ANALYSE DE BESOIN ET METHODOLOGIE 20](#_Toc187678701)

[I. Spécification des Besoins et des Données 21](#_Toc187678702)

[1. Analyse et Spécification des Besoins 21](#_Toc187678703)

[2. Sources de Données 21](#_Toc187678704)

[**2.1.** Twitter 21](#_Toc187678705)

[**2.2.** Reddit 22](#_Toc187678706)

[**2.3.** Autres sources 22](#_Toc187678707)

[3. Processus de Collecte 23](#_Toc187678708)

[**3.1.** Utilisation des API dédiées 23](#_Toc187678709)

[**3.2.** Automatisation avec des scripts Python 23](#_Toc187678710)

[**3.3.** Gestion des limitations d’accès 23](#_Toc187678711)

[**3.4.** Enrichissement des données 24](#_Toc187678712)

[2. Pré-processing des Données 25](#_Toc187678713)

[CHAPITRE3 : IMPLEMENTATION 30](#_Toc187678714)

.

# INTRODUCTION

Les réseaux sociaux ont radicalement transformé la manière dont les individus interagissent, partagent leurs expériences et expriment leurs émotions. Ces plateformes, bien qu’elles soient des espaces d’échange et de connexion, deviennent également des lieux où des traces des réalités les plus intimes de leurs utilisateurs peuvent être détectées. En particulier, les victimes potentielles de violence domestique, physique ou numérique utilisent parfois ces espaces pour se livrer, partager leur détresse ou, inconsciemment, révéler des indices sur leur vécu quotidien.

La violence domestique et physique, profondément enracinée dans les dynamiques sociales et familiales, reste souvent cachée dans la sphère privée. Les victimes, par peur, honte ou isolement, hésitent à chercher de l’aide directement. Cependant, certaines trouvent dans les réseaux sociaux un moyen de s’exprimer. À travers leurs publications, commentaires ou interactions, elles peuvent laisser transparaître des signaux émotionnels ou comportementaux révélateurs : expressions de tristesse, de peur, de désespoir ou encore des récits indirects de situations qu’elles vivent au quotidien. Ces indices, bien qu’ils puissent sembler anodins ou cryptiques pour un observateur humain, contiennent des informations précieuses qui, si elles sont correctement interprétées, pourraient permettre d’intervenir avant que la situation ne dégénère.

Les réseaux sociaux regorgent de contenus qui peuvent refléter des expériences de violence sous différentes formes. Ces violences, bien qu'exprimées dans des espaces virtuels, ont des répercussions réelles sur la vie des individus. Les principales formes de violences identifiables en ligne incluent :

* **Violence physique :** Même si elle n’est pas directement visible sur les réseaux sociaux, elle est souvent évoquée à travers des récits de coups, blessures ou maltraitances physiques infligées par un partenaire, un membre de la famille ou une autre personne.
* **Violence morale :** Ce type de violence, parfois subtil, inclut l’humiliation, la manipulation, les insultes ou les comportements visant à diminuer l’estime de soi d’une personne. Ces publications peuvent contenir des indices textuels révélant une souffrance psychologique profonde.
* **Violence numérique :** Également appelée cyberviolence, elle englobe le cyberharcèlement, la divulgation de données personnelles (doxxing), l’intimidation en ligne, ou encore la diffusion non consentie de contenus privés. Les victimes peuvent laisser des indices dans leurs messages, comme des références à des attaques subies en ligne.
* **Violence conjugale :** Elle peut combiner les formes mentionnées ci-dessus et s'exprime souvent à travers des récits d’expériences dans des relations toxiques ou abusives.
* **Harcèlement :** Cela inclut le harcèlement sexuel, les menaces ou la poursuite d’une personne sur plusieurs plateformes, souvent accompagné de commentaires explicites ou répétitifs.

À ce jour, les recherches en intelligence artificielle (IA) appliquées aux violences se concentrent principalement sur l’identification des auteurs de comportements nuisibles, comme les agresseurs en ligne ou les discours haineux. Ces approches, bien que cruciales, ne s’intéressent que partiellement aux victimes, notamment celles qui ne dénoncent pas explicitement leur situation. Cela laisse un vide significatif dans la recherche et les pratiques actuelles. Comment détecter qu’une personne est une victime potentielle à travers ce qu’elle publie sur les réseaux sociaux ? Comment analyser ses écrits pour comprendre les signes d’une détresse liée à la violence dans son quotidien ?

Les défis sont nombreux. D’un point de vue technique, les contenus publiés sur les réseaux sociaux sont souvent informels, fragmentés et fortement influencés par des facteurs linguistiques et culturels. Développer des modèles d’IA capables de traiter ce type de données tout en capturant des signaux émotionnels subtils demande une expertise avancée en traitement automatique du langage naturel (NLP) et en apprentissage machine (ML). D’un point de vue social, il est crucial de fournir des outils non seulement performants, mais aussi accessibles aux associations, plateformes et institutions engagées dans la prévention des violences. D’un point de vue éthique, la protection de la vie privée, la minimisation des biais algorithmiques et la transparence dans l’interprétation des résultats sont des priorités absolues.

Ce mémoire s’inscrit dans cette problématique en proposant le développement d’un modèle d’IA capable de détecter des victimes potentielles de violence domestique, physique ou numérique à travers l’analyse de leurs écrits sur les réseaux sociaux. Ce travail ambitionne de contribuer à la prévention proactive des violences en fournissant un outil capable d’interpréter les signaux émotionnels et comportementaux qui émanent des publications en ligne.

Les objectifs spécifiques incluent :

* Développement d’un modèle robuste et éthique permettant d’identifier des indices de détresse émotionnelle dans des publications textuelles issues des réseaux sociaux.
* Validation des performances du modèle sur des jeux de données réels et variés, en prenant en compte les dimensions linguistiques et culturelles.
* Proposition de recommandations pratiques, destinées aux acteurs sociaux, pour l’utilisation et l’interprétation des résultats produits par ce modèle.

À travers ce travail, nous espérons fournir une réponse novatrice à un problème critique et insuffisamment exploré. Ce mémoire se positionne à l’intersection de la technologie, des sciences sociales et de l’éthique, en vue de concevoir des solutions concrètes pour identifier et protéger les individus vulnérables, en donnant aux signaux numériques une voix porteuse de prévention et de changement.

# CHAPITRE I : ETAT DE L’ART

# ETAT DE L’ART SUR LE DOMAINE

## CONTEXTE GENERALE

La détection des victimes potentielles de violences domestiques, numériques ou physiques à travers les réseaux sociaux constitue un domaine de recherche émergent, à la croisée de l’intelligence artificielle, des sciences sociales et de l’éthique. Ce domaine repose sur l’observation que les comportements humains en ligne peuvent révéler des indices précieux sur les situations de vulnérabilité ou de détresse, même lorsqu’ils ne sont pas explicitement formulés.

Le domaine d’application de cette recherche s’inscrit dans un contexte global marqué par l’augmentation des violences en ligne et hors ligne, exacerbées par l’adoption massive des technologies numériques. Les réseaux sociaux, en tant que miroir des dynamiques sociales, offrent un terrain fertile pour l’analyse de ces phénomènes. Cependant, cette opportunité s’accompagne de défis majeurs, notamment en termes de traitement des données massives et de respect des principes éthiques fondamentaux.

Plusieurs problématiques spécifiques émergent dans ce domaine. Premièrement, les données des réseaux sociaux sont souvent bruitées et informelles, ce qui complique leur analyse. Deuxièmement, les indices de vulnérabilité, tels que des changements dans les interactions sociales ou des expressions émotionnelles particulières, sont souvent implicites et subtils. Enfin, les biais algorithmiques dans les modèles d’intelligence artificielle peuvent entraîner des discriminations ou des erreurs dans l’identification des victimes potentielles.

L’objectif de cette section est de situer ce travail dans le cadre des recherches existantes et de mettre en lumière les approches, méthodologies et outils développés dans le domaine. En examinant les solutions actuelles, leurs forces et leurs limites, cette analyse vise à identifier les opportunités d’innovation pour proposer des solutions performantes, éthiques et adaptées au contexte complexe des violences en ligne et hors ligne.

# Etat de l’art sur les travaux de recherches pertinents liés au sujet

L’identification proactive des victimes potentielles et la prévention des violences en ligne sont des champs de recherche émergents où l’intelligence artificielle joue un rôle clé. Cette section se base sur deux travaux majeurs qui abordent, directement ou indirectement, ces thématiques. Ces études offrent des méthodologies et des perspectives complémentaires, allant de l’analyse de comportements en ligne à l’exploration des discours publics autour de la violence. Nous les présentons dans un ordre de pertinence, du plus directement applicable au domaine étudié au moins directement applicable mais inspirant.

## Approches multimodales pour la détection d’anomalies sur les réseaux sociaux numériques

L’étude intitulée "Approches multimodales pour la détection d'anomalies sur les réseaux sociaux numériques" explore des méthodes innovantes pour identifier des comportements inhabituels ou nuisibles sur les plateformes sociales. Ce travail s’appuie sur l’utilisation de données multimodales, combinant des informations textuelles, des graphes sociaux et des métadonnées, pour proposer des approches semi-supervisées et non supervisées adaptées à l’analyse des grandes quantités de données produites sur les réseaux sociaux.

### Objectifs et Methodologie

L’objectif principal est d’identifier les utilisateurs présentant des comportements atypiques, tels que les comptes malveillants, les bots ou les individus diffusant des contenus inappropriés. L’approche multimodale adoptée permet d’exploiter différentes sources d’information pour enrichir l’analyse et améliorer la précision des détections. Les messages sont analysés à l’aide de techniques de traitement automatique du langage naturel (NLP), pour détecter des discours violents, haineux ou trompeurs. Les graphes sociaux, représentant les relations entre les utilisateurs, sont étudiés pour repérer des anomalies dans les structures communautaires, tandis que les métadonnées, comme la fréquence des messages ou les horaires de publication, apportent des éléments supplémentaires pour caractériser les comportements. Ces différentes modalités sont ensuite fusionnées afin d’obtenir une vue globale et cohérente des anomalies détectées.

Les méthodologies employées dans cette étude incluent l’analyse des graphes à l’aide d’algorithmes tels que Louvain ou Label Propagation, permettant de détecter des communautés et d’analyser les interactions sociales. En parallèle, des algorithmes d’apprentissage non supervisé, comme Isolation Forest et Autoencoders, sont utilisés pour identifier des schémas comportementaux anormaux. Ces approches, qui ne dépendent pas exclusivement de données annotées, sont particulièrement adaptées dans un contexte où les données labellisées sont souvent limitées.

### Resultats

Les résultats de cette étude montrent que l’approche multimodale améliore considérablement les performances de détection par rapport à l’utilisation isolée d’une seule modalité. Par exemple, l’analyse des graphes seule peut ne pas identifier certains comportements subtils qui deviennent détectables grâce à l’ajout de données textuelles ou de métadonnées.

De plus, les méthodes développées permettent de détecter des comportements nuisibles existants tout en identifiant des signaux faibles qui peuvent indiquer une activité anormale émergente. Enfin, ces approches sont conçues pour être évolutives, rendant leur application pratique pour des plateformes sociales comptant des millions d’utilisateurs.

Cette étude offre une contribution significative au domaine de la détection d’anomalies sur les réseaux sociaux, en démontrant l’efficacité et la pertinence de combiner plusieurs types de données pour mieux comprendre et anticiper les comportements en ligne.

## Public Sentiment and Discourse on Domestic Violence during the COVID-19 Pandemic in Australia

### Objectifs

L’étude intitulée "Public Sentiment and Discourse on Domestic Violence During the COVID-19 Pandemic in Australia: Analysis of Social Media Posts" analyse les sentiments et les discours publics relatifs à la violence domestique pendant la pandémie de COVID-19. S’appuyant sur les publications des réseaux sociaux, cette recherche met en lumière les dynamiques sociales et émotionnelles en période de crise. Bien qu’elle ne cible pas directement la détection de comportements nuisibles, elle fournit des outils méthodologiques et des résultats précieux pour mieux comprendre comment les réseaux sociaux peuvent refléter les réalités de la violence domestique.

Les objectifs principaux de l’étude sont multiples. Tout d’abord, elle cherche à analyser les sentiments dominants exprimés par les utilisateurs, comme la peur, la colère ou l’empathie, afin de comprendre comment ces émotions évoluent dans un contexte de pandémie mondiale. Ensuite, elle exploite les réseaux sociaux, tels que Twitter et Facebook, en tant que baromètres sociaux, capturant les tendances émergentes, les préoccupations collectives et les attitudes changeantes vis-à-vis de la violence domestique. Enfin, elle utilise des outils de traitement automatique du langage naturel (NLP) pour extraire des thèmes récurrents, identifier des mots-clés associés à la violence domestique, et détecter des tendances dans les discussions en ligne.

### Méthodologie

Les publications sont analysées à l'aide de techniques de traitement automatique du langage naturel (NLP) pour extraire des thèmes, détecter les tendances, et identifier des mots-clés récurrents associés à la violence domestique.

#### Collection et prétraitement des données :

Les messages provenant de plateformes sociales ont été collectés à l'aide de mots-clés spécifiques liés à la violence domestique. Les données ont ensuite été nettoyées pour supprimer les doublons, les messages hors sujet et les données bruitées (émoticônes ou abréviations excessives).

#### Analyse de sentiment :

Des outils comme VADER ou des modèles basés sur le machine Learning ont été employés pour attribuer des scores de sentiment (positif, négatif ou neutre) aux messages.

Des algorithmes de modélisation thématique tels que LDA (Latent Dirichlet Allocation) ont été utilisés pour extraire les thèmes principaux des discussions, par exemple :

* Augmentation de la violence domestique pendant les confinements.
* Demande accrue de soutien pour les victimes.
* Sentiments de peur et d'isolement.

### Resultats

**Efficacité de la détection multimodale** :

La combinaison des modalités améliore significativement les performances de détection par rapport à l’analyse isolée de chaque modalité. Par exemple :

* L’analyse des graphes seule peut manquer certains comportements subtils détectables via le texte.
* Les métadonnées enrichissent l’analyse textuelle en fournissant des informations sur la temporalité et les schémas d’activité.

**Changement des tendances sentimentales :**

Les publications montrent une augmentation des sentiments négatifs (peur, colère) pendant les périodes de confinement strict. En parallèle, des expressions de soutien et d'empathie envers les victimes sont également apparues.

**Visualisation**

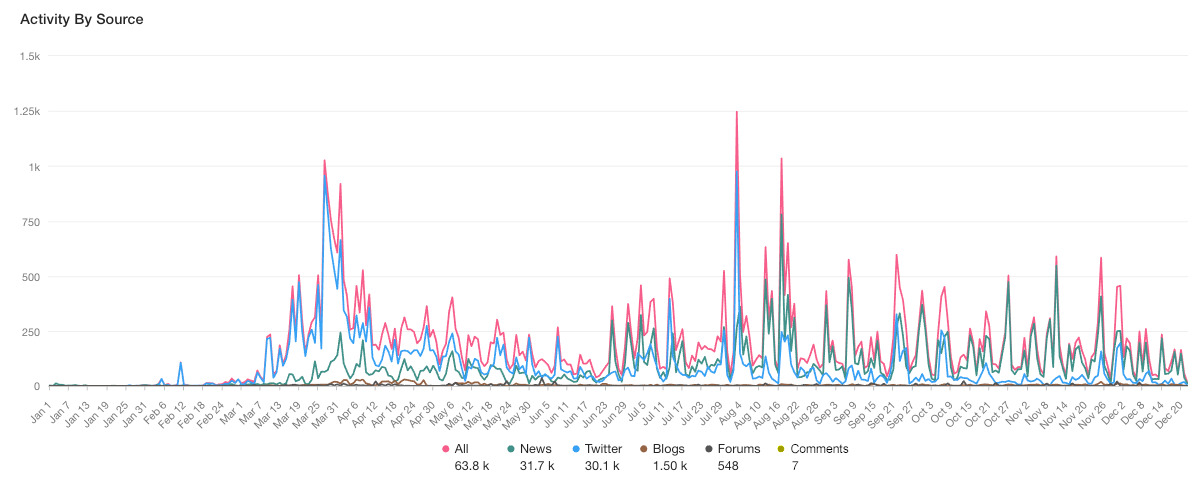
****

Figure 1: Resultat de 'Public Sentiment and Discourse on Domestic Violence During the COVID'

Dans l'échantillon final de données, on comptait au total 63 800 publications provenant des réseaux sociaux et des médias d'information (Figure 1). Parmi ces publications, notre analyse a révélé que la violence domestique était mentionnée en moyenne 179 fois par jour. Cela comprenait 30 100 tweets, 31 700 articles de presse, 1 500 billets de blog, 548 messages de forum et 7 commentaires (publiés sur des sites d’actualités et des blogs).

# Identification des zones manquantes liés au sujet

## Limites identifiées

### Projet 1

* Dépendance aux données disponibles :

Les performances des modèles dépendent fortement de la qualité et de la quantité des données disponibles, notamment les graphes sociaux, qui ne sont pas toujours accessibles pour des raisons de confidentialité.

* Biais dans les données multimodales :

Certaines modalités (comme les métadonnées) peuvent introduire des biais, par exemple en privilégiant des comportements associés à des fuseaux horaires spécifiques.

* Manque de focalisation sur les victimes :

Bien que les méthodes soient efficaces pour détecter des comportements nuisibles, elles se concentrent principalement sur les agresseurs ou les anomalies structurelles sans aborder la question des victimes potentielles.

### Projet 2

* Biais des données :

Les données des réseaux sociaux ne représentent pas toujours l'ensemble de la population. Les utilisateurs actifs sur ces plateformes sont souvent jeunes ou technologiquement connectés, ce qui peut biaiser les résultats.

* Langage informel et ambiguïté :

Les publications sur les réseaux sociaux contiennent souvent des messages implicites ou des termes subjectifs, ce qui complique leur analyse automatique.

* Absence de suivi longitudinal :

Bien que l'étude capte un instantané des sentiments publics, elle ne permet pas d’analyser les évolutions à long terme des discours sur la violence domestique.

## Opportunités d’innovation

### Projet 1

* Innovation vers des approches centrées sur les victimes :

Les techniques développées pourraient être étendues pour inclure des indicateurs de vulnérabilité, permettant de détecter non seulement les comportements nuisibles, mais aussi les utilisateurs susceptibles d’être affectés.

* Amélioration des modèles non supervisés :

Intégrer des mécanismes d’auto-apprentissage basés sur les retours des utilisateurs pour affiner les modèles en continu.

* Applications interdisciplinaires :

Les approches multimodales pourraient être appliquées à d’autres domaines, comme la détection de fausses nouvelles ou la prévention des discours haineux.

### Projet 2

La combinaison de l’analyse thématique (pour comprendre les discours) et de l’analyse de sentiment (pour capter les émotions) constitue une méthode puissante pour étudier des problématiques sociales complexes.

* Focus potentiel sur les victimes :

Les méthodes utilisées pourraient être adaptées pour mieux identifier les groupes vulnérables à la violence domestique et pour développer des modèles prédictifs permettant d'intervenir de manière proactive.

* Approches hybrides pour enrichir l'analyse :

L’intégration de données multimodales (images, vidéos) et de métadonnées pourrait compléter l’analyse textuelle pour capter des signaux faibles indiquant un risque accru pour certaines populations.

Ces deux travaux offrent des approches complémentaires qui peuvent inspirer l’identification des victimes potentielles et la prévention des violences en ligne grâce à l’IA. Les approches multimodales présentées par Ben Chaabene sont directement applicables pour détecter des anomalies comportementales, mais elles nécessitent une adaptation pour cibler les signaux de vulnérabilité. De leur côté, les méthodologies d’analyse de sentiment et de discours utilisées dans l’étude sur la violence domestique permettent d’explorer les états émotionnels des utilisateurs, une dimension essentielle pour anticiper les situations de victimisation.

En combinant ces deux perspectives, il devient possible de développer des modèles intégrant à la fois des signaux sociaux (par exemple, interactions inhabituelles) et émotionnels (sentiments exprimés dans les messages) pour mieux identifier les utilisateurs vulnérables et prévenir les violences en ligne de manière proactive.

## ****Pertinence du Sujet****

La détection proactive des signes de violence est un enjeu crucial dans le contexte numérique actuel, où des millions d’utilisateurs partagent quotidiennement des publications pouvant inclure des appels à l’aide.

Cependant, les lacunes identifiées telles que le manque d’analyse contextuelle, l’absence d’adaptabilité linguistique et culturelle, et la limitation des approches actuelles à des modèles uniquement réactifs soulignent la nécessité d’un travail approfondi. Mon projet s’inscrit dans cette perspective et ambitionne de contribuer à la résolution de ces problématiques.

Pour répondre à ces enjeux et combler les lacunes identifiées, le prochain chapitre présentera les choix méthodologiques effectués, depuis la collecte des données jusqu’au développement des modèles. Cette méthodologie détaillera les étapes permettant de transformer les objectifs théoriques en solutions concrètes et applicables.

# 

# CHAPITRE II : ANALYSE DE BESOIN ET METHODOLOGIE

# Spécification des Besoins et des Données

## Analyse et Spécification des Besoins

L’objectif principal est de concevoir un système capable d’identifier les publications en ligne qui manifestent, de manière explicite ou implicite, des indices de violence, qu'elle soit physique, morale ou numérique. Cette solution vise à apporter une aide proactive aux victimes potentielles.

* 1. Besoin fonctionnel :

Mettre en place un outil automatisé capable de détecter les messages impliquant des situations de violence sur les réseaux sociaux.

#### Besoin technique :

* Collecter des données textuelles fiables provenant de plateformes sociales telles que Twitter et Reddit, en tenant compte des limites des API.
* Utiliser des algorithmes d’apprentissage automatique et profond pour traiter les textes.
* S’assurer que les résultats soient compréhensibles pour les utilisateurs finaux, notamment les ONG et autres parties prenantes.

## Sources de Données

Pour mener à bien ce projet, les données textuelles nécessaires proviennent principalement de plateformes de réseaux sociaux, qui offrent une richesse d'expressions personnelles, d'expériences partagées, et parfois de signaux de détresse implicites ou explicites. Les principales sources identifiées sont :

### Twitter

* + 1. Description :

Twitter est une plateforme majeure pour la communication publique en temps réel. Les utilisateurs partagent des réflexions, des expériences ou des réactions, souvent en lien avec des sujets d'actualité ou des préoccupations personnelles.

#### Mode de collecte :

* Utilisation de mots-clés spécifiques tels que : domestic violence, abuse, harassment, toxic relationship, cyberbullying.
* Exploitation de l'API Twitter ou de librairies comme Tweepy et snscrape pour extraire des tweets.

#### Requêtes adaptées :

* Filtrage par langue (ex. : anglais ou français).
* Définition d'une période temporelle pour obtenir des données récentes et pertinentes.

### Reddit

* + 1. Description : Reddit, une plateforme de forums de discussion, permet aux utilisateurs de partager leurs expériences et de demander des conseils au sein de communautés spécifiques.

#### Mode de collecte :

* Extraction des discussions des sous-reddits pertinents, tels que :
* r/relationships : Discussions sur les relations, souvent marquées par des témoignages de conflits ou de violences.
* r/abuse : Un espace dédié aux personnes partageant des expériences de violences ou d'abus.
* Utilisation de PRAW (Python Reddit API Wrapper) pour collecter les données.

#### Requêtes adaptées :

* Cibler les publications contenant des mots-clés spécifiques.
* Récupérer à la fois les publications initiales et les commentaires pour un contexte complet.

### Autres sources

Blogs et plateformes collaboratives :

* + 1. Description : Des blogs personnels ou des plateformes collaboratives (ex. : Medium, forums spécialisés) où les utilisateurs partagent des récits détaillés de leurs expériences.

#### Mode de collecte :

* Scraping des articles pertinents à l'aide d'outils comme BeautifulSoup ou Scrapy.
* Recherche manuelle pour identifier les plateformes populaires dans des contextes spécifiques (linguistiques ou culturels).

*Exemples* :

Articles décrivant des expériences personnelles sur des blogs d’auteurs spécialisés dans les questions de violences ou de droits humains.

* + 1. Approche générale pour la collecte :
* Filtrage des données : Mise en place de filtres pour limiter les publications non pertinentes ou hors contexte.
* Préservation des métadonnées : Récupération des informations utiles (date de publication, auteur, etc.) pour enrichir l’analyse.
* Gestion des limitations : Respect des limites imposées par les API des plateformes pour éviter les blocages.

Ces sources fourniront un corpus diversifié et représentatif, essentiel pour entraîner un modèle capable de détecter et d’interpréter les signaux de violence dans des contextes variés.

## Processus de Collecte

### Utilisation des API dédiées

#### Twitter API :

* Description : API officielle de Twitter pour récupérer les tweets correspondant à des mots-clés spécifiques.
* Actions :
* Configuration des clés d'authentification (API Key, API Secret, Access Token).
* Construction de requêtes en fonction des mots-clés (ex. : domestic violence, abuse).
* Application de filtres pour limiter les résultats par langue, localisation, et date.
* Outil : Utilisation de librairies Python telles que Tweepy ou snscrape pour interagir avec l'API.

#### Reddit API :

* Description : API officielle de Reddit pour collecter des données à partir de sous-reddits pertinents.
* Actions :
* Configuration avec PRAW (Python Reddit API Wrapper).
* Récupération des publications et des commentaires des sous-reddits ciblés (ex. : r/relationships, r/abuse).
* Application de filtres sur les mots-clés et les dates.

### Automatisation avec des scripts Python

#### Développement de scripts personnalisés :

* Écriture de scripts Python pour collecter automatiquement les données.
* Intégration des librairies (ex. : Tweepy pour Twitter, PRAW pour Reddit).
* Automatisation de l'extraction via des tâches programmées (ex. : utiliser cron sous Linux ou Task Scheduler sous Windows).
* Gestion du pipeline de collecte :
* Extraction des données.
* Transformation pour uniformiser les formats (JSON, CSV).
* Stockage des résultats dans des fichiers structurés ou une base de données (ex. : SQLite, PostgreSQL).

### Gestion des limitations d’accès

#### Respect des quotas :

* Limitation du nombre de requêtes en fonction des règles des API (Twitter API impose des restrictions sur le nombre de requêtes par minute).
* Implémentation d’attentes (sleep time) pour éviter les blocages.

#### Gestion des erreurs :

* Détection et gestion des erreurs courantes (ex. : Rate Limit Exceeded, Connection Timeout).
* Reconnexion automatique en cas d’échec de requêtes.
* Utilisation d’alternatives :
* En cas de dépassement des quotas, exploitation de librairies comme snscrape pour Twitter, qui ne nécessite pas de clés d’API.

### Enrichissement des données

* + 1. Diversification des mots-clés :
* Utilisation d’une liste exhaustive de termes liés aux violences (domestic violence, toxic relationship, gaslighting, cyberbullying).
* Recherche de synonymes et d’expressions informelles pour améliorer la couverture.

#### Plages temporelles représentatives :

* Définition de périodes spécifiques (ex. : données récentes ou couvrant des événements particuliers).
* Récupération de données sur plusieurs mois ou années pour observer des tendances.

Résumé du Processus

Authentification via les API des plateformes.

Configuration des requêtes pour extraire les données pertinentes.

Automatisation et stockage des résultats.

Gestion proactive des limitations techniques.

Enrichissement des données pour assurer leur qualité et leur pertinence.

Ce processus garantit une collecte de données robuste, adaptée aux objectifs du projet et conforme aux contraintes techniques et éthiques.

## 2. Pré-processing des Données

Le pré-traitement des données est une étape essentielle pour garantir la qualité et la pertinence des textes analysés. Les données brutes collectées depuis les réseaux sociaux sont souvent bruitées et nécessitent un nettoyage rigoureux et une transformation adaptée pour une analyse efficace. Voici les étapes détaillées :

a. Nettoyage

L’objectif du nettoyage est de supprimer les éléments inutiles ou non significatifs pour améliorer la clarté et la cohérence des données textuelles.

* Étapes :
* Suppression des caractères spéciaux (ex. : #, @, $, %, &, \*, etc.).
* Retrait des liens URL (ex. : https://t.co/xyz).
* Suppression des emojis et autres symboles non textuels.
* Conversion du texte en minuscule pour uniformiser les données.
* Suppression des espaces multiples ou des lignes vides.

c. Lemmatisation et Stemming

La Lemmatisation permet de :

* Réduire des mots à leur forme canonique (ex. : marchant devient marcher).
* Préserver le sens des mots pour une analyse contextuelle plus précise.

Pour ce faire, l’utilisation de librairies comme spaCy ou NLTK est requises.

Le Stemming permet de :

* Réduire des mots à leurs racines (ex. : marchant, marche, marché deviennent march).
* Il est plus rapide mais parfois moins précis que la lemmatisation.
* Choix : Utiliser la lemmatisation si le contexte et la précision sont importants, sinon le stemming pour des besoins rapides.

d. Filtrage

L’Objectif du filtrage est de supprimer les éléments non pertinents ou répétitifs.

* Étapes :
* Suppression des stopwords (mots fréquents mais sans signification propre, comme le, et, ou, de).
* Filtrage des mots trop courts ou trop longs (ex. : conserver uniquement les mots de 3 à 15 caractères).
* Suppression des doublons dans les données.

b. Tokenisation

Le but de la tokenisation est de découper les textes en unités manipulables comme des mots (tokens) ou des phrases.

* Étapes :
* Utilisation de librairies comme NLTK, spaCy, ou transformers pour extraire les mots.
* Prise en compte des ponctuations et des séparateurs pour éviter une découpe incorrecte.

e. Vectorisation

Cette etape cruciale consiste à transformer les textes en représentations numériques exploitables par des modèles d’apprentissage.

* Méthodes :

TF-IDF (Term Frequency-Inverse Document Frequency) :

Pondération basée sur la fréquence des mots dans un document et leur rareté dans le corpus.

Idéal pour des modèles simples ou classiques.

Word2Vec ou GloVe :

Génération de vecteurs basés sur les relations sémantiques entre les mots.

Capturent mieux les contextes, utile pour des modèles avancés.

Embeddings BERT ou RoBERTa :

Représentations contextuelles riches générées par des modèles de type Transformers.

Conviennent pour des tâches complexes nécessitant une compréhension fine du contexte.

Résumé du Processus

Nettoyage : Élimination des éléments non pertinents pour clarifier les textes.

Tokenisation : Fractionnement des textes en unités manipulables.

Lemmatisation/Stemming : Normalisation des mots pour une analyse cohérente.

Filtrage : Suppression des mots inutiles et non significatifs.

Vectorisation : Conversion des textes en données numériques prêtes pour l’apprentissage.

Ce pipeline de pré-traitement garantit que les données textuelles sont adaptées et optimisées pour une analyse efficace par les algorithmes choisis.

I. Méthodologie

1. Choix des Algorithmes et Modèles

Pour répondre aux objectifs du projet, une combinaison d'algorithmes classiques et de modèles d'apprentissage profond est envisagée. Ces choix permettent de capturer à la fois les caractéristiques simples et contextuelles des textes.

a. Apprentissage Automatique Classique

Les algorithmes classiques seront utilisés comme base de référence pour comparer les performances des modèles avancés :

Logistic Regression (Régression Logistique)

Pourquoi : Simplicité et efficacité pour des données bien structurées avec des classes binaires ou multiclasses.

Support Vector Machines (SVM)

Pourquoi : Adapté pour la séparation linéaire ou non linéaire des données textuelles à l’aide de kernels.

Random Forest

Pourquoi : Modèle robuste qui combine plusieurs arbres de décision pour réduire l'overfitting.

K-Nearest Neighbors (KNN)

Pourquoi : Simple et efficace pour des données de faible dimension.

XGBoost

Pourquoi : Modèle puissant basé sur le boosting, souvent performant pour les données textuelles après vectorisation.

Multilayer Perceptron (MLP)

Pourquoi : Réseau de neurones simple capable de capturer des relations non linéaires.

b. Apprentissage Profond (Deep Learning)

Pour capturer les relations contextuelles et sémantiques plus complexes dans les textes :

Long Short-Term Memory (LSTM)

Pourquoi : Capacité à modéliser les dépendances temporelles dans les séquences textuelles, utile pour les phrases longues.

Exemple d’application : Analyse des mots dans le contexte d’une phrase pour détecter des signaux implicites.

Transformers (BERT, RoBERTa)

Pourquoi : Modèles d’état de l’art pour extraire des représentations riches et contextuelles du texte.

BERT (Bidirectional Encoder Representations from Transformers) :

Capture les relations bidirectionnelles dans un texte, par exemple :

Texte : "I feel trapped and need help urgently."

Contexte : Les mots "trapped" et "help" sont associés à un signal de détresse.

RoBERTa (Robustly Optimized BERT) :

Une version optimisée de BERT, adaptée pour des tâches complexes nécessitant une meilleure gestion des grands ensembles de données.

Justification des Choix

Algorithmes Classiques :

Rapides à entraîner et à évaluer.

Requiert moins de ressources informatiques, utile pour des itérations rapides.

Permettent une meilleure interprétabilité des résultats.

Apprentissage Profond :

Requis pour modéliser les nuances sémantiques et les dépendances complexes dans des données non structurées.

Utilisation des pré-entraînements sur de grandes bases de données textuelles (ex. : BERT) pour transférer les connaissances.

L’utilisation combinée de ces méthodes garantit une analyse complète, depuis les approches simples et rapides à implémenter jusqu’aux modèles avancés nécessitant une puissance de calcul importante pour des performances élevées.

# CHAPITRE III : IMPLEMENTATION

# Resultats

## Analyse exploratoire de donnees (EDA)

Objectifs :

Identifier les tendances générales dans les données : Comprendre la répartition des données collectées selon des critères tels que les mots-clés, les langues, ou les plateformes (ex. : Twitter, Reddit).

Détecter les anomalies ou les valeurs manquantes : Identifier les incohérences dans les données, comme des posts incomplets, des doublons ou des champs vides.

Explorer les fréquences des mots ou des thèmes liés aux violences : Analyser la prévalence des termes ou expressions associés à des situations de détresse ou de violence.

Exemples d’Analyse :

Histogrammes pour visualiser les fréquences des mots-clés :

Exemple : Un histogramme illustrant la fréquence des mots comme "abuse", "violence", et "harassment" montre que "abuse" est le mot le plus utilisé parmi les données collectées.

Code Python (avec matplotlib ou seaborn) pour générer un histogramme :

HISTOGRAME

Nuages de mots :

Visualisation des mots les plus fréquents associés aux publications collectées.

Exemple : Le nuage de mots montre que les termes "help", "violence", et "support" sont fréquemment associés.

NUAGE DE MOTS

L’EDA fournit une compréhension fondamentale des données collectées, identifiant les patterns significatifs et préparant le terrain pour les étapes suivantes de prétraitement et de modélisation.

## Resultats des modeles

Après l’entraînement des modèles, leurs performances ont été évaluées sur un jeu de test afin de comparer leur efficacité dans la détection des messages liés à des situations potentielles de violence. Les résultats obtenus permettent de distinguer les performances des modèles classiques et des modèles d’apprentissage profond.

Modèles Classiques :

-XgBoost et KNN sont les modeles classiques avec de moins bonnes performances pour une precision respectives de 87% et 92%, un rappel de 83% et 90%

-Logistic Regression et Support Vector Machine (SVM) ont atteint une précision de 91 % contre un rappel de 85% pour logistic Regression et 84% pour SVM, montrant une bonne capacité à détecter les messages pertinents, bien que certains cas aient été manqués.

-Decision Tree et MLP se sont mieux comportés, avec une précision égale a 95% , un rappel respectif de 81 % et , indiquant une meilleure gestion des faux négatifs.

Random Forest a obtenu les meilleures performances parmi les modèles classiques, avec une précision de 96 % et un rappel de 92 %, reflétant une capacité solide à classer correctement les messages.

Modèles d’Apprentissage Profond :

LSTM (Long Short-Term Memory) a fourni les résultats les moins satisfaisants, avec une précision de 70% et un rappel de %, démontrant sa capacité à capturer le contexte des messages.

BERT a surpassé les modèles classiques et LSTM, avec une précision de 97 % et un rappel de 97 %, grâce à sa compréhension approfondie des relations textuelles.

Visualisation :

Matrices de confusion :

Ces matrices montrent clairement les erreurs de classification pour chaque modèle. Par exemple, les faux positifs (messages incorrectement classés comme 'Potential danger') et les faux négatifs (messages pertinents manqués) sont visualisés pour identifier les limites spécifiques de chaque algorithme.

Courbes ROC (Receiver Operating Characteristic) :

Les courbes ROC ont permis de comparer les performances globales des modèles en termes de compromis entre le taux de vrais positifs et de faux positifs. RoBERTa a présenté la plus grande surface sous la courbe (AUC), confirmant sa supériorité.

Synthèse :

Les résultats montrent que les modèles d’apprentissage profond, en particulier BERT et Random Forest, surpassent nettement les modèles classiques en termes de précision et de rappel. Ces performances témoignent de l’importance des approches modernes basées sur des transformers pour traiter des données textuelles complexes et sensibles comme celles associées à la détection des violences en ligne.

## Interpretation des modeles

## Comparaison avec la littérature

## Discussion