|  |
| --- |
| **Membres du Groupe 2 :** Liban Souleiman Abdillahi & Nasteho Mahamoud Ibrahim |
| **Soumission de Proposition d'Idée** |

**1. Idée du Projet :**

**Titre : Détection automatique du cyberharcèlement sur les réseaux sociaux à l’aide de l’intelligence artificielle**

**Problème à résoudre :** Le cyberharcèlement est un phénomène en constante augmentation, particulièrement parmi les jeunes utilisateurs des réseaux sociaux. Il peut entraîner des conséquences graves sur la santé mentale, allant de l’anxiété à la dépression, voire au suicide.

**Objectif spécifique :** Développer un modèle d’intelligence artificielle capable d’identifier automatiquement les messages contenant des propos harcelants sur les plateformes sociales, afin de permettre une intervention précoce et de protéger les utilisateurs vulnérables.

**2. Pertinence par rapport aux Objectifs de Développement Durable (ODD):**

Ce projet s’aligne principalement sur l’ODD 3 : Bonne santé et bien-être, en contribuant à la prévention des troubles mentaux liés au cyberharcèlement. Il soutient également l’ODD 16 : Paix, justice et institutions efficaces, en promouvant des environnements en ligne sûrs et inclusifs. Elle contribue aussi a l’ODD 4 (Education de Qualité) en protégeant les élèves et les étudiants des effets négatifs du harcèlement en ligne.

**3. Exemples de Littérature :**

* “**Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT**”

Cet article présente une approche combinant plusieurs modèles d’apprentissage profond, notamment une version améliorée de BERT, pour détecter le cyberharcèlement sur les réseaux sociaux. Précision atteinte : 97,4 % sur Twitter.

* “**Cyberbullying Detection System on Social Media Using Supervised Machine Learning**”

Les auteurs utilisent des techniques d’apprentissage supervisé (SVM, régression logistique) avec des données annotées de Twitter pour construire un modèle de détection efficace.

**4. Description de Vos Données :**

Les données utilisées pour ce projet proviennent de sources publiques telles que Kaggle, Twitter et d'autres plateformes de recherche ouvertes. Parmi les jeux de données les plus pertinents figurent :

Cyberbullying Detection Dataset (Kaggle) : environ 30 000 à 50 000 messages annotés selon leur caractère harcelant ou non.

Hate Speech and Offensive Language Dataset : plus de 24 000 tweets classés en trois catégories : harcèlement, langage offensant, et neutre.

Ces données sont généralement au format CSV et contiennent des colonnes telles que le texte du message, l’étiquette (harcèlement ou non), et parfois des métadonnées (date, utilisateur, etc.).

**Étapes de prétraitement prévues :**

Nettoyage du texte : suppression de la ponctuation, des caractères spéciaux, des liens, des hashtags et des mentions (@).

**Conversion en minuscules pour uniformiser les données.**

Suppression des stop words (mots vides comme "le", "et", "de", etc.).

**Lemmatisation :** réduction des mots à leur forme de base (ex. : "insultes" → "insulte").

**Tokenisation :** découpage du texte en unités (mots ou sous-mots) pour l’analyse par le modèle.

**Encodage :** transformation des textes en vecteurs numériques via des modèles comme BERT.

Ces étapes sont essentielles pour garantir la qualité des données en entrée du modèle d’apprentissage profond, et ainsi améliorer la précision de la détection du cyberharcèlement.

**5. Approche (Apprentissage Machine ou Apprentissage Profond (Deep Learning)) :**

L’approche retenue est l’apprentissage profond, notamment les modèles de type BERT (pré-entraînés), adaptés pour capter le contexte sémantique des messages. Cela permet une détection plus précise des propos ambigus ou dissimulés.