# Partie 3 : Concept et Plan de Mise en Œuvre du Projet

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

Membres de l’équipe :

1. Liban Souleiman Abdillahi

2. Nasteho Mahamoud Ibrahim

## Note de Concept

### 1. Vue d'ensemble du projet

Le projet vise à développer un système intelligent capable de détecter automatiquement les propos harcelants sur les réseaux sociaux. Ce phénomène, en constante augmentation, affecte particulièrement les jeunes et peut avoir des conséquences graves sur la santé mentale. Notre solution s’aligne sur plusieurs Objectifs de Développement Durable (ODD), notamment l’ODD 3 (Bonne santé et bien-être), l’ODD 4 (Éducation de qualité) et l’ODD 16 (Paix, justice et institutions efficaces). En permettant une détection précoce du cyberharcèlement, notre système contribuera à créer un environnement numérique plus sûr et inclusif.

### 2. Objectifs

L’objectif principal est de concevoir un modèle d’intelligence artificielle capable d’identifier automatiquement les messages contenant des propos harcelants. Les objectifs spécifiques incluent :  
- Collecter et prétraiter des données textuelles issues des réseaux sociaux.  
- Entraîner un modèle de type BERT pour la classification des messages.  
- Évaluer la performance du modèle à l’aide de métriques standards (précision, rappel, F1-score).  
- Déployer une interface simple permettant d’alerter les modérateurs ou utilisateurs.

### 3. Contexte

Le cyberharcèlement est un problème mondial qui touche toutes les tranches d’âge. De nombreuses initiatives ont vu le jour, mais la plupart reposent sur des systèmes de modération manuelle ou semi-automatique. Ces approches sont limitées par leur scalabilité et leur subjectivité. L’intelligence artificielle, et plus particulièrement l’apprentissage profond, offre une alternative prometteuse grâce à sa capacité à traiter de grandes quantités de données et à détecter des schémas complexes. Notre projet s’inscrit dans cette dynamique en proposant une solution automatisée, précise et évolutive.

### 4. Méthodologie

Nous utiliserons une approche basée sur l’apprentissage profond, en particulier les modèles de type Transformer comme BERT (Bidirectional Encoder Representations from Transformers). Le pipeline comprendra les étapes suivantes :  
- Prétraitement des données textuelles (nettoyage, tokenisation, lemmatisation).  
- Vectorisation des textes à l’aide de BERT.  
- Entraînement d’un classifieur supervisé (fine-tuning de BERT).  
- Évaluation du modèle sur un jeu de test.  
- Déploiement via une interface web ou API REST.

### 5. Diagramme de conception architecturale

L’architecture du système comprend les composants suivants :  
- Collecteur de données : récupère les messages depuis les réseaux sociaux ou jeux de données publics.  
- Module de prétraitement : nettoie et prépare les textes.  
- Modèle BERT : encode les textes et prédit les étiquettes (harcèlement ou non).  
- Interface utilisateur : permet de visualiser les résultats et d’alerter les modérateurs.  
- Base de données : stocke les messages analysés et les prédictions.  
Un diagramme illustrant ces composants pourra être ajouté dans la version finale du rapport.

### 6. Sources de données

Nous utiliserons des jeux de données publics tels que ceux disponibles sur Kaggle ou Twitter, contenant des milliers de messages annotés. Ces données sont pertinentes car elles reflètent des cas réels de cyberharcèlement. Le prétraitement inclura la suppression des caractères spéciaux, la normalisation, la tokenisation, et la lemmatisation pour préparer les textes à l’analyse.

### 7. Revue de littérature

Les travaux existants montrent que les modèles basés sur BERT surpassent les approches classiques pour la détection du cyberharcèlement. Par exemple, Muralidhar et al. (2024) ont obtenu une précision de 94 % sur des données Twitter. D’autres études, comme celles de Khafajeh (2024) et Aggarwal & Mahajan (2024), ont démontré l’efficacité des architectures hybrides combinant BERT avec des classifieurs comme SVM. Notre projet s’appuie sur ces résultats pour proposer une solution robuste et adaptée au contexte local.

## Plan de Mise en Œuvre

### 1. Technology Stack

Langages : Python  
Bibliothèques : Transformers (Hugging Face), Scikit-learn, Pandas, Numpy, Matplotlib  
Frameworks : PyTorch ou TensorFlow  
Outils : Jupyter Notebook, Google Colab, GitHub, Streamlit (pour l’interface), Heroku (pour le déploiement)

### 2. Chronologie

Le projet sera réalisé sur une période de 12 semaines :  
- Semaine 1-2 : Collecte et exploration des données  
- Semaine 3-4 : Prétraitement et annotation  
- Semaine 5-7 : Entraînement du modèle  
- Semaine 8-9 : Évaluation et ajustement  
- Semaine 10-11 : Développement de l’interface  
- Semaine 12 : Déploiement et documentation  
Répartition des tâches :  
- Liban : collecte des données, entraînement du modèle, documentation  
- Nasteho : prétraitement, développement de l’interface, évaluation

### 3. Étapes importantes

- Collecte de données annotées  
- Prétraitement complet des textes  
- Fine-tuning du modèle BERT  
- Évaluation avec précision > 90 %  
- Déploiement d’une interface fonctionnelle

### 4. Défis et atténuations

Défis potentiels :  
- Qualité des données : nous utiliserons plusieurs sources pour diversifier les exemples.  
- Performance du modèle : nous testerons plusieurs variantes de BERT et ajusterons les hyperparamètres.  
- Contraintes techniques : recours à Google Colab pour l’entraînement sur GPU, et à des outils open source pour le déploiement.

### 5. Considérations éthiques

Le projet respecte la confidentialité des données en anonymisant les messages. Nous veillerons à éviter les biais algorithmiques en équilibrant les classes et en testant le modèle sur des données variées. L’objectif est de protéger les utilisateurs sans porter atteinte à leur vie privée.

### 6. Références

- Muralidhar, A. (2024). BERT-Based Detection of Cyberbullying in Online Texts.  
- Khafajeh, H. (2024). Cyberbullying Detection in Social Networks Using Deep Learning.  
- Aggarwal, R., & Mahajan, A. (2024). Shielding Social Media: BERT and SVM Unite.  
- Hasan, M. et al. (2023). A Review on Deep-Learning-Based Cyberbullying Detection.  
- Gomez, R. et al. (2022). Curating Cyberbullying Datasets: a Human-AI Collaborative Approach.  
- Kumar, A. et al. (2024). Bias and Cyberbullying Detection Using Transformers.  
- Rahman-Laskar, A. et al. (2024). Cyberbullying Detection in a Multi-classification Codemixed Dataset.  
- Philipo, J. et al. (2024). Cyberbullying Detection: Exploring Datasets, Technologies, and Approaches.  
- Saini, R. et al. (2023). Cyberbullying Detection on Social Networking Sites Using BERT.  
- Aravindhan, R. (2025). Deep Learning Algorithms for Robust Cyberbullying Detection.  
- Altayeva, A. et al. (2024). Hybrid Deep Learning Architecture for Cyberbullying Detection.  
- Balakrishnan, V., & Kaity, M. (2023). Cyberbullying detection and machine learning: A systematic literature review.