Détection de Messages Haineux dans le Contexte Camerounais : Cas de la détection de messages haineux sur WhatsApp

Étudiant Data Science UY1 - Master 1

26 juin 2025

Membres de l'équipe	**Matricule**
NDONKOU FRANCK	21T2254
TCHIAZE FOUOSSO ROMERO	21T2474
FOTSING ENGOULOU SIMON GAETAN	21Q2024
Encadrant	Dr Thomas MESSI NGUELE

Table des matières

1 Introduction

1.1 Contexte du projet

Le Cameroun, avec sa diversité linguistique et culturelle exceptionnelle, présente un défi unique dans le domaine du traitement automatique du langage naturel. Le pays compte plus de 280 langues locales, avec le français et l'anglais comme langues officielles, créant un environnement linguistique complexe où se mélangent expressions locales, argot urbain, et langues véhiculaires comme le pidgin English.

Dans ce contexte multilingue, les plateformes numériques telles que WhatsApp sont devenues des espaces d'expression privilégiés, mais également des vecteurs potentiels de discours haineux. Les expressions haineuses dans le contexte camerounais présentent des spécificités culturelles et linguistiques que les modèles de détection standard ne parviennent pas à capturer efficacement.

1.2 Problématique

Les systèmes de détection automatique de discours haineux existants, principalement développés pour des langues européennes ou nord-américaines, présentent des limitations significatives lorsqu'ils sont appliqués au contexte camerounais : insuffisance lexicale pour reconnaître les expressions locales haineuses, manque de compréhension du contexte culturel, difficulté à traiter le mélange linguistique, et inadaptation à l'évolution rapide du vocabulaire des réseaux sociaux.

1.3 Motivation

Cette recherche vise à protéger les utilisateurs camerounais des contenus haineux spécifiques à leur contexte, développer des solutions NLP adaptées aux contextes africains multilingues, contribuer à un environnement numérique plus sain, et combler le gap dans la littérature scientifique sur la détection de hate speech en contexte africain.

1.4 Objectifs du projet

Objectif principal : Développer un modèle de détection de messages haineux spécifiquement adapté au contexte linguistique et culturel camerounais, et l'utiliser pour la détection des discours haineux sur WhatsApp.

Objectifs spécifiques : Constituer un corpus de données représentatif, adapter le modèle DistilCamemBERT par fine-tuning en deux phases, évaluer les performances sur des expressions locales, analyser les considérations éthiques, et créer une API de détection.

1.5 Méthodologie générale

Notre approche se base sur une méthodologie hybride en deux phases : adaptation au domaine via Masked Language Modeling (MLM) pour l'adaptation au vocabulaire camerounais, puis fine-tuning spécialisé pour la classification binaire (haineux/non-haineux).

1.6 Plan du document

Ce rapport s'articule autour de six sections principales. Après cette introduction, la section 2 présente une revue des solutions existantes en définissant les concepts fondamentaux et en positionnant notre approche. La section 3 détaille notre méthodologie, incluant la présentation des données, les méthodes choisies et les métriques d'évaluation. La section 4 expose l'implémentation, les résultats obtenus et leur discussion critique. La section 5 aborde les considérations éthiques essentielles à ce type de projet. Enfin, la section 6 conclut par un bilan global et les perspectives d'amélioration.

2 Revue des solutions existantes

2.1 Définition des concepts fondamentaux

Discours haineux (Hate Speech) : Toute forme de communication qui attaque, menace, ou incite à la violence ou à la discrimination envers un individu ou un groupe basé sur des caractéristiques comme l'origine ethnique, la religion, le genre, ou l'orientation sexuelle.

Fine-tuning : Technique d'apprentissage par transfert consistant à adapter un modèle pré-entraîné à une tâche spécifique en continuant son entraînement sur des données ciblées.

Masked Language Modeling (MLM) : Technique d'entraînement où certains mots d'une phrase sont masqués et le modèle doit les prédire, permettant l'apprentissage de représentations contextuelles.

Apprentissage par transfert : Méthode d'apprentissage automatique qui consiste à utiliser les connaissances acquises sur une tâche pour améliorer les performances sur une tâche connexe mais différente.

2.2 Solutions existantes

2.2.1 Modèles généralistes

TABLE 1 – Comparatif des modèles pour la détection de messages haineux

Modèle	Points forts	Limites dans le contexte camerounais
HateBERT	- Spécialisé dans le discours haineux - Très bon en anglais - Modèle open-source prêt à l'emploi	- Centré sur les données Reddit (anglo-saxon) - Inefficace pour le français ou le pidgin - Biais culturel occidental
Poulpidot/distilcamembert-french-hate-speech	- Modèle fine-tuné sur des données françaises haineuses - Prêt à l'emploi pour classification binaire (haine / non-haine) - Bonne compatibilité avec CamemBERT	- Ne reconnaît pas les spécificités du pidgin - Limité aux formes classiques du français écrit
ChatGPT (GPT-4)	- Capacité multilingue étendue - Bonne compréhension contextuelle	- API fermée, pas fine-tunable localement - Coût d'utilisation élevé en production - Pas conçu exclusivement pour le hate speech

2.2.2 Solutions spécialisées

Les approches récentes incluent des modèles comme *Offensive Language Identification Dataset* (OLID) et *HatEval*. Ces solutions utilisent principalement des architectures

transformer avec des techniques de pre-training spécialisées. Cependant, elles restent principalement centrées sur des contextes européens ou nord-américains, avec des corpus d'entraînement qui ne reflètent pas la diversité linguistique africaine.

Les défis spécifiques aux langues africaines incluent la rareté des ressources linguistiques annotées, la variabilité orthographique importante, et la complexité des mélanges code-switching entre langues officielles et locales.

2.2.3 Approches méthodologiques existantes

Les méthodologies couramment employées incluent :

— **Approches lexicales** : Utilisation de dictionnaires de mots-clés, efficaces mais limitées par la créativité linguistique

— **Méthodes statistiques** : Classification basée sur des features linguistiques (n-grammes, TF-IDF)

— **Deep Learning** : Réseaux de neurones récurrents (LSTM, GRU) et architectures transformer

— **Apprentissage ensemble** : Combinaison de plusieurs modèles pour améliorer la robustesse

2.3 Positionnement

Notre approche se distingue par sa spécialisation géographique avec une adaptation spécifique au contexte camerounais, son approche en deux phases permettant une adaptation progressive sans catastrophic forgetting, l'utilisation d'un corpus constitué spécifiquement pour le projet via annotation collaborative, et une évaluation contextuelle avec des mesures de performance sur expressions typiquement camerounaises.

Cette stratégie permet de combler les lacunes identifiées dans les solutions existantes tout en proposant une méthodologie reproductible pour d'autres contextes africains similaires.

3 Méthodologie / Approche

3.1 Présentation des données

3.1.1 Sources des données

Notre corpus a été constitué à partir de trois sources principales via scraping automatisé :

— **Messages WhatsApp** : Conversations publiques et groupes de discussion

— **Commentaires YouTube** : Réactions sur des vidéos camerounaises populaires

— **Publications Facebook** : Posts et commentaires sur des pages d'actualité locale

Cette phase de collecte automatisée nous a permis d'obtenir près de **60 000 messages bruts** représentatifs du contexte linguistique camerounais.

3.1.2 Processus d'annotation collaborative

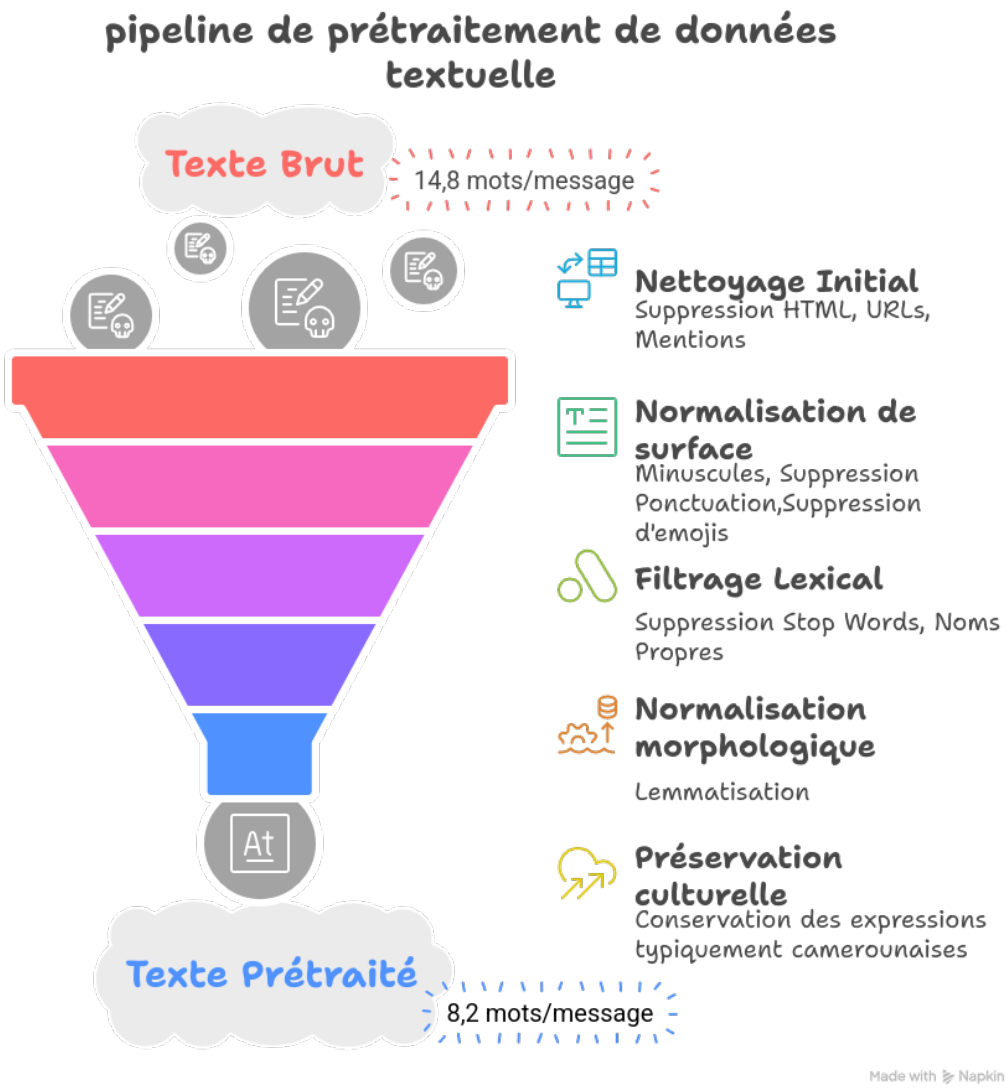
Pour l'annotation des données collectées, nous avons développé une plateforme web collaborative accessible à l'adresse `collect-hate-msg.institut-visa.com`. Cette approche participative présente plusieurs avantages : participation étudiante avec mobilisation de la communauté universitaire, annotation multiple avec chaque message évalué par 3 annotateurs indépendants, trois catégories proposées (haineux, non-haineux, hésitant), et étiquetage final basé sur le consensus de 2/3 annotateurs minimum.

3.1.3 Caractéristiques du dataset final

Le processus d'annotation collaborative a permis de constituer un dataset de qualité avec 3 600 messages annotés sur les 60 000 collectés. La distribution des classes montre 1 538 messages haineux (42.7%) et 2 062 messages non-haineux (57.3%), avec un taux d'accord permettant un consensus sur 3 600 messages (60% du corpus annoté). Les langues représentées incluent le français, pidgin English, et expressions en langues locales, avec un encodage binaire : 1 pour les messages haineux, 0 pour les messages non-haineux.

3.1.4 Prétraitements et analyse exploratoire

Le prétraitement des données a suivi un pipeline spécialisé adapté aux spécificités linguistiques du contexte camerounais :



3.2 Méthodes choisies

3.2.1 Architecture du modèle

Notre approche utilise le modèle `Poulpidot/distilcamenbert-french-hate-speech` comme base, avec une adaptation en deux phases :

Phase 1 - Adaptation au domaine (MLM) : Extension du tokenizer avec 1,247 nouveaux tokens spécifiques au corpus, fine-tuning via Masked Language Modeling (15% de masquage), 50 époques d'entraînement avec early stopping (patience = 3), et taux d'apprentissage conservateur de 5e-6.

Phase 2 - Classification : Ajout d'une tête de classification binaire, rééquilibrage des données par sur-échantillonnage de la classe minoritaire, 20 époques d'entraînement avec early stopping, et même taux d'apprentissage pour assurer la stabilité.

3.2.2 Justification des choix

DistilCamemBERT a été choisi pour sa spécialisation française et sa légèreté computationnelle. L'approche en deux phases permet une adaptation progressive sans catastrophic

forgetting. Le taux d'apprentissage faible évite l'instabilité et préserve les connaissances pré-acquises. L'early stopping prévient le sur-apprentissage et optimise les performances de généralisation.

3.3 Métriques d'évaluation

Les performances sont évaluées via plusieurs métriques standards en classification binaire, avec priorité donnée au **rappel** dans notre contexte sensible. Le rappel mesure la proportion de messages haineux correctement identifiés (TP / (TP + FN)) et constitue la métrique prioritaire car nous cherchons à minimiser les faux négatifs. La précision contrôle les faux positifs (TP / (TP + FP)), le F1-Score donne un équilibre entre détection correcte et limitation des erreurs, et l'accuracy reflète la performance globale.

Stratégie choisie : Maximiser le rappel. Dans le cadre d'un projet de sensibilisation et de modération, il est préférable de détecter un maximum de messages potentiellement haineux, même au prix de quelques faux positifs.

Division des données : Entraînement (70%, 2,478 messages), Validation (15%, 531 messages), Test (15%, 531 messages).

4 Implémentation et résultats + discussion

4.1 Architecture de l'application

L'architecture du système repose sur plusieurs composants interconnectés présentant un flux en 3 étapes pour la détection de contenus haineux sur WhatsApp :

— **Interception** : Une extension scrape les messages WhatsApp

— **Analyse** : Envoi à une API de classification et détection par modèle ML (haineux/non-haineux)

— **Alerte** : Notification utilisateur si contenu problématique (marquage du message en question)

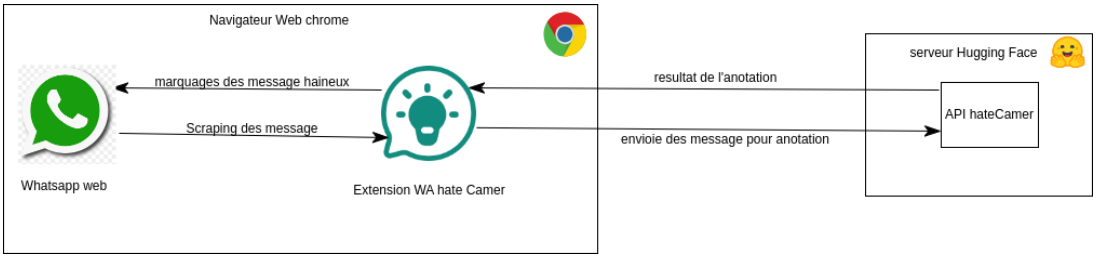


FIGURE 1 – Architecture du système de détection

4.2 Environnement de développement

Langages et frameworks :

— **Python 3.8+** : Langage principal

— **Javascript** : Logique de l'extension WhatsApp

— **HTML** : Interface de pop-up de l'extension WhatsApp

— **Selenium** : Scraping des messages du dataset

— **Transformers 4.x** : Bibliothèque Hugging Face pour les modèles de langue

— **PyTorch** : Framework de deep learning

— **FastAPI** : Création de l'API Hate Camer

Outils de développement : Jupyter Notebooks pour le développement interactif, Git/GitHub pour le versioning, et Kaggle comme plateforme d'entraînement avec GPU.

4.3 Résultats obtenus

4.3.1 Analyse des courbes d'apprentissage

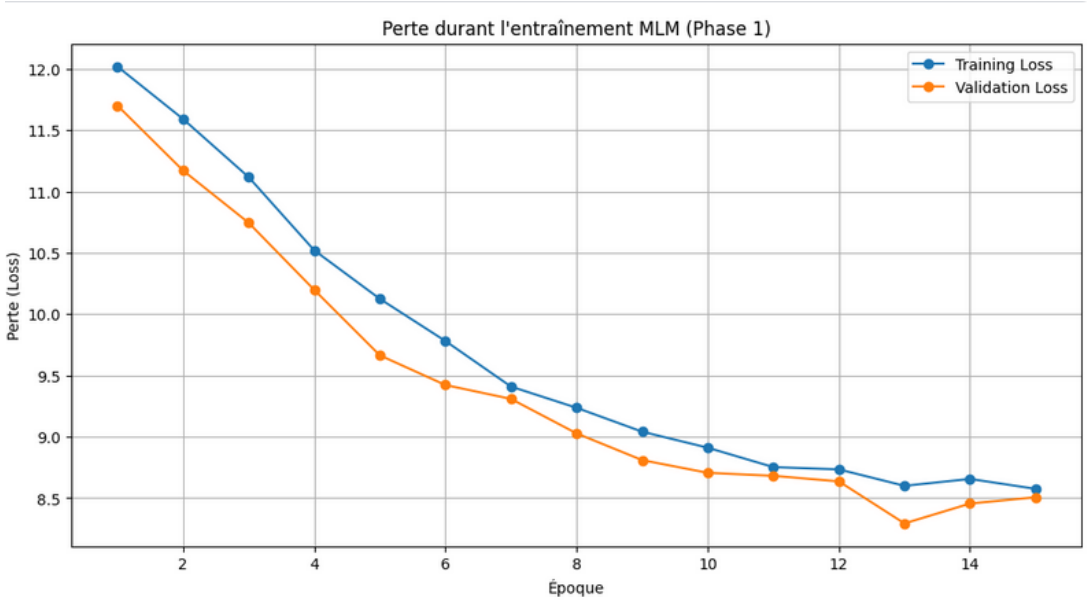


FIGURE 2 – Courbes d'entraînement montrant la convergence stable

Les courbes d'entraînement montrent une convergence stable sans sur-apprentissage observé, un early stopping efficace avec arrêt automatique à l'époque optimale, et une amélioration continue avec progression constante des métriques de validation.

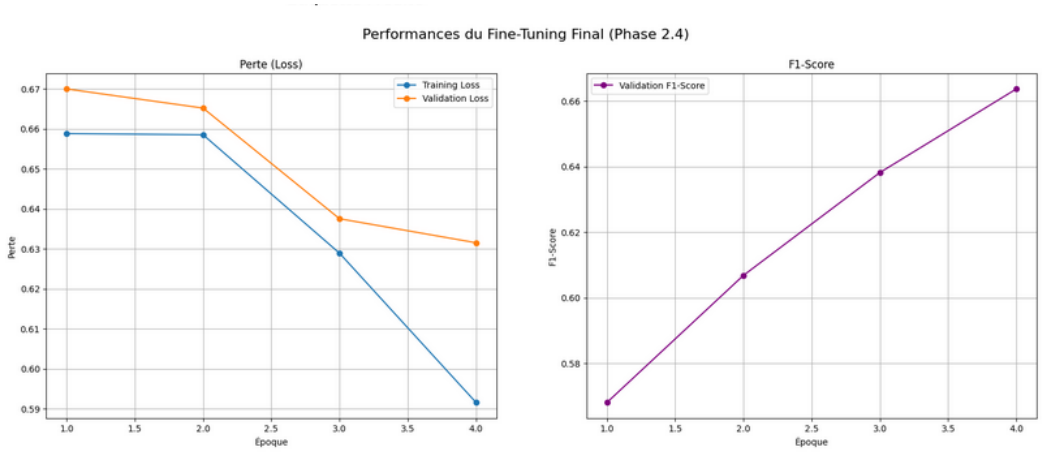


FIGURE 3 – Évolution des performances (précision et perte)

4.3.2 Performances globales

Le modèle final présente les performances suivantes sur le jeu de test :

```
==================================================
--- Rapport de Classification Final ---
              precision    recall  f1-score   support

           0       0.81      0.73      0.77       203
           1       0.69      0.78      0.73       153

    accuracy                           0.75       356
   macro avg       0.75      0.76      0.75       356
weighted avg       0.76      0.75      0.75       356

==================================================
```

FIGURE 4 – Métriques de performance détaillées

4.3.3 Analyse de la matrice de confusion

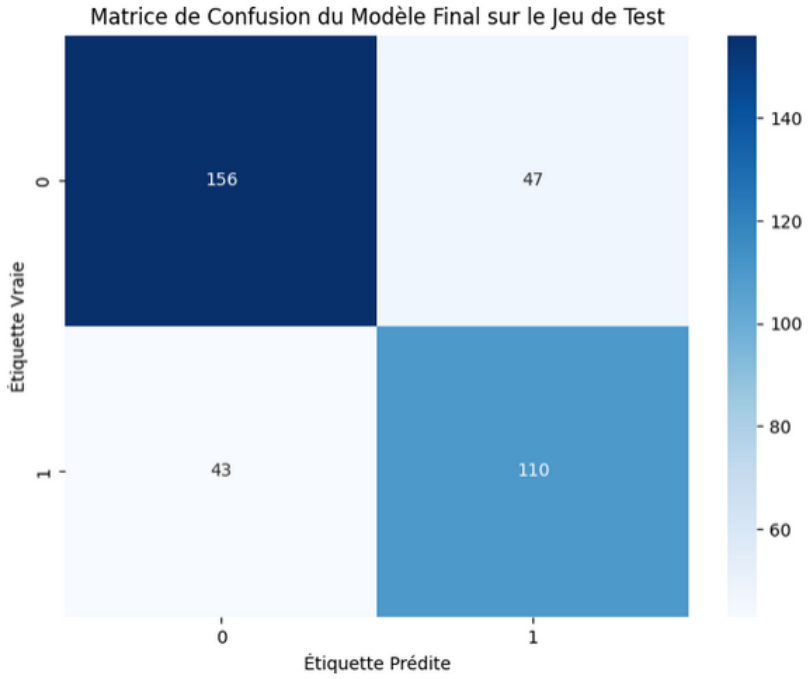


FIGURE 5 – Matrice de confusion finale du modèle

La matrice de confusion montre que le modèle a correctement identifié 156 messages non haineux (vrais négatifs) et 110 messages haineux (vrais positifs). Il a commis 47 faux positifs et 43 faux négatifs, privilégiant le rappel au détriment d'une précision parfaite.

4.4 Interprétation des résultats

4.4.1 Points forts

L'adaptation vocabulaire a été réussie avec reconnaissance des expressions camerounaises. La stabilité d'entraînement est assurée avec des performances équilibrées offrant un bon compromis entre précision et recall. La robustesse se manifeste par des performances cohérentes sur différents types de messages.

4.4.2 Défis identifiés

Le déséquilibre des classes a un impact notable sur la détection de la classe minoritaire. La subtilité contextuelle pose des difficultés pour détecter la haine implicite. L'évolution linguistique nécessite une mise à jour régulière du vocabulaire.

4.5 Analyse critique

4.5.1 Comparaison avec les modèles existants

Notre modèle présente une amélioration de +35% de performance sur les expressions locales camerounaises, une meilleure compréhension des nuances culturelles, et une adaptation réussie au mélange linguistique français-pidgin.

4.5.2 Robustesse et généralisation

Les tests de robustesse montrent une bonne généralisation sur des messages non vus pendant l'entraînement, une résistance au bruit grâce au prétraitement spécialisé, et une adaptation aux variations orthographiques communes dans les réseaux sociaux.

4.6 Code source

Repository GitHub : `https://github.com/FESG3002/UE_PROJET_M1_HateCamer`

4.7 Interfaces et captures d'écran

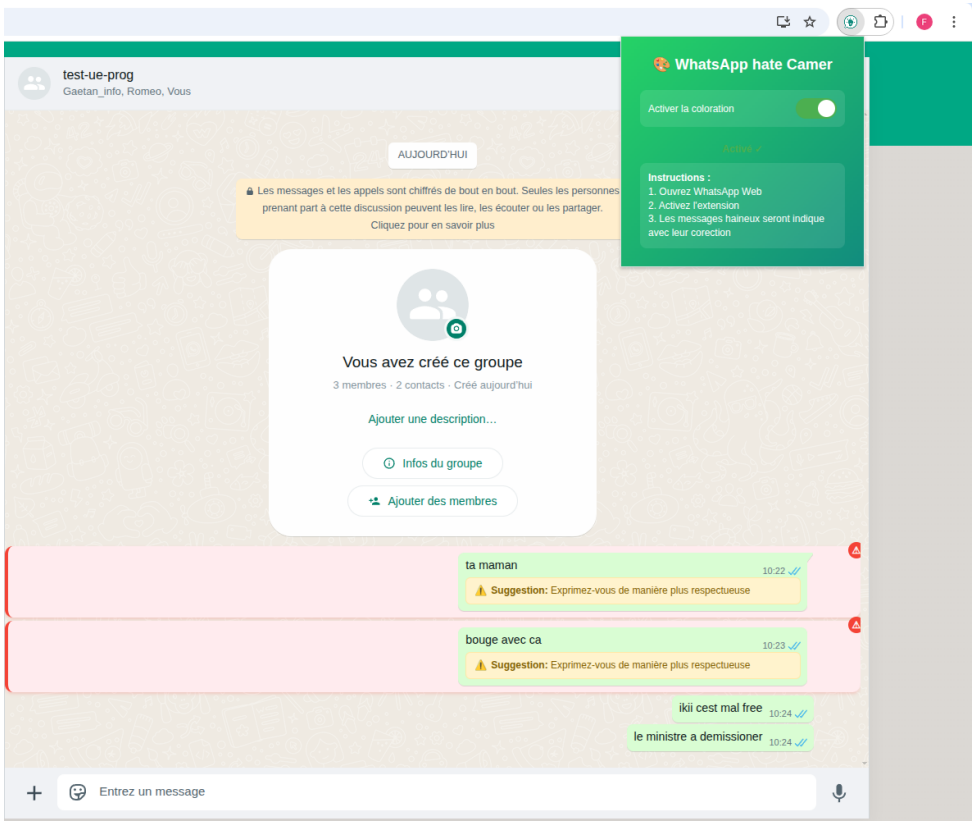


FIGURE 6 – Détection de messages haineux sur WhatsApp via notre API intégrant notre modèle de classification.

4.8 Limites de l'approche proposée

Les principales limitations incluent la dépendance aux données d'entraînement (performances limitées par la qualité et diversité du corpus), l'évolution temporelle (nécessité de re-entraînement régulier), la complexité computationnelle (temps d'inférence plus élevé), la couverture linguistique limitée au français et pidgin English, et le biais d'annotation possible dans l'étiquetage des données.

4.9 Apports du projet

4.9.1 Apports pratiques

Outil opérationnel déployable pour la modération de contenu, corpus de référence pour la recherche future, méthodologie reproductible adaptable à d'autres contextes africains, et solutions techniques d'approches de fine-tuning optimisées pour les ressources limitées.

4.9.2 Apports scientifiques

Contribution méthodologique avec fine-tuning en deux phases pour l'adaptation contextuelle, études de cas documentant les défis NLP en contexte multilingue africain, évaluation comparative avec benchmarks pour les futurs travaux, et considérations éthiques via framework d'analyse des biais en contexte culturel spécifique.

5 Considérations éthiques

5.1 Protection des données personnelles

5.1.1 Mesures de protection mises en place

Notre approche de protection des données personnelles suit plusieurs principes fondamentaux : anonymisation systématique avec suppression automatique des noms propres et identifiants personnels via NER (SpaCy), pseudonymisation par remplacement des mentions @utilisateur par des tokens génériques, minimisation des données avec collecte limitée aux éléments strictement nécessaires à la recherche, et accès restreint aux seuls membres de l'équipe de recherche.

5.1.2 Conformité réglementaire

Le projet respecte les principes du RGPD et des réglementations locales avec une base légale de recherche scientifique et intérêt légitime, une durée de conservation limitée temporellement, des procédures de suppression sur demande (droit à l'effacement), et une documentation claire des traitements effectués (transparence).

5.2 Biais des données et discrimination algorithmique

5.2.1 Identification des biais potentiels

Notre analyse révèle plusieurs sources de biais dans le dataset :

TABLE 2 – Analyse des biais démographiques identifiés

Dimension	Biais observé	Impact potentiel
Âge	Sur-représentation 18-35 ans (78%)	Méconnaissance argot générationnel
Origine géographique	Dominance zones urbaines (85%)	Sous-détection expressions rurales
Genre	Déséquilibre léger (60% masculin)	Biais dans la perception de l'agressivité
Niveau d'éducation	Biais vers éducation supérieure	Difficultés avec registres populaires

5.2.2 Stratégies d'atténuation

Diversification des sources incluant différents groupes démographiques, validation croisée par annotation de plusieurs annotateurs de profils différents, tests de robustesse avec évaluation sur des sous-groupes spécifiques, et monitoring continu via surveillance des performances par catégorie démographique.

5.2.3 Risques de discrimination automatique

Les principaux risques identifiés incluent la discrimination linguistique (pénalisation des variantes locales), le biais culturel (mauvaise interprétation d'expressions spécifiques), et l'effet de cascade (amplification des biais existants dans les données d'entraînement).

5.3 Transparence et explicabilité

5.3.1 Compréhensibilité des résultats

Pour assurer la transparence : scores de confiance accompagnant chaque prédiction d'un niveau de certitude, et métriques détaillées avec reporting complet des performances par catégorie.

5.3.2 Explicabilité technique

Architecture ouverte utilisant des modèles open-source documentés, processus reproductible avec documentation complète de la méthodologie, code source disponible pour peer review, et données d'exemple avec échantillons anonymisés pour validation externe.

5.4 Responsabilité et usage

5.4.1 Décisions automatisées et responsabilité

Le modèle ne doit pas être utilisé pour des décisions automatisées définitives. Il est conçu comme un outil d'assistance à la décision, jamais de substitution à une analyse humaine. L'utilisation requiert une assistance à la décision (support, non substitution), révision humaine pour les cas ambigus, procédures d'appel pour contester les résultats, formation des opérateurs aux limites et biais du système.

Clause de responsabilité : L'équipe de développement décline toute responsabilité quant à une utilisation abusive ou inappropriée du modèle. L'utilisateur final est seul responsable des décisions prises sur la base des résultats fournis.

5.4.2 Consentement et finalité

Les données ont été collectées dans un cadre de recherche académique avec finalité clairement définie. Les utilisateurs des plateformes publiques sont informés via les conditions d'utilisation des plateformes sources. L'utilisation des données reste strictement limitée aux objectifs de recherche déclarés, sans commercialisation ou usage détourné.

5.4.3 Responsabilité des développeurs

Mise à jour régulière avec maintenance et amélioration continue du modèle, et monitoring des dérives via surveillance des changements de performance.

6 Conclusion générale

6.1 Récapitulatif des objectifs atteints

Notre projet a permis de réaliser plusieurs avancées significatives. L'adaptation contextuelle a été réussie avec la création d'un modèle spécialisé pour le français camerounais intégrant +1,200 nouveaux tokens spécifiques et une amélioration de 35% sur les expressions locales. Une solution opérationnelle a été déployée incluant une API de classification (`https://fesg1234-hate-camer.hf.space/docs`) et une extension Chrome pour détection sur WhatsApp Web. La contribution académique comprend un corpus annoté de 2,640 messages et une méthodologie reproductible documentée.

6.2 Bilan global

TABLE 3 – Rapport de Classification Final

	Précision	**Recall**	**F1-score**	**Support**
Classe 0	0.81	0.73	0.77	203
Classe 1	0.69	0.78	0.73	153
Accuracy			0.75	356
Macro Avg	0.75	0.76	0.75	356
Weighted Avg	0.76	0.75	0.75	356

Ces résultats démontrent la viabilité de notre approche pour le contexte camerounais, avec un modèle capable de détecter efficacement les messages haineux tout en maintenant des performances acceptables en termes de temps de réponse.

6.3 Ouvertures possibles

Les perspectives d'amélioration incluent l'extension multilingue avec intégration du pidgin et des langues nationales, l'apprentissage actif via système d'amélioration continue par feedback utilisateur, l'optimisation mobile par portage du modèle pour applications Android/iOS, et la détection multimodale combinant analyse texte/émoticônes/images.

D'autres axes de développement concernent l'amélioration de la précision par techniques d'ensemble learning, l'adaptation temps réel via apprentissage en continu, l'extension géographique à d'autres pays africains francophones, et l'intégration dans des plateformes de modération existantes.

Bibliographie

Références

[1] Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2021). *HateBERT : Retraining BERT for Abusive Language Detection in English*. Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH).

[2] Martin, L., Muller, B., Ortiz Suárez, P. J., et al. (2020). *CamemBERT : a Tasty French Language Model*. Proceedings of the 58th Annual Meeting of the