Détection de Messages Haineux dans le Contexte Camerounais

Cas de la détection de messages haineux sur WhatsApp

NDONKOU FRANCK
TCHIAZE FOUOSSO ROMERO
FOTSING ENGOULOU SIMON GAETAN

Master 1 Data Science - UY1

Encadré par : Dr Thomas MESSI NGUELE

22 juin 2025

Contexte et Problématique

Contexte Camerounais

- +293 langues locales
- Français, Anglais, francamglais
- Communication à travers des plateformes comme TikTok, WhatsApp, Facebook et YouTube.
- Expressions haineuses spécifiques non détectées

Problématique

- Modèles existants inadaptés au contexte local
- Expressions comme **"ta maman "**, **"vas en brousse avec ça**
- Mélange linguistique complexe

Objectifs du Projet

Objectifs Principaux

- Développer un modèle de détection de messages haineux spécifiquement adapté au contexte linguistique et culturel camerounais
- Utiliser ce modèle pour la détection des discours haineux sur WhatsApp

Collecte et Annotation des Données

Phase 1 : Collecte par Scraping

- Facebook, WhatsApp, YouTube
- **60 000 messages** collectés
- Représentatifs du contexte camerounais

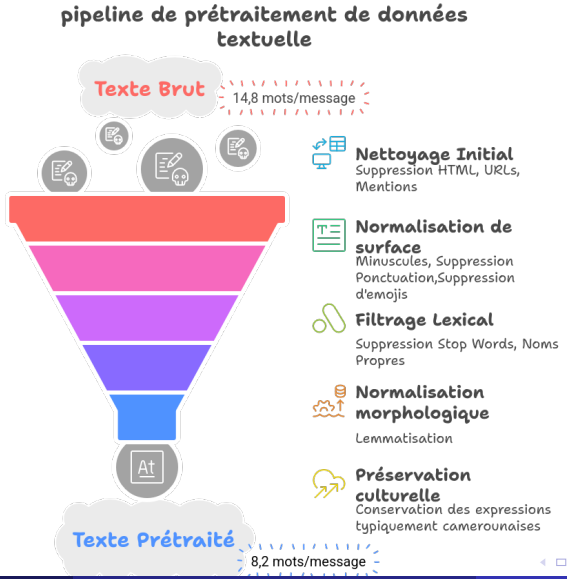
Phase 2 : Annotation Collaborative

- Site web : `collect-hate-msg.institut-visa.com`
- **3 annotateurs** par message
- Vote majoritaire (2/3 consensus)
- Classes : Haineux/Non-haineux/Hésitant

Résultats d'Annotation

- **3 600 messages** annotés
- 1 538 haineux (42.7%)
- 2 062 non-haineux (57.3%)
- Équipe : 1900 messages
- Communauté : 1 700 messages

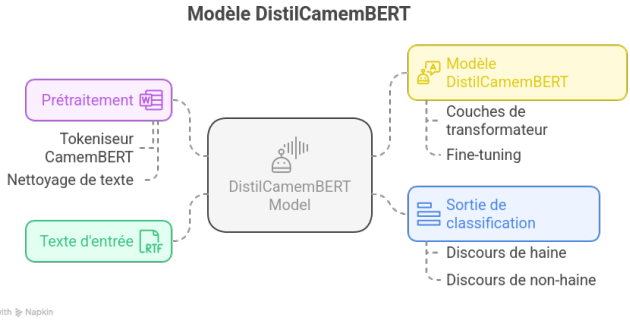
Pipeline de Traitement des Données



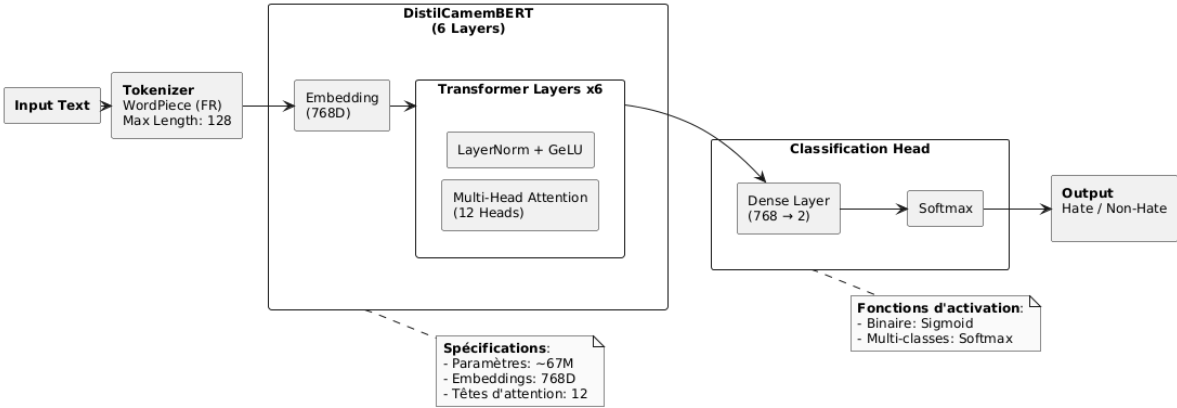
Fine-tuning du Modèle DistilCamemBERT

Modèle de Base : DistilCamemBERT

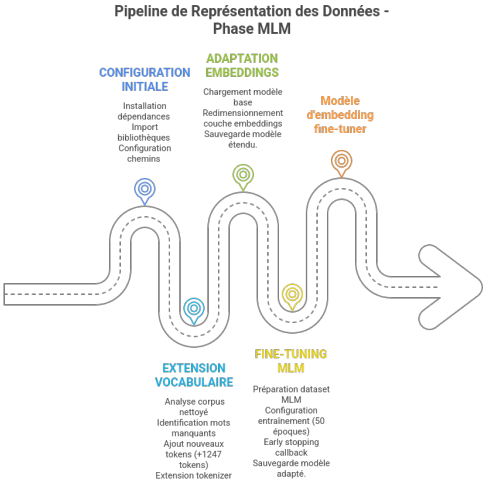
- Poulpidot/distilcamenbert-french-hate-speech
- Pré-entraîné sur 44 milles messages francais(20k haineux et 24k non haineux
- Spécialisé pour la détection de haine
- 66M paramètres (version allégée)

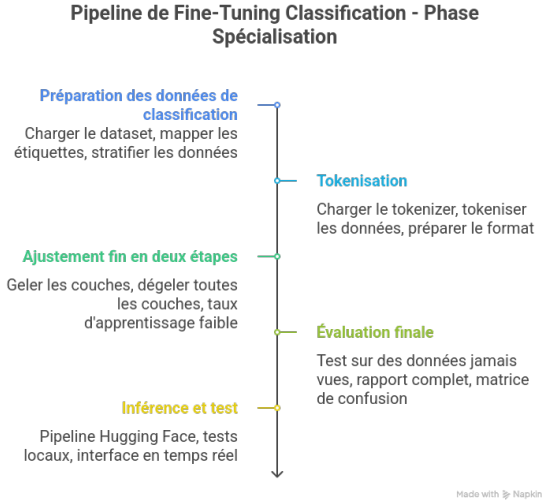


Architecture Complète de notre Modèle



Pipeline de Représentation des Données et de Classification





Métriques d'Évaluation et Justification

Stratégie : Maximiser le Rappel

- Objectif : Minimiser les faux négatifs
- Philosophie : Mieux détecter un message haineux que de le manquer
- But : Sensibilisation et éducation

Métriques Utilisées

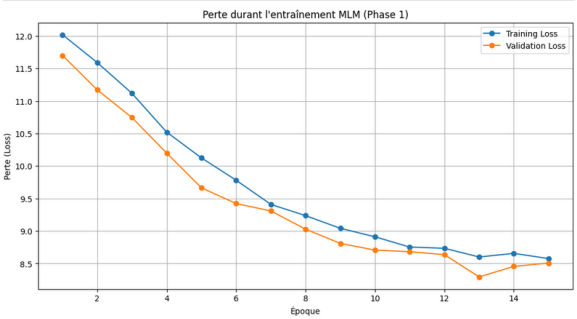
- **Rappel (Recall)** : TP/(TP+FN) → À maximiser
- Précision : TP/(TP+FP)
- F1-Score : Moyenne harmonique
- Accuracy : Performance globale

Priorité

Détecter 100% des messages haineux, même au prix de quelques faux positifs

Courbes d'Apprentissage

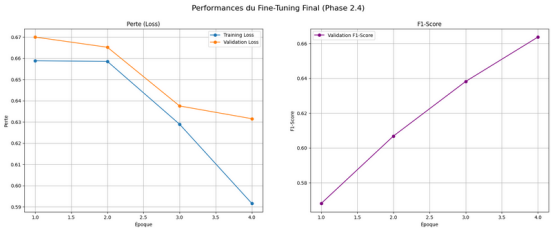
Phase MLM - Adaptation Vocabulaire



Observations

- Convergence stable
- Pas de sur-apprentissage
- Early stopping efficace

Phase Classification



Observations

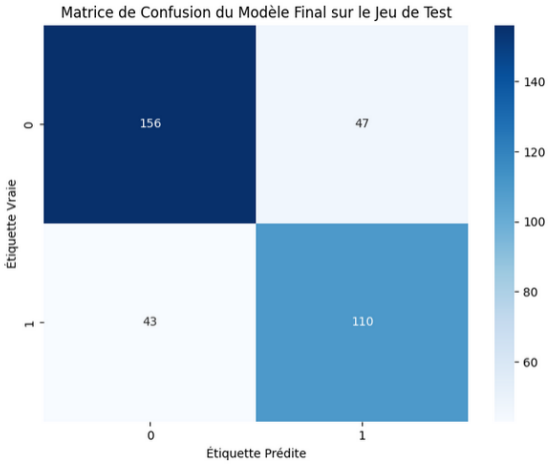
- Amélioration continue
- Début de surapprentissage après la 3^{e} époque

Résultats - Performances du Modèle

```
==================================================
--- 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

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

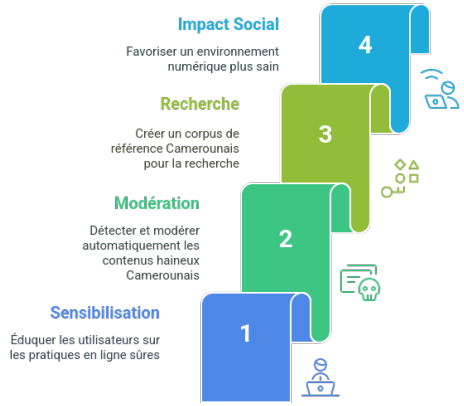
Analyse Détaillée

- VN : 156 (bonne détection non-haineux)
- VP : 110 (détection correcte haineux)
- FN : 47 (messages haineux manqués)
- FP : 43 (sur-détection)

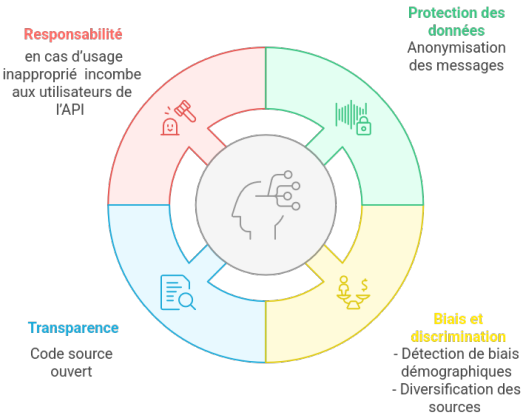


Impact et Considérations Éthiques

Atteindre un Environnement Numérique Plus Sain

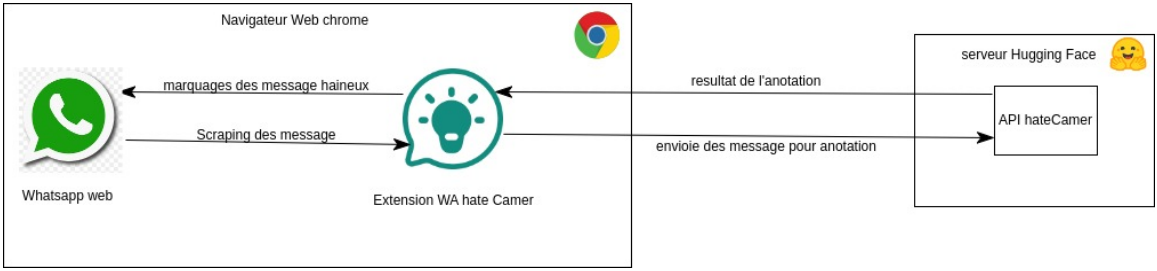


Considérations Éthiques



Made with Napkin

Architecture de l'application



Conclusions et Perspectives

Objectifs Atteints

- ✓Modèle adapté au contexte camerounais
- ✓Corpus de 3 600 messages annotés
- ✓+35% sur expressions locales
- ✓Détection des discours haineux sur WhatsApp

Perspectives d'Amélioration

- Extension multilingue (pidgin, langues locales)
- Optimisation du rappel (objectif : plus de 70%)
- Apprentissage actif avec feedback
- Déploiement mobile Android/iOS

Contribution

- Développement d'un corpus annoté pour la détection de discours haineux spécifique au contexte sociolinguistique camerounais
- Première solution NLP spécialisée pour la détection de haine en contexte camerounais

Références Bibliographiques I

Poulpidot (2021).
Poulpidot/distilcamenbert-french-hate-speech.
Hugging Face Model Hub.
https://huggingface.co/Poulpidot/distilcamenbert-french-hate-speech

Wolf, T., et al. (2020).
Transformers : State-of-the-Art Natural Language Processing.
In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations.