Projet IA

Modèle d´IA permettant la détection de messages toxiques dans un Webchat.

Groupe 5 : BERGONZI Vinicius, LENING Steve, MASI Alessio, THEUBO Ghislain

Idée du projet

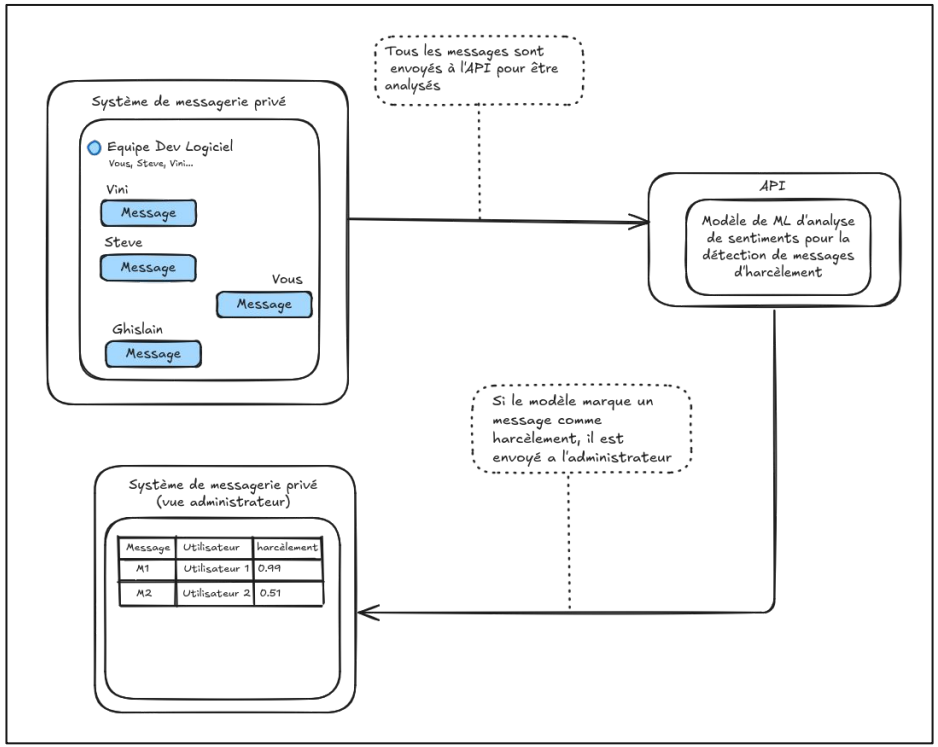
L'objectif de ce projet est de développer un Modèle d'IA qui va permettre d'évaluer de façon automatique chaque message envoyé dans une application de Chat interactive. Pour chacun de ces messages, le modèle vérifie s'il s'agit oui ou non de harcèlement en calculant une probabilité de harcèlement comprise entre 0 et 1. Si la valeur retournée pour un message est >= 0,5 , il s'agit potentiellement de harcèlement et pour le confirmer, on envoie ce type de messages (messages suspects) à un administrateur qui va trancher et effectuer une action qui pourrait consister à supprimer le message du Chat par exemple. Le modèle est entraîné sur un jeu de données équilibré contenant 50% de messages négatifs et 50% de messages positifs. On espère ainsi avoir un modèle dont le résultat soit le plus cohérent possible, quoique le modèle ne pourra toujours pas identifier des situations comme les critiques constructives ou le second degré, d'où le choix de laisser la décision finale à un humain.

Architecture

Le système est composé de:

- Un système de messagerie privée de l'entreprise (vue utilisateur)
- Un système de messagerie privée de l'entreprise (vue administrateur)
- Une API contenant le modèle utilisé pour la détection de harcèlement

Le tout dans une architecture **Client-Serveur**.



Modèle et entraînement

Modèle de base : Deberta-v3-small

- 44 Millions de paramètres.
- 6 couches et 768 tailles cachées.

Jeu de données : spécialisé pour évaluation des messages du chat

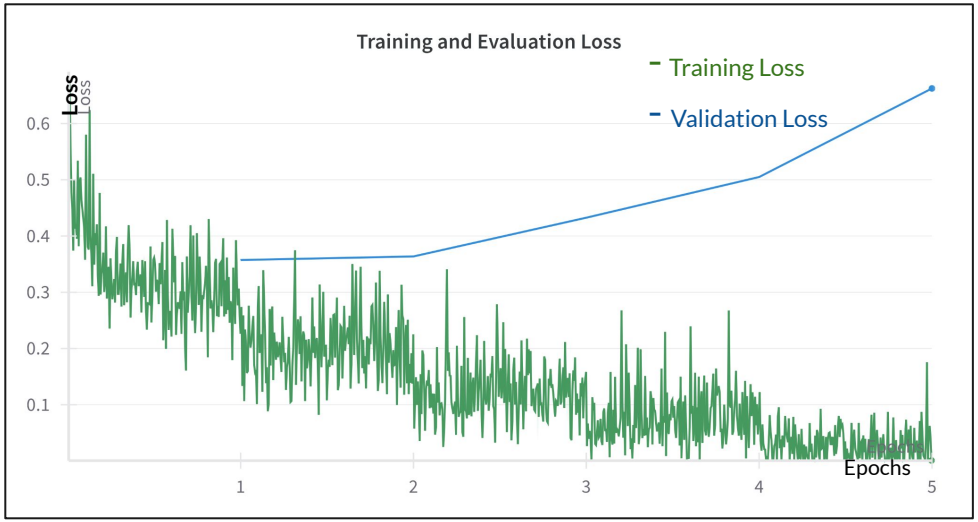
- 38000 valeurs équilibrées entre harcèlement (50%) et non harcèlement (50%).
- **Source des données :** https://huggingface.co/datasets/ifmain/text-moderation-410K

Entraînement :

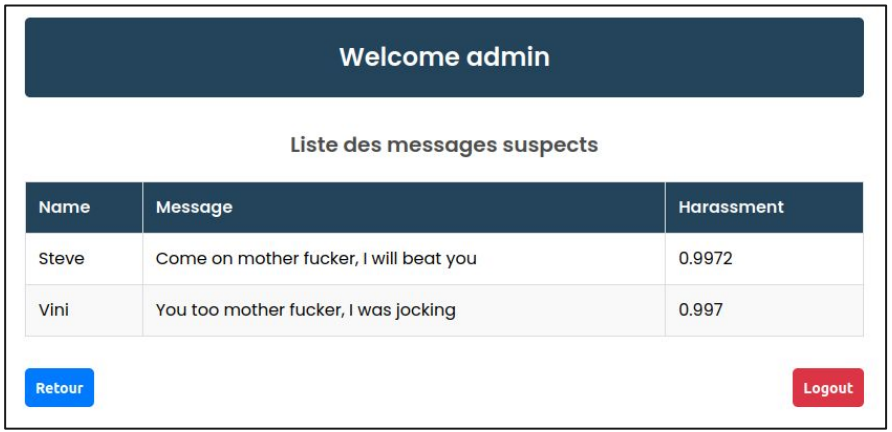
- 5 époques
- Évaluation de l'erreur faite par la valeur du **loss**

Les résultats de l'entraînement sont encourageants :

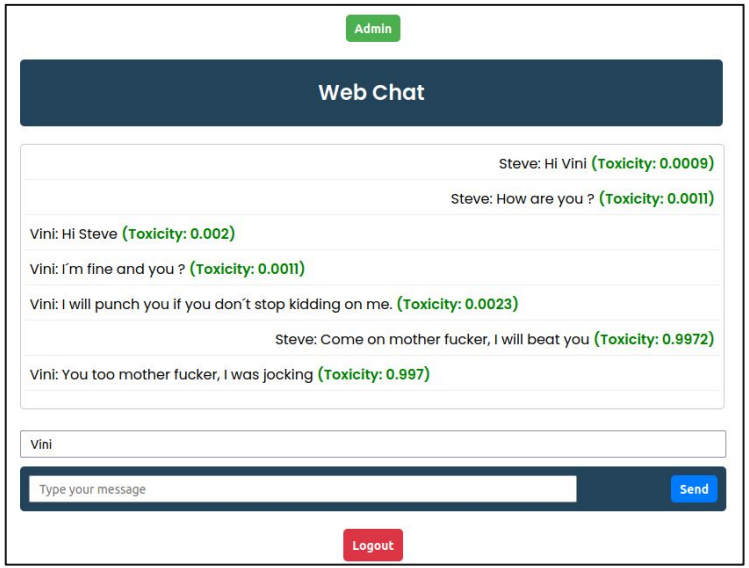
- F1 score : 90.1%
- Accuracy : 95%



Démonstration



Vue administrateur avec l'ensemble des messages dont la toxicité est >= 0,5



Vue utilisateur contenant l'historique des messages échangés entre Vini et Steve

Merci de votre attention !