Benchmark pour la classification de commentaires toxiques sur le jeu de données Civil Comments

Corentin Duchêne*, Henri Jamet*, Pierre Guillaume*, Réda Dehak*

* EPITA Speaker and Language Recognition Group (ESLR),
Laboratoire de Recherche de l'EPITA (LRE), France
{prénom.nom}@epita.fr

Résumé. La détection des commentaires toxiques sur les réseaux sociaux est devenue essentielle pour la modération automatique des messages. Dans cet article, nous présentons une comparaison d'un large éventail de modèles sur un ensemble de données multi-labels de discours haineux. Nous prenons en compte dans notre comparaison le temps d'inférence, les performances et le biais en utilisant différentes métriques. Nous avons découvert que tous les modèles BERT ont des performances similaires, indépendamment de leur taille, des optimisations ou du langage utilisé pour le pré-entraînement. Les réseaux BiLSTM restent un bon compromis entre la performance et le temps d'inférence. Le modèle RoBERTa utilisant la fonction Focal Loss pour l'entraînement demeure le moins biaisé de tous. Comme prévu, le modèle DistilBERT a le temps d'inférence le plus faible des modèles BERT. Enfin, tous les modèles sont affectés par le biais d'association des identités à la toxicité. Les modèles BERT, RNN et XLNet y sont moins sensibles que les CNN et les Compact Convolutional Transformers.

1 Introduction

La détection des commentaires toxiques sur les médias sociaux s'est avérée essentielle pour la modération automatique du contenu. Selon le ministre français de l'éducation, 18% des élèves français ont été victimes de harcèlement sur les réseaux sociaux en 2021. Dans le même temps, le nombre de publications sur ces plateformes n'a cessé d'augmenter. En 12 ans, le nombre de tweets par jour a été multiplié par dix pour atteindre 500 millions aujourd'hui [1].

Cela montre que la détection rapide et ciblée des commentaires toxiques sur les réseaux sociaux est devenue un enjeu crucial pour assurer la cohésion de la société. Par conséquent, cela ne peut se faire qu'en automatisant la modération en ligne.

De nos jours, les types de modèles les plus performants en matière de classification de textes et représentant l'état de l'art sont des modèles basés sur des Transformers (Vaswani et al., 2017) tels que le modèle BERT (Devlin et al., 2019). Plusieurs modifications de ce modèle ont été proposées, Liu et al. (2019) ont fine-tuné un modèle BERT préentraîné pour identifier les discours offensants, catégorisant automatiquement les types de haine et identifiant la cible de ces commentaires.

1. https://www.internetlivestats.com/twitter-statistics/, Statistiques d'utilisation de Twitter - Internet Live Stats.