Projet EISTI 2020

Commentaires toxiques

Description

Les menaces, le harcèlement, les insultes en ligne sont de plus en plus répandus et de nombreuses personnes cessent de s'exprimer et renoncent à donner leurs opinions. Pour lutter contre ce nouveau fléau les plateformes doivent efficacement classifier les conversations.

L'objectif de ce projet est de contruire un outil pour détecter les conversations/comportements négatifs en ligne (commentaires toxiques, c'est-à-dire les commentaires impolis, irrespectueux ou susceptibles de faire quitter la discussion à quelqu'un).

Résultat attendu

Construire un modèle qui à partir d'un texte en entrée va détecter son niveau de toxicité (plusieurs types sont possibles dans un texte). Le modèle doit être capable de détecter différents types de toxicité comme les menaces, l'obscénité, les insultes, etc. Le modèle doit prédire une probabilité pour chaque classe (type de toxicité). Proposer une application/démonstrateur.

Données

L'ensemble de donnée provient des pages de discussion de Wikipedia. Il contient plusieurs dixaines de milliers de commentaires qui ont été labélisés. Les types sont : `toxique`, `sévèrement toxique`, `obscène`, `menace`, `insulte`, `haineux`.

Avertissement : l'ensemble de données pour ce projet contient des textes qui peuvent être considérés comme profanes, vulgaires ou offensants.

Téléchargement des données :

https://drive.google.com/open?id=1Ob9SA5_uQ-rw9OJvuueDrkpQkTqa7aXe