# NLP – Modération de Chat

Nakache Eithan
Ziane Camil
Hadj-Said Samy
Perez Jason
Six Briac
Bellamy Baptiste

**Sujet**

Modération en temps réel des chats sur une plateforme de contenu, par l'intermediaire de modèle NLP

# Table des matières

# 1. Introduction

Ce projet se propose de développer des modèles du traitement du language naturel dédiés à la modération en temps réel des chats sur une plateforme de contenu. L'objectif principal est d'assurer la modération des messages des utilisateurs afin de prévenir la diffusion de contenus inappropriés, tels que les insultes, les spams et autres formes de communications indésirables.

### Dataset

Le dataset choisit pour ce projet est le **Jigsaw Toxic Comment Classification**. Ce dataset recense un grand nombre de commentaires anglais, provenant de Wikipedia, labellisés par des humains en fonction de leur toxicité. Ce dernier provient d'une compétition Kaggle et peut être obtenu à partir du lien suivant : jigswaw toxic comment classification challenge.

### Objectif

Plusieurs modèles de machine learning et de deep learning seront développés et entraînés dans le but de prédire la pertinence et l'acceptabilité des messages. Une fois ces modèles développés, nous procéderons à une évaluation rigoureuse de leurs performances par comparaison mutuelle. L'ultime étape de ce projet consistera en la création d'une preuve de concept sous la forme d'une interface web. Cette interface permettra de modérer en direct les messages échangés sur un chat de live streaming, démontrant ainsi l'applicabilité pratique de nos recherches.

# 2. Présentation du jeu de données

### Objectifs et Financement du Dataset

Le dataset de classification des commentaires toxiques de JIGSAW a été créé pour promouvoir la recherche sur la détection de la toxicité dans les commentaires en ligne. Il vise à identifier des comportements indésirables tels que les commentaires toxiques. Ce corpus a été collecté dans le but de développer des technologies et des méthodes capables de modérer automatiquement ces types de contenu nuisible sur les plateformes en ligne.

Le projet a été financé par Jigsaw (anciennement connu sous le nom de Google Ideas) et Google, dans le cadre d'un concours organisé sur la plateforme Kaggle. Ce partenariat a non seulement mis à disposition les ressources nécessaires, mais a également encouragé la communauté globale des data scientists à résoudre ce problème urgent de modération des contenus toxiques sur Internet.

### Contexte et Caractéristiques des Données du Dataset

Les commentaires inclus dans le dataset proviennent des pages de discussion de Wikipedia. Ces discussions sont menées en anglais par des contributeurs qui échangent sur les améliorations à apporter aux articles et sur les modifications nécessaires. Ces échanges sont caractéristiques des interactions collaboratives typiques sur Wikipedia, où les utilisateurs débattent de la véracité, de la neutralité et de la complétude des articles.

Le format du texte est écrit et prend la forme de communications en ligne non formelles mais structurées. Cela signifie que les commentaires, bien que rédigés dans un cadre informel, suivent une certaine structure logique et sont orientés vers des objectifs spécifiques de collaboration et de modification de contenu.

### Démographie des auteurs

Les informations démographiques spécifiques sur les auteurs des commentaires ne sont pas fournies.

### Processus de collecte

Le dataset comprend environ 160 000 commentaires pour l'entraînement et 60 000 commentaires pour le test. Ces données ont été extraites dans le but de représenter divers comportements toxiques, bien que la méthode exacte d'échantillonnage n'ait pas été spécifiée.

Étant issues de plateformes ouvertes, les questions de consentement sont gérées dans le cadre des normes de Wikipedia concernant la publication de commentaires publics. Toutefois, les détails spécifiques concernant le consentement des auteurs ne sont pas divulgués.

En ce qui concerne le prétraitement, les données ont été anonymisées et les informations personnellement identifiables ont été supprimées pour protéger la vie privée des utilisateurs.

### Processus d'annotation

Le dataset est structuré autour de plusieurs catégories d'annotations qui permettent de définir la nature de la toxicité des commentaires. Celles-ci incluent : **Toxique**, **Très toxique**, **Obscène**, **Menace**, **Insulte**, **Haine identitaire**. Ces catégories ont été choisies pour couvrir un large éventail de comportements toxiques potentiellement rencontrés dans les commentaires en ligne.

La méthode d'annotation repose sur l'intervention de multiples annotateurs pour chaque commentaire. Cette approche vise à maximiser la fiabilité des annotations. Le recours à plusieurs annotateurs permet de réduire les biais individuels et d'améliorer la précision générale des données annotées, assurant ainsi que les modèles de machine learning entraînés avec ce dataset peuvent fonctionner de manière efficace et équitable.

### Distribution

Le dataset est disponible à des fins de recherche non commerciales. Les utilisateurs doivent généralement accepter des conditions d'utilisation qui limitent l'utilisation commerciale et la redistribution.

# 3. Statistiques descriptives du jeu de données

## Division du dataset

Le dataset a été divisé en trois parties : entraînement, validation et test. Voici la répartition du nombre de commentaires dans chacune de ces parties :

| **Catégorie** | **Nombre de commentaires** |
|---|---|
| Train | 127,656 |
| Validation | 31,915 |
| Test | 63,978 |

TABLE 3.1 – Répartition du nombre de commentaires

## Répartition des labels

Les commentaires toxiques sont minoritaires dans l'ensemble des données. En effet il y a **10.2%** de commentaire globalement non-toxique. Cela peut poser des problèmes lors de l'entraînement des modèles, car les classes minoritaires peuvent être sous-représentées et donc mal apprises. Il y a aussi une répartition inégale au sein des labels de toxicité On peut remarquer que la somme des pourcentages n'est pas égale à 100% car un commentaire peut avoir plusieurs labels. On est donc dans un problème de classification multi-labels.

| **Label** | **Pourcentage** |
|---|---|
| toxic | 94.3% |
| severe_toxic | 9.8% |
| obscene | 51.9% |
| threat | 3.1% |
| insult | 48.2% |
| identity-hate | 8.6% |

TABLE 3.2 – Répartition des labels sur les commentaires globalement toxiques

## Corrélation entre les labels

On peut remarqué que les labels de toxicité sont fortement corrélés entre eux. On peut dès à présent anticiper une difficulté du modèle à distinge un commentaire toxique d'un commentaire obscene (**74%** de corrélation). Cela représente un point à prendre en compte lors de la conception du modèle. La matrice de corrélation est inclus dans l'appendice A.

## Longueur des commentaires

La distribution de la longueur des commentaires est très variée. En effet l'écart type est bien plus élevé que la moyenne. La longueur moyenne des commentaires est de **395** caractères, avec un écart-type de **593** caractères. Cela peut poser des problèmes lors de la conception du modèle. Il est donc important de prétraiter les données pour normaliser la longueur des commentaires.

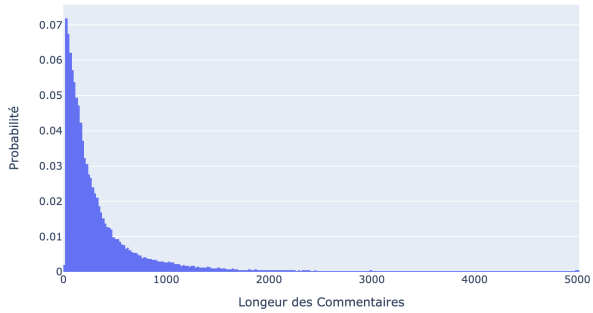


FIGURE 3.1 – Distribution de la longueur des commentaires

## Distribution des mots

Dans le jeu de données, il y a un total de environs 180 000 mots uniques. Sans surprises, les mots les plus fréquents sont les mots de liaison et les mots vides. Mais en vue de la source du dataset les mots **articles**, **wikipedia** et **pages** apparaissent aussi très souvent dans les commentaires (top 26, 30, 31 respectivement). En effet les utilisateurs peuvent citer des sources pour appuyer leurs propos. Dans les commentaires toxiques, on trouve des insultes, des mots vulgaires et des mots discriminatoires. On peut visualiser ces derniers en utilisant WordCloud. Cela représente une représentation visuelle des mots les plus utilisés. En annexe A, on peut voir une représentation WordCloud générée à partir des commentaires du jeu de données filtré selon le type de toxicité.

# A. Appendix

## A.1 Corrélation entre les labels

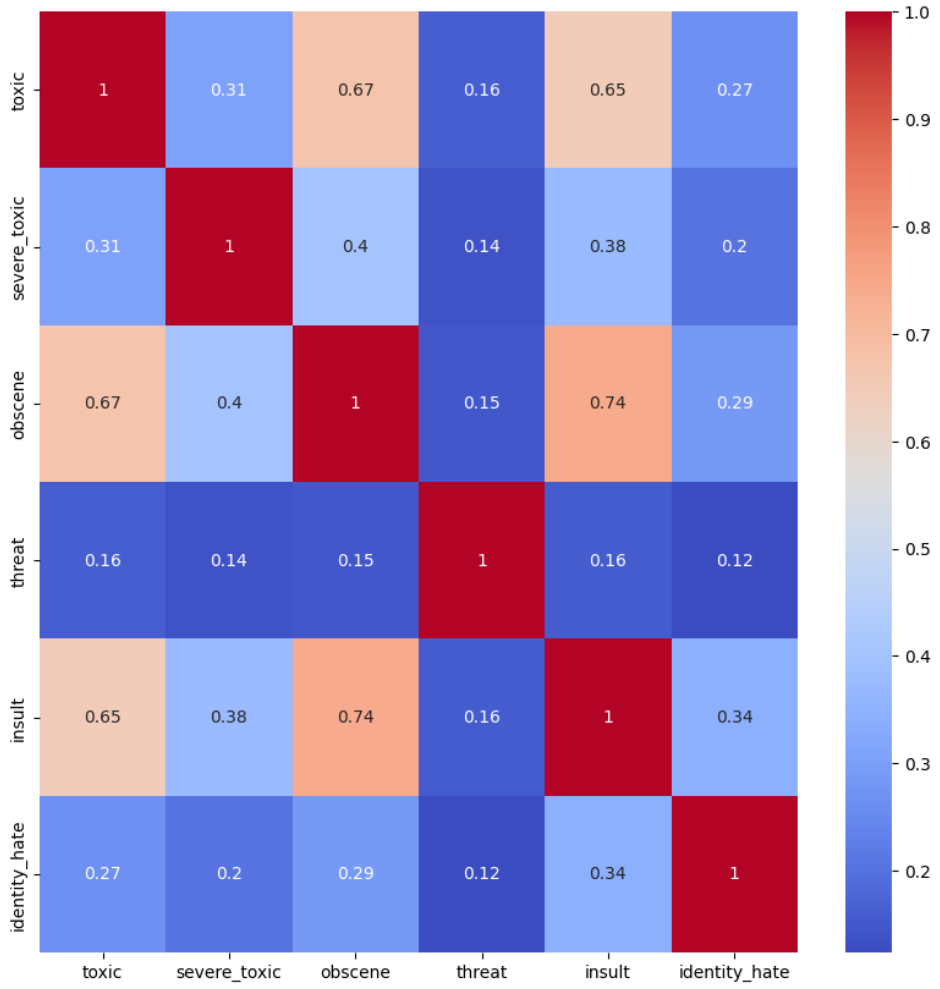


FIGURE A.1 – Matrice de corrélation des labels du dataset

## A.2 Distribution des mots


(a) Word Cloud – Toxique

(b) Word Cloud – Haine identitaire

(c) Word Cloud – Obscene

(d) Word Cloud – Insult

FIGURE A.2 – Word Clouds