Compte – Rendu Semaine 4

Au niveau des données :
Nous avons ajouté une nouvelle colonne, ce qui nous fait un total de 5 colonnes : Identifiant, Texte (en anglais), Traduction, Type, Catégorie. Cette semaine, nous avons surtout travaillé sur les deux dernières colonnes : *Type* et *Catégorie*. La colonne *Type* contient deux classes : *haineux* et *non haineux*, indiquant si le texte constitue un discours haineux ou non. Pour la colonne *Catégorie*, lorsqu'un texte est haineux, il est associé à l'une des six catégories suivantes : Racisme, Sexisme, Homophobie, Islamophobie, Validisme et Xénophobie. Nous avons réparti les données de manière équilibrée dans nos deux bases : dans la base d'entraînement, nous avons 1200 textes haineux et 1200 textes non haineux, avec 200 textes par catégorie pour les textes haineux, soit un total final de 2400 textes. Dans la base de test, nous avons 120 textes haineux et 120 textes non haineux, avec 20 textes par catégorie pour les textes haineux, soit un total final de 240 textes.

Visualisation :
Nous avons réalisé plusieurs visualisations pour mieux comprendre les données et faciliter le travail ultérieur. Nous avons utilisé notamment des nuages de mots et différents types de graphiques (plots).

Code :
Pour la partie code/Python, nous avons apporté plusieurs améliorations. Nous avons commencé par mettre en place une validation croisée et effectué des prédictions pour vérifier si le modèle détecte correctement les catégories, notamment le racisme. À la fin de l'exécution, l'algorithme génère un fichier CSV contenant les résultats, que l'on peut retrouver sur notre dépôt Git.