Compte rendu semaine 2 – Projet Machine Learning : Détection du harcèlement

Semaine 1 :

Au début de la semaine, nous n'avions pas encore choisi définitivement le thème du projet. Nous hésitions notamment avec une analyse linguistique des pronoms, mais nous avons rencontré des difficultés concernant la disponibilité et la qualité des bases de données. Après plusieurs discussions et après avoir réfléchi aux possibilités techniques, nous avons finalement décidé de travailler sur la détection automatique du harcèlement en ligne à l'aide du machine learning. Ce sujet nous a semblé intéressant, surtout pour son importance sociale, et il permet d'utiliser des jeux de données publics déjà annotés.

Nous avons choisi deux jeux de données disponibles sur Kaggle : Linguistic Analysis of Hateful Speech Dataset et YouTube Toxicity Data. Ces bases contiennent des messages (commentaires, publications) ainsi que des indications sur la présence ou non de contenu haineux ou toxique. Notre objectif était d'utiliser ces données pour créer un modèle capable de prédire si un message relève du harcèlement.

Semaine 2 :

Le Travail de cette semaine :

- Nettoyage et tri des données : Nous avons analysé tous les jeux de données et sélectionné les plus pertinents. Pour qu'un jeu soit exploitable, il fallait au minimum une colonne contenant le texte et une colonne avec la catégorie associée à ce texte. Nous avons supprimé toutes les colonnes inutiles afin de ne conserver que les textes et les labels. Nous avons également vérifié la cohérence des catégories et supprimé les éventuelles valeurs manquantes.

- La base d'entraînement : Ce fichier a été créé à partir des deux jeux de données issus de Kaggle. Cependant, il contient une quantité plus importante de données provenant du premier jeu de données (Hateful…). Cela peut créer un déséquilibre entre les catégories. C'est pour cette raison que la semaine prochaine nous prévoyons d'ajouter davantage de données afin d'éviter d'éventuels problèmes lors des prédictions.

- La base test : Elle a été créée de la même manière que la base d'entraînement, mais en plus petite quantité. Nous avons pris environ 1/10 de chaque jeu de données. Concrètement, nous avons récupéré 10 lignes par catégorie de harcèlement dans chaque jeu afin de les ajouter à cette base. Il est important de préciser que les lignes présentes dans le fichier base_test ne figurent pas dans le fichier base_entrainement. Nous avons bien veillé à ce que la base de test soit totalement indépendante de la base d'entraînement pour pouvoir évaluer le modèle sur des données qu'il n'a pas vues pendant l'apprentissage.

- Le code pour tester la cohérence des 2 bases : Nous avons mis en place une vectorisation TF-IDF pour transformer les textes en vecteurs numériques. Nous avons choisi de convertir les textes en minuscules, de supprimer les stop words en anglais et de limiter le nombre de caractéristiques à 5 000 afin d'éviter une

dimension trop importante. Nous avons ensuite entraîné le modèle uniquement sur la base d'entraînement. Une fois l'apprentissage terminé, nous avons utilisé la base de test pour obtenir les prédictions. Pour évaluer les performances, nous avons utilisé un rapport de classification ainsi qu'une matrice de confusion. Cela nous permet d'analyser la précision, le rappel, le F1-score et les erreurs, notamment les faux positifs et les faux négatifs. Enfin, nous avons exporté les résultats dans un nouveau fichier Excel appelé Base_Test_Predictions.xlsx, qui contient les textes de la base de test, les labels réels et les prédictions du modèle.

En résumé, cette semaine nous a donc permis de choisir définitivement notre sujet, de préparer les données, de créer les bases d'entraînement et de test, de mettre en place le modèle et de réaliser une première évaluation de notre système de détection du harcèlement.