PARCOURS INGENIEUR MACHINE LEARNING

PROJET 8 : PARTICIPEZ A UNE COMPETITION KAGGLE !

Objectif : Exposer les résultats obenus et les conclusions associées

RESULTATS ET CONCLUSION

Pour rappel, l'objectif de ce projet était de venir classifier des commentaires en fonction de leur toxicité. Nous étions sur une problématique de NLP multilingue avec utilisation d'un accélérateur de type TPU.
La donnée texte a été nettoyée, des opérations de data augmentation ont été appliquées et nous avons principalement utilisé des modèles pré-entrainés pour essayer de réaliser au mieux cette classification.

Chacun des tests réalisés m'ont permis d'obtenir un score sur la plateforme Kaggle qui a pu évolué au fil des améliorations et modifications apportées. Parfois favorablement, parfois non. Ainsi, j'ai pu passer d'un score public à 0.8674 pour atteindre mon maximum à **0.9346** en 11 soumissions.

Ce score a été obtenu sur de la donnée nettoyée, sans data augmentation et avec un dataset d'entrainement d'environ 500 000 observations (utilisation des 2 fichiers d'entrainement disponibles). Le modèle pré-entrainé choisi est **XLM-Roberta** (jplu/tf-xlm-roberta-large) au détriment des modèles BERT et DistilBERT. L'ajout d'une couche supplémentaire de Dropout n'a pas montré son efficacité et le modèle a été entrainé en 2 fois : 3 époques sur je jeu d'entrainement (en anglais) et 6 époques sur le jeu de validation (multilingue).

En comparaison avec les meilleurs résultats du concours, cela me place dans la 2ème partie de tableau autour de la 1000ème place (sur 1650), le meilleur score obtenu étant de 0.9556.

Par rapport aux discussions de la communauté, j'en ai déduit que le modèle pré-entrainé choisi (XLM Roberta) semblait être le modèle le plus approprié à notre problématique. En effet, ce modèle, basé sur BERT (Robust optimized BERT approach), est un ré-entrainement de BERT avec des améliorations sur la méthodologie et avec beaucoup plus de data et de temps de compute. Il est particulièrement adapté aux problématiques multilingues. Il est donc logique d'avoir de meilleurs résultats avec ce modèle qu'avec BERT ou DistilBERT.
Je pense donc que le modèle pourrait être encore amélioré grâce à 2 leviers : le volume de data pris en compte pour l'entrainement et le preprocessing des données. En effet, j'ai pu constater qu'une data nettoyée permettait d'avoir de meilleurs résultats. Peut être qu'une seconde phase de nettoyage plus poussée pourrait me permettre de faire évoluer le score à la hausse. Aussi, je n'ai utilisé qu'une petite partie du 2ème jeu d'entrainement mis à disposition. Néanmoins, les temps de traitement assez importants pour 1 run m'ont invités à ne pas prendre plus de données en considération. Il aurait peut être été intéressant de compléter mon jeu de données avec la totalité du second fichier dans l'espoir de voir le score s'améliorer une nouvelle fois.