JIGSAW MULTILINGUAL TOXIC COMMENT CLASSIFICATION

PAR :

NAMA NYAM GUY ANTHONY

24 Juin 2020

Résultats et conclusions

Le projet de la compétition ***Kaggle*** a porté sur la conversation en IA. Plus spécifiquement, il s'agissait de prédire la probabilité qu'un commentaire soit toxique. La donnée d'entraînement contient une colonne de commentaires ***comment_text*** fournies en anglais provenant du *Wikipedia talk page comments* et *Civil Comments*. Les commentaires de la donnée d'entraînement et de validation (multilingue) ont été classés comme toxique ou pas (0/1) dans la colonne ***toxic***. La donnée de test contient une colonne de commentaires ***comment_text*** composés de plusieurs différentes langues.

Dans ce contexte multilingue, nous avons utilisé deux modèles multilingues ***Bert*** (bert_multi_cased_L-12_H-768_A-12/2) et ***XLM-Roberta*** (jplu/tf-xlm-roberta-large). Le modèle *Bert* a été rapidement mis de coté en comparaison sur la métrique d'évaluation (auc) au modèle *XLM-Roberta*.

Le résultat obtenu dans cette compétition (score public) avec le modèle *jplu/tf-xlm-roberta-large* est de : ***0.9238***. Résultat obtenu juste avec ***6 soumissions***. Déjà sur ce dernier, il faut dire que le nombre de soumission est faible comparé au trois premiers de cette compétition qui tourne au delà de 300 soumission atteignant même la barre de 385.

Le résultat est jugé moyen dut à une ***mauvaise organisation*** ; méconnaissance de la plateforme kaggle, utilisation entière de la donnée dans la phase de preprocessing qui ont multiplié considérablement les temps d'exécutions, une connaissance réduite du fonctionnement interne du modèle XLM-Roberta, les contraintes de la compétition.
La phase de preprocessing nécessitait davantage d'opérations avancées impliquant les lois de distribution du texte pour l'ensemble des dataset(train, validation et test). A cela, nous pouvons également rajouter la construction de la partie *classifieur* dans la phase de modélisation.

Pour finir, je dirais que c'était une expérience enrichissante d'échange avec la communauté kaggle et je me revois revenir prochainement pour une autre compétition. Les fondamentaux de la plateforme pour les compétitions Kaggle ont été acquis ainsi que les contraintes. L'utilisation de deux dépôts importants de modèles NLU et NLG (***HuggingFace*** et ***TFHub***). L'utilisation des ***TPUs***, la construction de modèles multilingues et la stratégie d'entraînement de ces modèles ; voilà autant de connaissances acquises lors de la compétition.