Introduction

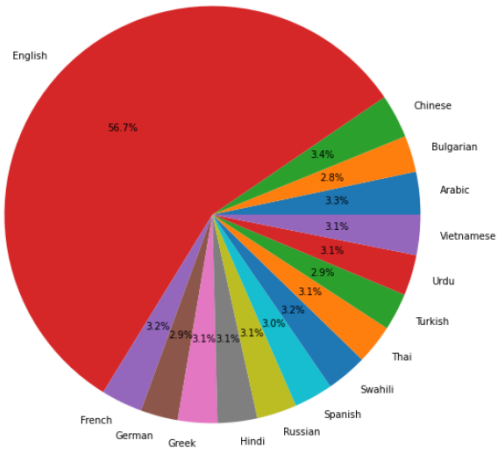
L'objectif dans ce projet est la participartion à une compétition `Kaggle`. Kaggle est une plateforme web qui accueille la plus grande communauté de Data Science au monde. La plateforme offre un environnement Jupyter Notebooks personnalisable et donnant accès à une grande quantité de données et de codes publiés par la communauté. Le choix s'est porté sur une compétition réelle et en cours en *Natural Language Inference* (NLI) dont le jeu de données associé est *Contradictory, My Dear Watson*.

En NLI, on cherche à construire des systèmes capables d'élaborer la reconnaissance d'implications entre deux textes (*Recognising Textual Entailment*, RTE). Soient deux textes **P** (*premise*) et **H** (*hypothesis*), cette tâche consiste à juger si **P** confirme ou infirme **H**. Il existe ainsi trois types d'implications : confirmation, infirmation ou neutralité.

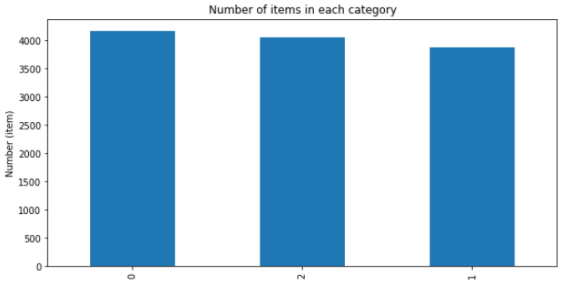
Modélisation

Le jeu de donnée

Sur la figure 1 est montrée la distribution des différentes langues et des trois classes dans le jeu d'entraînement. Il y 15 langues qui sont représentées où l'anglais constitue plus de 56% du dataset, les autres langues sont présentent à hauteur d'environ 3% chacune. D'autre part, les classes dans le jeu de données sont plutôt équilibrés, avec chacun des trois labels représentant le tiers.



(a) Distribution des différentes langues présentent dans le jeu d'entraînement.



(b) Distribution des trois labels dans le jeu d'entraînement.

Figure 1: Distribution (a) des langues et (b) des labels dans le jeu d'entraînement.

Modèle final

Il existe de nos jours plusieurs modèles d'apprentissage Deep Learning basés sur les architectures des *Transformers*. Ces dernières années, il est possible d'obtenir d'excellents résultats avec des modèles multilingues capables de traiter plusieurs langues. Les plus populaires sont *mBERT* et *XLM-Roberta*, basés sur le modèle *BERT* décrit comme marquant le début d'une nouvelle ère pour le NLP.

XLM-Roberta semble être le modèle multilingue le plus précis, et notre choix s'est porté sur l'une des variantes de ce modèle : `'joeddav/xlm-roberta-large-xnli'`. Le prétraitement des données en amont est subtil et est extrêmement important, malgré qu'il puisse paraître trivial, pour obtenir de meilleures performances. Cependant, nous n'avons pas procédé au nettoyage du contenu textuel. La segmentation (*tokenization*) a été d'autre part réalisée avec soin. En entré au modèle, on fourni les textes **P** et **H** espacé/séparé avec les jetons spéciaux [CLS] et [SEP] comme illustré sur la figure 2.

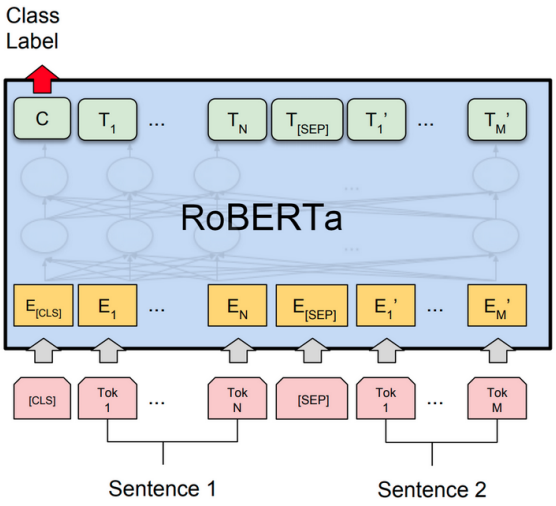


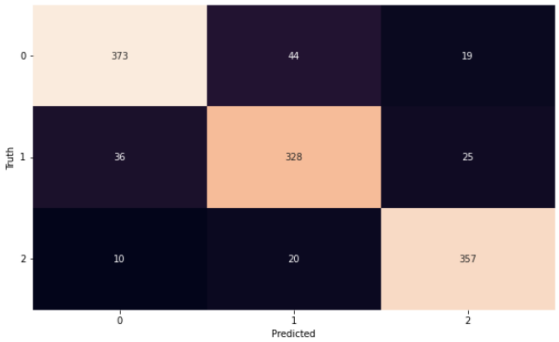
Figure 2: Configuration des séquences en inputs au modèle.

Résultats & Conclusion

Le notebook utilisé lors de cette compétition est public et est disponible à l'adresse : `https://www.kaggle.com/code/alihigo/nli-xlm-roberta`.

À partir du jeu d'entraînement original, un jeu de validation représentant 10% de ce dernier a été crée. On peut ainsi mesurer les performances de notre modèle sur ce nouveau jeu. On rappelle que le jeu de test fourni par la plateforme n'est pas labelisé et on ne peut donc pas s'en servir pour mesurer les performances. Sur la figure 3 sont montrés les performances *transfer learning* avec le modèle `'joeddav/xlm-roberta-large-xnli'`, qui obtient un score global de 0.8733 sur le jeu de test à la soumission.

Une approche alternative consiste à exploiter les modèles monolingue de l'état de l'art pour palier aux limitations des modèles multilingues. L'idée serait de traduire le contenu non-anglais en anglais et envoyer le contenu anglais au modèle dans le but d'améliorer les performances.



	precision	recall	f1-score	support
0	0.890215	0.855505	0.872515	436.000000
1	0.836735	0.843188	0.839949	389.000000
2	0.890274	0.922481	0.906091	387.000000
accuracy	0.872937	0.872937	0.872937	0.872937
macro avg	0.872408	0.873724	0.872852	1212.000000
weighted avg	0.873069	0.872937	0.872784	1212.000000

Figure 3: Matrice de confusion (À gauche) et résultat des performances du modèle final (À droite) sur le jeu de validation.