RAPPORT D'EXPERIENCE TP 05 - CLASSIFICATION DE TEXTES

Dans un premier temps, en charge la dataset d'allociné comme déjà fait dans le TP 05 au classe, puis on commence le preprocessing du data pour que cette dernière soit valide comme input des models.

le preprocessing des modèles "NBSVM", "LOGREG" et "FASTTEXT" est différent que pour les modèles "BERT" et "CAMEMBERT":

On doit seulement changer l'argument "preprocess_mode": soit "Standard" pour tout les models sauf pour les models BERT c'est "preprocess_mode='bert'".

Model Ensemble :

Il existe plusieurs méthodes pour faire le modèle ensembling, j'ai choisi le "Hardvoting" et "Sample Averaging Method".

Sur les deux méthodes, on reçoit des bonnes résultats comme vu sur les sorties dans le fichier ".ipynb" sur la partie "Model Ensembling"

Data augmentation :

En utilisant la bibliothèque "textattack", l'implémentation est facile depuis la documentation, y a que quelques variables á changer pour que ca convient nos données.

Les résultats sont affichés sur la partie "Data Augmentation" du code.

Partie 3 - TP RICHARD - NER Detection & Classification using NER as features. :

Tout d'abord, on doit extraire les features "NER" des donnés, et pour cela on utilise le modèle "camember-ner".

Après l'extraction, j'ai rajouter pour chaque "review" 4 colonnes (per, loc, misc, org) qui déterminent le nombre de NER dans le texte ou le "review" comme affiché dans le tableau au-dessous

texte	per	loc	misc	org
"je m'appelle Othmane"	1	0	0	0

On pourrait ajouter d'autres méta-informations comme le début du tag NER et sa fin aussi.

À la fin, j'ai concaténé les variables d'input qui sont déjà passés par le preprocessing avec les méta-données. J'avais pour chaque texte ou "review", 404 features :
400 du texte après le preprocessing et +4 de NER tag
Les résultats sont affichés sur dernière partie du TP : "PARTIE 3 - TP RICHARD"

J’ai utilisé que 10% de la data pour ce test mais quand même j’ai reçu de meilleures résultat que le model camembert à base des features du texte seulement sans les features “NER”