Rayan Samy RAMOUL (raysamram@gmail.com)

Test en Deep Learning

Question 1 :

Décrivez un pipeline de détection de langage abusif sur un réseau social. Vous pouvez utiliser des schémas. (votre réponse ne doit pas dépasser une page)

Une pipeline de détection de langage offensant sur réseaux sociaux serait ainsi :

- Tout d'abord il faut un jeu de données, il y a plusieurs jeux de données existants pour chaque langue et disponibles sur des sites comme kaggle ou UCI Machine Learning. On peut aussi créer soi-même notre dataset et ce, en scrappant par exemple des commentaires sur des produits amazon ou encore des critiques sur des sites (cinéma par exemple) et en associant leurs texte au degré de négativité correspondant à la note.
- Le pre-processing consiste ensuite à effectuer une tokenization des phrases afin de les split en mots, une lemmatization pour extraire des bases de chaque mot contextualisés.
- Il faut ensuite purifier le jeu de données des caractères qui ne sont pas intéressants pour cette classification (localisations, noms propre .. etc.)
- Ensuite il faut créer une représentation vectorielle et numérique significative pour ça nous avons 2 possibilités : création au travers d'algorithmes classiques comme TF-IDF, ou utiliser un modèle qui va créer cette représentation au travers du word embedding (en utilisant par exemple word2vec) et en faire un pooling pour n'avoir qu'un seul vecteur.
- Enfin y ajouter un padding de 0 pour égaliser toutes les longueurs.
- Après cela vient l'étape machine learning, pour cela déjà on effectue un split entre les sets train/test/validation.
- Sur le jeu de train on va entrainer un modèle de machine learning (SVM ou Random Forest par exemple) puis on évaluera la performance du modèle sur le jeu de test.
- Si la performance (évaluée au travers de metrics comme la précision ou la cross validation pour prendre en compte les possibles déséquilibres de classes) n'est pas satisfaisante on ajuste les hyper-paramètres spécifiques au modèle.
- Dans un second temps si on n'arrive toujours pas à de bon résultats on peut alors faire appel à des réseaux de neurones profonds comme des LSTM, qui couplés à un word embedding sauront garder une temporalité et une information historique sur la phrase/le paragraphe, ou alors comme suggéré par certaines publications faire un tri plus précis des mots (ne prenant en compte que des mots avec assez de répétition) ou en combinant convnets + LSTM.
- On finit par effectuer des vérifications manuelles en calculant pour des exemples tirés des réseaux sociaux (que ce soit scrappé dans un second temps, dans un autre réseau social pour vérifier la généralisation, ou d'un autre jeu de données) pour étudier l'efficacité du modèle et on effectue une analyse des variations de phrases (suppression / ajout de mots) qui font varier les probabilités et la confidence de notre modèle en ses prédictions afin de déterminer les éventuels biais existants (et trier en conséquence le jeu de train).

Question 2 :

Expliquez le processus permettant d'utiliser un algorithme hybride CNN-LSTM pour détecter un contenu agressif dans une publication.

- Preprocessing (en suivant un processus assez similaire à celui expliqué dans la question 1) et avoir un word embedding pour chaque mot d'une phrase avec un padding pour avoir une taille fixe.
- Pour chaque mot on calcule son embedding équivalent, on le concatene à celui des autres mots de la phrase pour créer une matrice.
- Utiliser un modèle pour l'extraction de caractéristiques des phrases permettant de décerner des représentations spatiales.
- Cet embedding est ensuite utilisé comme input du modèle de convolution pour avoir en output une matrice de taille : nombre de caractéristiques extraites * nombre de mots fixé.
- Le modèle LSTM est utile pour pouvoir garder une information temporelle et séquentielle lors des prédictions (comme des informations de type n-gram, mais qui pourrait être amélioré en faisant appel à des modèles d'attention) , il reçoit alors en input l'output du CNN (par conséquent sa taille d'input et nombre de layers est fixé au nombre de mots fixés) et côté output une sortie binaire avec un softmax affectant des probabilités complémentaires entre message haineux ou non.

Question 3 :

Fichier notebook joint à cet email.