Faculté d'Informatique SII 2023/2024
Département IASD **Module** AARN

Projet de fin de semestre

Introduction

Pour ce projet, l'objectif est d'utiliser les notions apprises durant ce semestre pour résoudre un problème de sciences de données. Les étudiants en binôme devront implémenter un classifieur de tweet dans le contexte de l'analyse des sentiments.

Étapes à réaliser :

Étape 1 : préparation des données

Avant de commencer une tâche d'apprentissage automatique, il est généralement judicieux de jeter un œil à des exemples de l'ensemble de données. La Figure 1 montre un exemple de tweets avant et après prétraitement.

Before preprocessing	After preprocessing
#trump clone will be online for few hours, feel free to live chat on twitch with him: #deeplearning #deepfakeâe| https://t.co/quxwoazwd8	clone online hour feel free live chat twitch
New music from @brendan_m96 on the way!! ŸŽ¶Ÿ′ Ÿ"¥â³Ÿ Ž¶#deepfake #NewSingle #cantwait https://t.co/nCFAgEpGtz	music
@Fakepix @disclosetv Manipulated. Why is there such desperation to twist and turn photos and print to serve a far râe| https://t.co/v8aKOHz8Eu	manipulated desperation twist turn photo print

Figure 1. Exemple d'un tweet

A cet effet, la normalisation ou nettoyage qui doit être implémentée dans les étapes suivantes. Il faut noter qu'on ne garde que la caractéristique texte du dataset initial, et qui va nous servir à construire notre modèle.

Nettoyage des tweets :

Minuscule : l'intégralité du tweet devra être convertie en minuscules.

Suppression de balises HTML : Toutes les balises HTML devront être supprimées.

Radicalisation de mots : Les mots devront être réduits à leur forme radicale. Par exemple, "discount",

Pour toute question, envoyez un mail à l'adresse hw.moulai@gmail.com
Dr. H.MOULAI

"discounts", "discounted" et "discounting" devront être tous remplacé par " discount", et "include", "includes", "included", et "ncluded" devront être tous remplacés par « includ ».

Suppression des non-mots : les non-mots et la ponctuation devront être supprimés. Tous les espaces blancs (onglets, nouvelles lignes, espaces) devront être remplacés par un seul espace.

Suppression des mots vide : les mots tel que « a, the, this,… » devront être supprimés.

D'autres prétraitements peuvent être rajoutés à ce niveau.

2.1.1 Construction du vocabulaire

Après le prétraitement des tweets, une liste de mots représentera chaque tweet. L'étape suivante consiste à choisir les mots que nous aimerions utiliser dans notre classificateur et que nous voudrions laisser de côté.

La liste complète du vocabulaire devra être sauvegardée dans un fichier, exemple vocab.txt.

Dans cette liste de vocabulaire seulement les mots qui apparaissent au moins K fois dans le corpus de tweets devront être gardés. K devra être choisi empiriquement.

Une fois **la liste de vocabulaire** obtenu, il faudra mapper chaque mot dans le tweet prétraité à son index dans **une liste d'index de mots** (qui contient l'index du mot dans la liste de vocabulaire).

Ceci est fait en cherchant le mot dans la liste de vocabulaire et trouver si le mot existe. Si oui, il devra être ajouté dans la variable index des mots. Si le mot n'existe pas, et n'est donc pas dans le vocabulaire, le mot devra être ignoré.

2.2 Extraction de caractéristiques

L'extraction de caractéristiques devra convertir chaque tweet en un vecteur dans R^n. Pour ce projet, nous utiliserons n = # mots de vocabulaire liste.

Il existe deux manières de représenter le vecteur caractéristique, une représentation binaire et une représentation par comptage.

Représentation binaire des caractéristiques : la caractéristique $x_i \in \{0,1\}$ d'un tweet correspond à savoir si le i-ème mot du dictionnaire apparaît dans le tweet. Autrement dit, $x_i = 1$ si le i-ème mot est dans le tweet et $x_i = 0$ si le i-ème mot n'est pas présent dans le tweet.

Représentation des caractéristiques par comptage : la caractéristique $x_i \in \{0,..,m\}$ d'un tweet correspond au nombre d'apparitions du i-ème mot du dictionnaire dans le tweet.

Étape 2 : Classification

Une fois les vecteurs caractéristiques obtenus ; il est possible d'utiliser tous les classifieurs appris durant ce semestre :

- Une comparaison des classifieurs devra être faite et les résultats discutés.
- Une modularité du code est exigée.
- L'utilisation des implémentations existantes (librairies) d'algorithmes de classification est permise, toutefois il est important de justifier le choix des libraires ainsi que des paramètres.
- L'utilisation d'algorithmes d'apprentissage profond pour cette tâche est grandement appréciée et une comparaison entre les approches classiques et celles de l'apprentissage profond est encouragée.

Consignes :

Les livrables du projet sont :

- Code source du projet à envoyer par mail à l'adresse **hw.moulai@gmail.com**.
- **Rapport de projet :** décrivant et justifiant les choix des approches et des libraires utilisées ainsi qu'une analyse (synthèse) des résultats obtenus. Le rapport en version papier doit être déposé dans ma boite aux lettres au niveau de la faculté.

Le délai de remise du projet est fixé au **11/05/2024**.