Challenge 2 : apprentissage supervisé

Prévision de la toxicité sur Youtube

Contexte :

Les médias écrits et audiovisuels sont un maillon essentiel du pluralisme politique. Leur numérisation et l'explosion des contenus informationnels en ligne favorisent *a priori* la formation d'opinions plurielles, garantes d'un bon fonctionnement de la démocratie. Pour autant, ce **pluralisme des contenus** et des sources n'est pas synonyme en soi de **qualité de l'information**. C'est à ces questions de pluralisme et de qualité de l'information en ligne que s'intéressent les chercheurs du projet national PIL (http://www.anr-pil.org/) démarré en 2018.
En tant que réseau social en ligne, Youtube constitue un espace public de partage d'information et d'interactions entre utilisateurs, aussi bien propice à la co-création de valeur qu'au développement de comportements toxiques, de propagande politique voire de contenus radicalisés. C'est dans ce contexte que les chercheurs de l'ANR PIL ont extrait une grande base de données de vidéos Youtube issues de 58 chaines de medias français, qui est mise à disposition pour votre challenge.

But du challenge :

L'objectif de ce challenge est de comprendre et prédire la notion de toxicité dans une vidéo Youtube. Il s'agit de déterminer dans quelle mesure l'espace des commentaires des médias français est touché par les débats brutaux et agressifs. Nous posons comme objectif de comprendre où et quand l'agressivité en ligne est retrouvée et quels facteurs la limitent et la favorisent. **D'abord, en abordant la question sous la forme d'un problème de régression visant à prédire le nombre d'insultes présentes dans les commentaires associés à une vidéo. Ensuite, en construisant puis prédisant un indice original de toxicité à partir des variables présentes dans la base.**

Description des données :

Le fichier *challenge_youtube_toxic.csv* comporte 46102 observations (1 ligne = 1 vidéo Youtube issue de 58 chaines de médias français, contenant des commentaires) et 28 variables caractérisant chaque vidéo. 1 vidéo est donc liée à une chaîne de médias (ex : CNews, Le Figaro, L'équipe), chaque chaîne de médias possédant plusieurs vidéos dans le jeu de données. Chaque chaîne analysée appartient à une catégorie institutionnelle (Presse nationale, régionale, magazine, médias alternatifs, pure-players, TV) et est aussi catégorisée par son type de contenu (« cœur », « niche » ou « partisan »).

Chaque observation, *i.e* ligne du tableau de données, représente ainsi une vidéo youtube caractérisée par les colonnes (*features*) suivantes :

[1] "video_id_court" : video id (short)
[2] "video_id" : video id
[3] "channel_id" : id de la chaîne où la vidéo est publiée
[4] "nbrMot" : nombre de mots au total dans tous les commentaires de la vidéo
[5] "nbrMotInsulte" : nombre de mots de type « insulte » au total dans tous les commentaires de la vidéo
[6] "nbrMotAllong" : nombre de mots allongés au total dans tous les commentaires de la vidéo
[7] "nbrMotMAJ" : nombre de mots en majuscule au total dans tous les commentaires de la vidéo
[8] "nbrExclMark" : nombre d'exclamations au total dans tous les commentaires de la vidéo
[9] "nbrQuestMark" : nombre d'interrogation au total dans tous les commentaires de la vidéo
[10] "nbrMotMoyenne" : nombre moyen de mots par commentaire
[11] "nbrMotInsulteMoyenne" : nombre moyen d'insultes par commentaire
[12] "nbrMotAllongMoyenne" : nombre moyen de mots allongés par commentaire
[13] "nbrMotMAJMoyenne" : nombre moyen mots en majuscule par commentaire
[14] "nbrExclMarkMoyenne" : nombre moyen d'exclamations par commentaire
[15] "nbrQuestMarkMoyenne" : nombre moyen d'interrogations par commentaire
[16] "thread_count" : nombre de commentaires top level
[17] "comment_count" : nombre de commentaires qui sont des réponses à un top level
[18] "message_count" : comptage de tous les commentaires postés (top level + dans les fils de discussions)
[19] "discussion_count" : nombre de fils de discussion après la vidéo : les commentaires *top-level* suivis d'au moins une réponse
[20] "distinct_authors_count" : nombre de commentateurs uniques
[21] "authors_3channels_count" : nombre de commentateurs actifs, participant aux discussions sur plus de trois chaînes différentes
[22] "liked_authors_count" : nombre de commentateurs populaires dont les commentaires sont les plus *likés* sur la plateforme : les membres avec le nombre de *likes* supérieur à la moyenne sur la plateforme
[23] "channel_name" : nom de la chaîne d'appartenance de la vidéo
[24] "subscriberCount" : nombre d'utilisateurs enregistrés sur la chaîne associée
[25] "viewCount" : nombre de vues de la chaîne
[26] "categorie_new" : catégorisation des chaînes par « Coeur, niche, partisans »
[27] "categ_inst" : catégorisation des chaînes par "Presse nationale, régionale, etc »
[28] "X" : colonne inutile, à supprimer.

Travail à effectuer

Le travail consiste à mettre en place une méthodologie complète d'apprentissage supervisé appliquant les méthodes vues en cours. Nous proposons de suivre les étapes suivantes et indiquons un barème indicatif pour chacune d'entre elles :

Statistiques descriptives et feature engineering (2 points)

1. Statistiques uni- et multi-dimensionnelles : évaluation de la qualité des données, compréhension de la structure, des liens entre variables
2. Recodage des variables, transformation, création éventuelle de nouvelles variables

Benchmark des méthodes de régression pour prédire le nombre d'insultes liées à une vidéo (3 points)

Le but de cette partie est de prédire la variable « nbrMotInsulte », en formulant donc le problème sous la forme d'une régression. Vous suivrez une méthodologie classique d'apprentissage supervisé (train/test), en appliquant différentes méthodes vues en cours :

- Régression logistique
- kNN Regression
- Support Vector Regression
- Arbres de régression et random forest
- Gradient Boosting
- Réseaux de neurones et deep learning

Il s'agit de proposer le meilleur modèle possible en termes de généralisation sur un ensemble de test, c'est-à-dire d'être capable de prédire le plus précisément possible le nombre d'insultes pour une nouvelle vidéo.

Création d'un nouvel indice de toxicité et classification (3 points)

C'est la partie créative du challenge, qui départagera les meilleures équipes. Nous vous demandons de créer une nouvelle variable, à partir des variables de départ, qui sera selon vous plus représentative du caractère toxique d'une vidéo. Votre indice composite pourra prendre la forme d'une variable binaire (ex : toxic 0/1) ou catégorielle (ex : toxic faible/moyen/fort). Vous transformez alors le problème de régression en un problème de classification et pourrez appliquer les méthodes vues en cours. Vous proposerez un modèle capable de prédire la toxicité d'une nouvelle vidéo et l'évaluerez en généralisation, sur un ensemble test. Vous pourrez appliquer si besoin une méthode de rééquilibrage de classes et calculerez les métriques de performances habituelles (matrices de confusion, F1 score etc.).

Qualité du code et de l'analyse (2 points)

Une attention particulière sera portée sur la qualité de votre analyse, vos idées sur le problème proposé. Un notebook très bien commenté sera le rendu minimum.