Projet Data Science
HMIN232 - 2020/2021

Classification d'assertions selon leur valeurs de véracité (*automatic fact-checking*)

Ce projet a pour but de proposer des modèles de classification supervisée d'assertions faites par des figures politiques selon leur valeur de véracité, ou autrement dit, de proposer une approche de *fact-checking* automatique.

1. **Données**

Le jeu de données utilisé est ClaimsKG. Il a été collecté à partir de sites de fact-checking (tels que www.politifact.org ou www.snopes.org) par le LIRMM en collaboration avec plusieurs équipes de recherche européennes. Il est décrit en détail ici : https://data.gesis.org/claimskg/site/.

Il est important de lire attentivement la description afin de bien comprendre les différentes composantes.

Les données, en format CSV, sont disponibles sur le lien suivant :

http://tinyurl.com/3smozlx6

Il est également possible de faire vos propres extractions à partir de l'interface web : https://data.gesis.org/claimskg/explorer/home

2. **Feature engineering**

Vous êtes invités à utiliser comme source de features le **texte** de chaque assertion (indexer les assertions et sélectionner les termes les plus informatifs) mais pensez également à utiliser les **métadonnées associées** (auteur de l'assertion, date, entités nommées mentionnées dans le texte de l'assertion ou bien dans l'articles de fact-checking l'accompagnant, sources externes et les références, les mots clés, etc.)

Pour préparer les données textuelles, pensez aux **pré-traitements** vus en cours : stop words filtering, lemmatization, n-grams, POS tagging, etc.

Pour aller plus loin, vous pouvez éventuellement penser aux mesures de crédibilité des auteurs ou des sources que l'on peut retrouver dans la littérature, par exemple dans :

- A. Kirilin and M. Strube. Exploiting a speaker's credibility to detect fake news. *DSJM*, 2018.

3. **Classification et sélection de variables**

Nous allons nous intéresser à trois tâches de classification :

1. {VRAI} vs. {FAUX} (deux classes)
2. {VRAI ou FAUX} vs. {MIXTURE} (deux classes)
3. {VRAI} vs. {FAUX} vs. {MIXTURE} (trois classes)

Dans les trois cas, il faudra classer les assertions en groupes selon les labels. Pensez bien à vérifier que les instances sont labellisées selon les catégories indiquées et éventuellement apportez les modifications nécessaires.

Attention, vos données d'apprentissage risquent de ne pas être équilibrées, i.e. il peut y en avoir plus d'une classe que de l'autre. Quelle solution proposeriez-vous ? Idée : Penser au *upsampling* et/ou au *downsampling*.

Vous pouvez utiliser les **modèles de classification** vus en cours, tels que les arbres de décision, les SVMs, le Naïve Bayes, les K-NN, etc. mais ne vous censurez pas et vous être libre de faire appel à d'autres approches de classification (par exemple, les réseaux de neurones).

Pour chacune de ces trois tâches, en plus de vos modèles de classification, préparer une liste de features discriminantes en ordre décroissant. Pour cela, vous pouvez vous appuyer sur des méthodes de **sélection de variables** (ou de features). Le plus important est de tirer les conclusions. Qu'en concluez-vous en comparant les listes obtenues pour les deux tâches ?

4. **Analyse**

La partie `analyse` de votre projet consiste à comparer empiriquement les différents choix faits au niveau de features, pré-traitements, modèles et data sampling par rapport à leur impact sur la qualité de la classification obtenue. Cette analyse sera présentée de manière synthétique et lisible (à l'aide des tables et/ou des courbes). Essayez de "creuser" les raisons du pourquoi et comment un modèle ou traitement ou ensemble de features seraient plus pertinents qu'un autre (dans ce contexte-là).

5. **Organisation et rendu**

    - Le travail s'effectuera en groupes de **3 ou 4 étudiants**.
    - Une soutenance orale de 15 minutes suivie de 10 minutes de questions est prévue à la fin du semestre, si cela sera possible en fonction de l'évolution de la crise sanitaire. La soutenance a pour objectif de présenter vos approches, vos choix et de mettre en avant également l'analyse des résultats que vous avez obtenu. Il est inutile de perdre du temps lors de la présentation sur les données initiales (qui sont communes) ni sur la problématique du projet ou bien la théorie des méthodes utilisées.
    - Le rendu final sera soumis sous la forme d'un fichier compressé (gzip) identifié par le numéro du groupe **à déposer sur Moodle au plus tard 3 jours avant la soutenance** consiste en :
      (1) Un rapport de **max 10 pages**

(2) Le notebook au pdf et ipynb de vos codes de l'ensemble des traitements automatiques

- Attention à bien mettre le prénom, nom et numéro d’étudiant de chaque personne du groupe dans les documents rendus.