Projet Machine Learning

HAI817 - 2021/2022

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

Projet en groupe (4 à 5 étudiants)

Ce projet s'inscrit dans le contexte de l'apprentissage supervisé, i.e. les données possèdent des labels. Il vise à trouver les modèles les plus performants pour prédire si des assertions (une assertion est une proposition que l'on avance et que l'on soutient comme vraie) faites par des hommes politiques (par exemple) sont vraies ou fausses.

1. **Les 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/ et dans l'article suivant (facilement trouvable sur Google) :

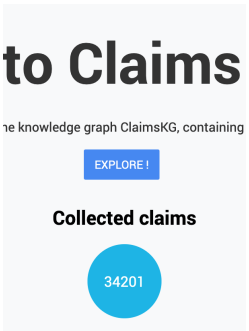
Tchechmedjiev, A., Fafalios, P., Boland, K., Gasquet, M., Zloch, M., Zapilko, B., ... & Todorov, K. (2019, October). ClaimsKG: a knowledge graph of fact-checked claims. In *International Semantic Web Conference* (pp. 309-324). Springer, Cham.

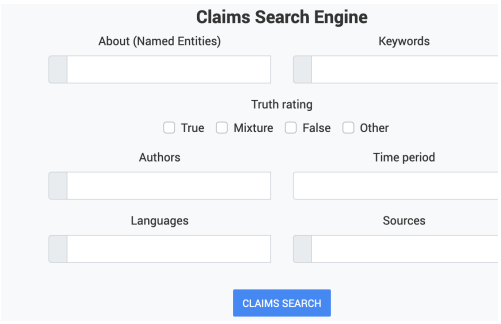
Attention : il est important de lire attentivement la description du jeu de données afin de bien comprendre à quoi correspondent les différents attributs.

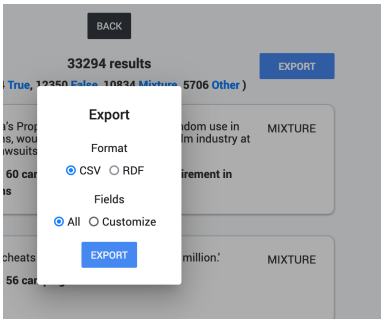
Vous pouvez extraire les données en choisissant leur tailles, équilibre et contenus au format CSV à partir de l'interface web : https://data.gesis.org/claimskg/explorer/home

(1) Cliquer sur "Explore"; (2) Choisir vos paramètres dans les facettes ou laisser vide et puis cliquer sur "Claims search"; 3) Cliquer sur "EXPORT", puis sur "Customize" - choisissez les champs (variables) à inclure dans vos données. L'outils permet une extraction de 10K claims maximum en une seule fois. Chaque groupe est libre de choisir son jeu de données.

(1) (2) (3)







2. **Ingénierie des données**

Le jeu de données contient bien entendu le **texte** de chaque assertion mais également le **texte de l'article** qui l'accompagne. Il est fortement conseillé d'utiliser ces deux informations comme source. De plus il existe de nombreuses **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, les topics, etc.) qui peuvent s'avérer importante pour améliorer les performances de vos modèles.

N'oubliez pas que pour préparer des données textuelles, il existe de nombreux pré-traitements (élimination des stop words, lemmatisation, n-grammes, etc.) vus en cours et disponibles dans les notebook (e.g. ingénierie des données textuelles).

De la même manière n'oubliez pas que comme vous ne connaissez pas les données, il est indispensable de tester plusieurs classifieurs pour voir celui ou ceux qui ont de meilleures performances (e.g. notebooks premières classification, classification de données textuelles) pour au final définir une chaîne de traitement complète adaptée à vos données.

Les notebooks sont là pour vous aider. N'hésitez pas à les consulter.

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. **Les tâches de classification**

Nous nous intéressons à trois tâches de classification (voir la **NB** plus bas) :

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 dans une classe que dans l'autre. Quelle solution proposeriez-vous ? Idée : Pensez à l'*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, les random forest, etc. Ne vous censurez pas, vous pouvez utiliser d'autres approches de classification (par exemple, les réseaux de neurones), si vous le souhaitez.

N’oubliez pas de bien évaluer vos modèles. L’accuracy n’est pas suffisante. Pensez à la matrice de confusion, au rappel, à la précision, à la F-mesure.

BONUS: Pour chacune des trois tâches de classification, en plus de vos modèles de classification, préparez 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 ?

NB: Les labels de véracité utilisés par les différents sites de fact-checking ont des formes linguistiques différentes. Nous vous conseillons de vous appuyer sur des labels normalisés qui contiennent 4 valeurs uniquement: “TRUE”, “FALSE”, “MIXTE” et “OTHER”. Les mappings entre les labels d’origine et ces 4 catégories sont disponibles ici:

https://github.com/claimskg/claimskg_generator/blob/master/claimskg/generator/ratings.py

4. **Analyse des erreurs, validation et comparaisons des modèles**

La partie `analyse` de votre projet consiste à comparer empiriquement les différents choix que vous avez pu faire dans la partie sélection des features, des prétraitements, des modèles utilisés, de l'échantillonnage, etc. par rapport à leur impact sur la qualité de la classification.

Cette analyse devra être présentée de manière synthétique et lisible à l’aide d’un tableau comparatif et/ou des courbes. Il est important d’essayer de “comprendre" les raisons des résultats obtenus en fonction des choix effectués (par exemple : Pourquoi ce modèle se comporte mieux ou moins bien qu’un autre ? Pourquoi la suppression des stop words améliore ou au contraire n’améliore pas les résultats ? etc). Cette prise de recul sera particulièrement prise en compte lors de l’évaluation.

5. **Organisation et rendu**

   - Le travail s'effectuera en groupes de **4 à 5 étudiants**.
   - Une soutenance orale de 15 minutes suivie de 10 minutes de questions est prévue à la fin du semestre. 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 **les noms des membres** du groupe **à déposer sur Moodle au plus tard 3 jours avant la soutenance** consiste en :
     (1) Un rapport de **max 8 pages**
     (2) Le notebook en 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.