Projet Machine Learning

HAI817 - 2024/2025

P. Poncelet, K. Todorov, E. Raoufi

Classification d'assertions venant d'X (Twitter) selon leur rapport à la science

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.

Attention : il est impératif d'inscrire la composition du groupe sur le Google sheet suivant : https://docs.google.com/spreadsheets/d/1f68uAT2f_gK7Bk47Iu4u6LuB1MckftdmD2_uF7lOjJc/edit?usp=sharing

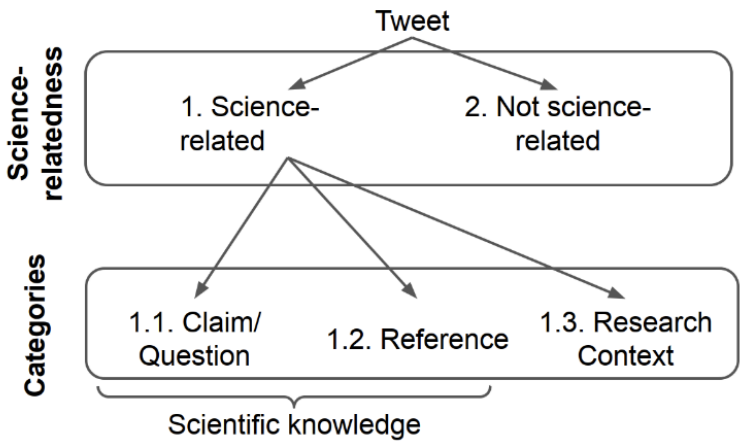
La personne de la première colonne est chargée de déposer le rendu.

1. **Les données**

Le jeu de données utilisé est SciTweets ("dataset scitweets" sur Moodle). Il a été collecté à partir de X (anciennement Twitter) par le LIRMM en collaboration avec l'institut de recherche en sciences sociales GESIS (Cologne, Allemagne). Le jeu de données est décrit en détail dans le papier suivant, facilement trouvable en pdf sur google :

Hafid, S., Schellhammer, S., Bringay, S., Todorov, K., & Dietze, S. (2022). SciTweets-A Dataset and Annotation Framework for Detecting Scientific Online Discourse*. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (pp. 3988-3992).*

Attention : il est important de lire attentivement la description du jeu de données afin de bien comprendre à quoi correspondent les différents attributs et les différents labels. Notamment, il faut bien comprendre la définition d'assertions relatives à la science. Les données sont labélisées selon la définition hiérarchique donnée dans l'article.



Comme vous pouvez le constater, il s'agit d'une tâche multi-classes (nous avons plusieurs classes) et multi-labels (une instance peut appartenir à plusieurs classes). Notamment, les tweets "scientifiques" peuvent appartenir à une, deux ou bien trois sous-catégories (1.1. CLAIM, 1.2. REF, et 1.3. CONTEXT), alors qu'un tweet

quelconque peut être classifié comme scientifique ou non (1. SCI et 2. NON-SCI).

2. **Ingénierie des données**

Le jeu de données contient bien entendu le **texte** de chaque assertion ainsi que les **labels**. Les données proviennent directement de X/Twitter sans filtrage, et contiennent alors des caractères spéciaux, des hashtags, des liens et des emojis. Vous pouvez faire le choix de remplacer les emoji par des chaînes de caractères à l'aide de dictionnaires ou bien de les éliminer.

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.

3. **Les tâches de classification**

Nous nous intéressons à trois tâches de classification :

1. {SCI} vs. {NON-SCI} (deux classes se situants au plus haut niveau de la hiérarchie représentée sur la figure plus haut)
2. {CLAIM, REF} vs. {CONTEXT} (deux classes), uniquement pour les données dans la catégorie SCI
3. {CLAIM} vs. {REF} vs. {CONTEXT} (trois classes), uniquement pour les données dans la catégorie SCI

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 ?

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**. Un travail ne peut pas être rendu si vous n'avez pas au préalable rempli la composition du groupe dans le Google Sheet qui est au début du document car vous devez récupérer le numéro de votre groupe.
   - Le rendu final sera soumis sous la forme d'un fichier compressé (gzip) identifié par **le numéro du groupe** du groupe (par exemple groupe1.gzip) **à déposer sur Moodle au plus tard le 5 mai 2025** consiste en :

     (1) Un rapport d'un **maximum de 8 pages (sans compter la page de garde).** Le rapport ne doit pas décrire les données ni les méthodes, mais se focaliser uniquement sur vos analyses et résultats.
     (2) Un lien YouTube ou autre avec une vidéo de **10 minutes max** par groupe reprenant les codes d'une soutenance :
     (3) 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.
   - **Tout rendu ne respectant pas les consignes sera fortement pénalisé.**