Master Bioinformatique

HAI817 – Machine learning 1 (méthodes classiques)

Professeurs : P. Poncelet, K. Todorov, E. Raoufi

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

---

Par : Tiziri Tamani (22415178), Ciarán Mahony (22400729), Raphaël Ribes (22401925), Dalia Belmadi (22208849)

Lien vers le projet : `https://github.com/RaphaelRibes/machine-learning/`

Table des matières

Résumé

L'information sur Internet et les réseaux sociaux pose de nouveaux défis en matière de traitement automatique du langage. Dans ce contexte, il devient essentiel de pouvoir automatiquement classifier le contenu textuel, notamment pour identifier des textes à caractère scientifique ou non. C'est dans cette optique que s'inscrit notre projet, qui vise à analyser des articles ou extraits textuels, et à déterminer s'ils sont liés au domaine scientifique ou non.

1- Résultats

A) Pré-traitement

Dans notre projet de classification de tweets en scientifiques et non scientifiques, nous avons appliqué trois types de classification et à chaque étape, un bon prétraitement s'est révélé indispensable pour améliorer la qualité des données et les performances des modèles. En effet, les tweets bruts contiennent souvent du bruit (liens, hashtags, mentions, emojis, etc.), ce qui peut perturber l'apprentissage.

Nous avons donc appliqué les étapes suivantes, adaptées à chaque tâche :

— Nettoyage textuel et des *stop words*
— Normalisation linguistique : Lemmatisation et tokenisation
— Vectorisation avec TF-IDF et paramétrage des n-grammes
— Gestion du déséquilibre

Des tests ont été effectués dans le cadre de cette classification en utilisant uniquement les attributs importants extraits à l'aide du classifieur Random Forest. Cependant, les résultats n'étaient pas très satisfaisants. Nous avons donc envisagé de réduire manuellement la taille maximale des features lors de la vectorisation, afin d'observer le comportement du modèle. Cette décision se justifie par la présence de bruit dans les données, bruit qui reste néanmoins représentatif pour distinguer les tweets scientifiques des non-scientifiques. C'est pourquoi nous avons choisi de ne pas effectuer une classification uniquement à partir de ces attributs.

Ces étapes ont été adaptées à chaque type de classification (Scientifique versus Non Scientifique, Affirmation et Référence versus Contexte, Affirmation versus Référence versus Contexte) afin d'obtenir les meilleures performances. Dans la suite de ce rapport, une comparaison entre les données brutes et les données après traitement sera présentée, accompagnée d'une justification des choix effectués. Par la suite pour chaque classification, on effectuera une évaluation comparative des performances de classification.

On regarde le nombre de @ et lien par tweet selon le type de tweet scientifique (figure 1).

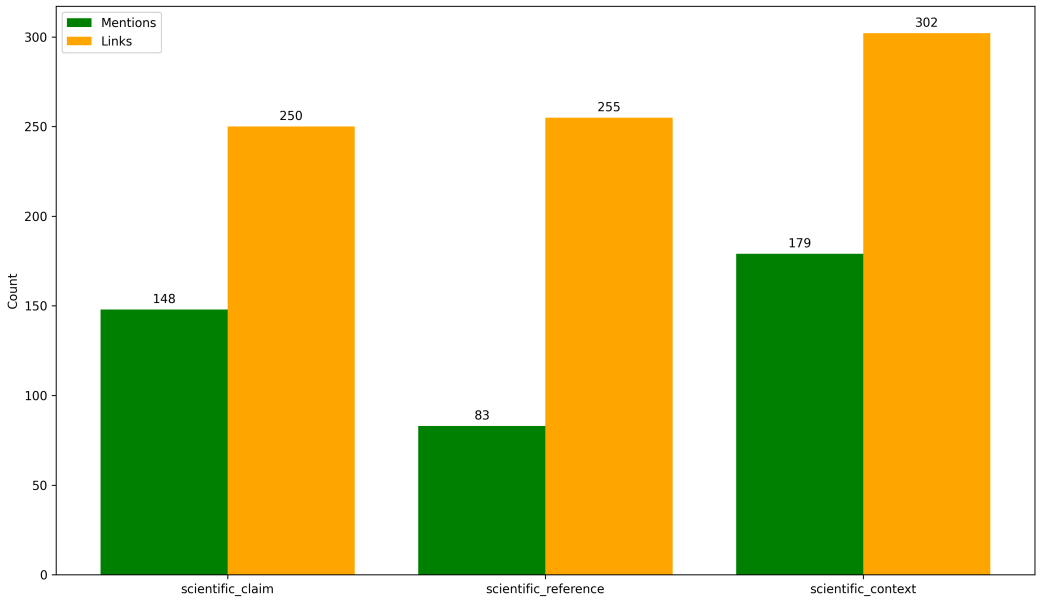

FIGURE 1 – Nombre de @ et de lien par tweet selon le type de tweet scientifique

Après avoir réalisé des tests d'indépendance de student sur chaque variable, les *p-values* ne descendent pas en dessous de 0.18 donc on ne remarque pas de différence significative entre les tweets scientifiques et non scientifiques. On peut ainsi

conclure que ces variables ne sont pas pertinentes pour la classification des tweets scientifiques, on les retirera de notre dataset.

B) Modélisation

Dans cette section, nous comparons les performances des modèles sur trois tâches de classification hiérarchiques successives. Par la suite, nous présentons les résultats de classification obtenus pour chaque tâche ainsi qu'une comparaison des performances des approches utilisées.
Pour chacunes des étapes, nous allons comparer différents modèles entre eux pour sélectionner le meilleur modèle. Par meilleur modèle, on entend la meilleure précision et le plus petit écart-type. Chaque modèle est testé par cross-validation sur 10 itérations.

C) Modèle 1 : Scientifique versus Non Scientifique

Dans cette tâche, nous avons comparé plusieurs modèles de classification afin de distinguer les tweets scientifiques des non-scientifiques. L'objectif était d'évaluer les performances de différentes approches d'apprentissage supervisé appliquées à un problème de classification binaire.

C).1 Préparation des données

Nous avons supprimé les mentions (@), les caractères spéciaux et converti l'ensemble des textes en minuscules, pour standardiser l'entrée pour éviter que des variations superficielles n'influencent la classification. Une liste de stopwords a été utilisée, combinant celle de NLTK et des termes propres à Twitter ("http", "https", "rt", "co", "amp", "via") qui sont fréquents mais peu informatifs pour déterminer la nature scientifique du contenu. Pour la vectorisation, nous avons utilisé la méthode TF-IDF avec des n-grammes de taille 1 à 2, un choix raisonnable pour capturer à la fois des mots individuels et des associations fréquentes, tout en limitant la complexité.

Étant donné le fort déséquilibre des classes (la classe SCI représentant moins de 10 %), la technique SMOTE a été utilisée pour équilibrer la distribution entre classes SCI et NON-SCI.

C).2 Analyse comparative des performances des différents modèles

Plusieurs modèles de classification ont été testés (tableau 1). Chacun a été optimisé à l'aide d'une recherche par grille (GridSearchCV) et évalué à l'aide d'une validation croisée à 10 plis (KFold), afin d'identifier les meilleures combinaisons d'hyperparamètres.

Table 1 – Performances des modèles - Tâche 1 (KFold = 10)

Modèle	Accuracy (figure 2)	Précision (figure 3)	Rappel (figure 3)	F1-score
Logistic Regression	0.9327 ± 0.0163	0.93	0.93	0.93
Naive Bayes	0.9118 ± 0.0197	0.91	0.90	0.89
k-NN	0.8111 ± 0.0286	0.82	0.79	0.78
Random Forest	0.8575 ± 0.0246	0.89	0.86	0.86
SVM	**0.9431 ± 0.0137**	**0.92**	**0.92**	**0.92**

Le modèle SVM semble être le plus performant pour cette tâche, affichant les meilleurs scores en précision, rappel et F1-score, ainsi qu'une accuracy élevée accompagnée mais un plus grand écart-type que la régression logistique figure 2. C'est pour cette raison de stabilité que l'on préfèrera la régression logistique car elle présente une baisse de performance négligeable comparé à la stabilité du modèle.

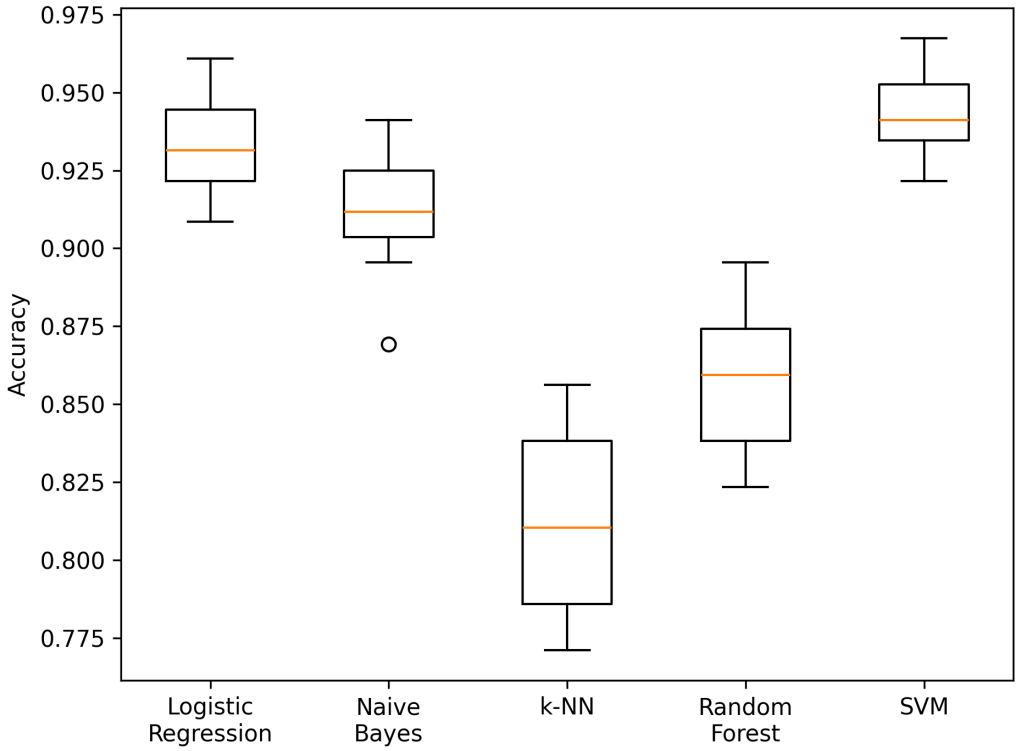

FIGURE 2 – Comparaison des modèles pour la classification des tweets scientifiques et non scientifiques.

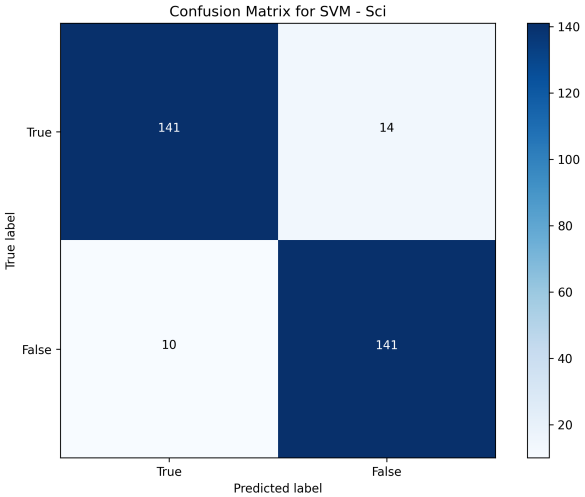

FIGURE 3 – Matrice de confusion du modèle SVM pour la classification des tweets scientifiques et non scientifiques.

D) Modèle 2 : Affirmation et Référence versus Contexte

À présent, nous allons considérer les classes "Affirmation" et "Référence" qui citent une affirmation scientifique d'une part, et la classe "Contexte" qui servent de contexte D'autre part, toujours afin de réaliser une classification binaire. Les tweets analysés sont uniquement ceux classés comme scientifiques dans la première tâche.

D).1 Préparation des données

Le prétraitement textuel a suivi une logique proche de celle adoptée dans la première tâche (minuscules, stop words, lemmatisation), mais sans suppression des mentions et caractères spéciaux, afin d'explorer s'ils peuvent porter une valeur contextuelle utile.

La vectorisation a été réalisée à l'aide de TF-IDF, en élargissant la fenêtre aux trigrammes (1 à 3-grammes) afin de mieux capturer les structures linguistiques propres aux énoncés de type affirmation ou citation.

Contrairement à la première tâche, nous n'avons pas appliqué SMOTE Ici, le déséquilibre est partiel dans la distribution des combinaisons de labels (tableau 2,) car la combinaison 1,1 est majoritaire, tandis que la combinaison 0,1 est fortement sous-représentée. La classification est multi-label (un tweet peut être à la fois CONTEXT et CLAIM/REF). Nous avons donc implémenté un rééchantillonnage manuel spécifique au multi-label, en nous basant sur l'identification de toutes les combinaisons de labels possibles et en suréchantillonnant celles qui étaient sous-représentées.

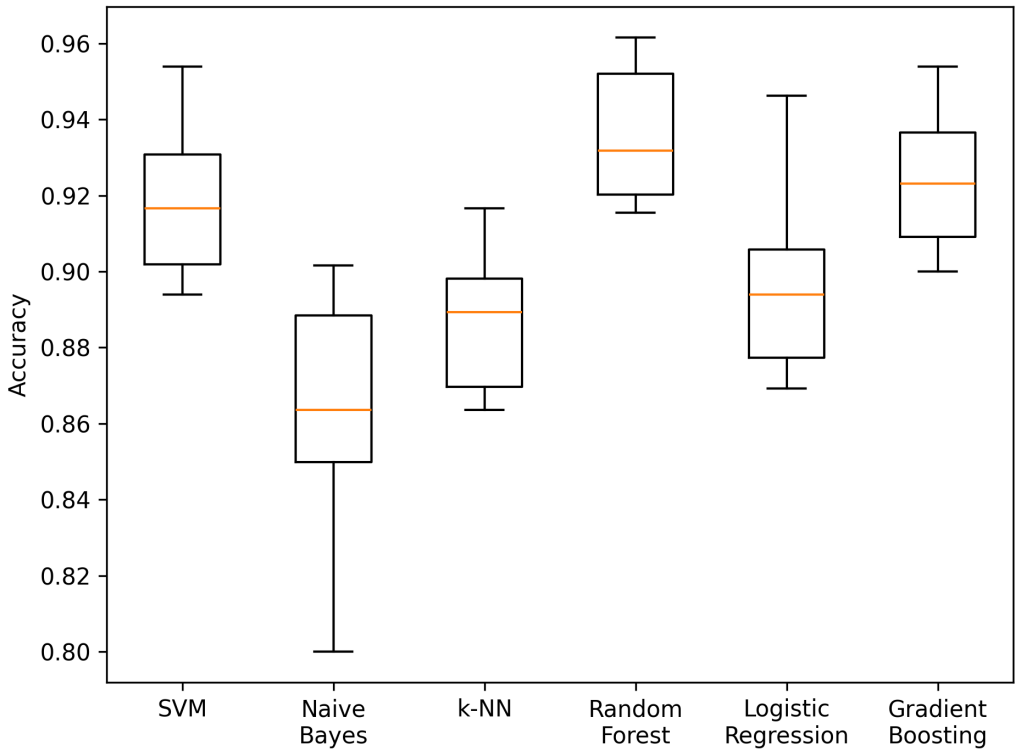

FIGURE 4 – Comparaison des modèles pour la classification des tweets claim et ref vs contexte.

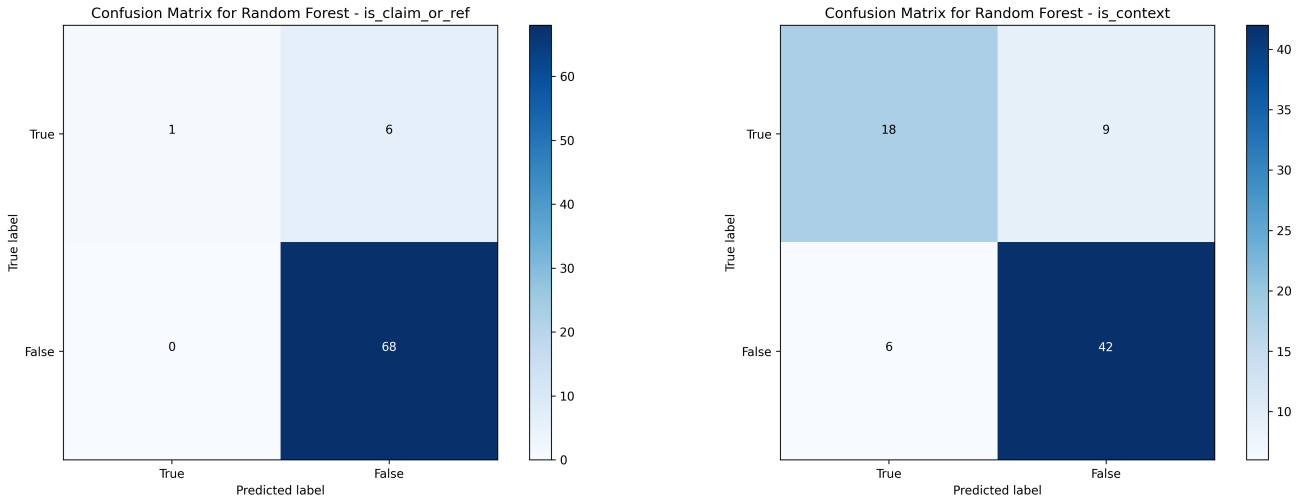

FIGURE 5 – Matrice de confusion du modèle Random Forest pour la classification des tweets claim et ref vs contexte.

E) Modèle 3 : claim vs ref vs contexte

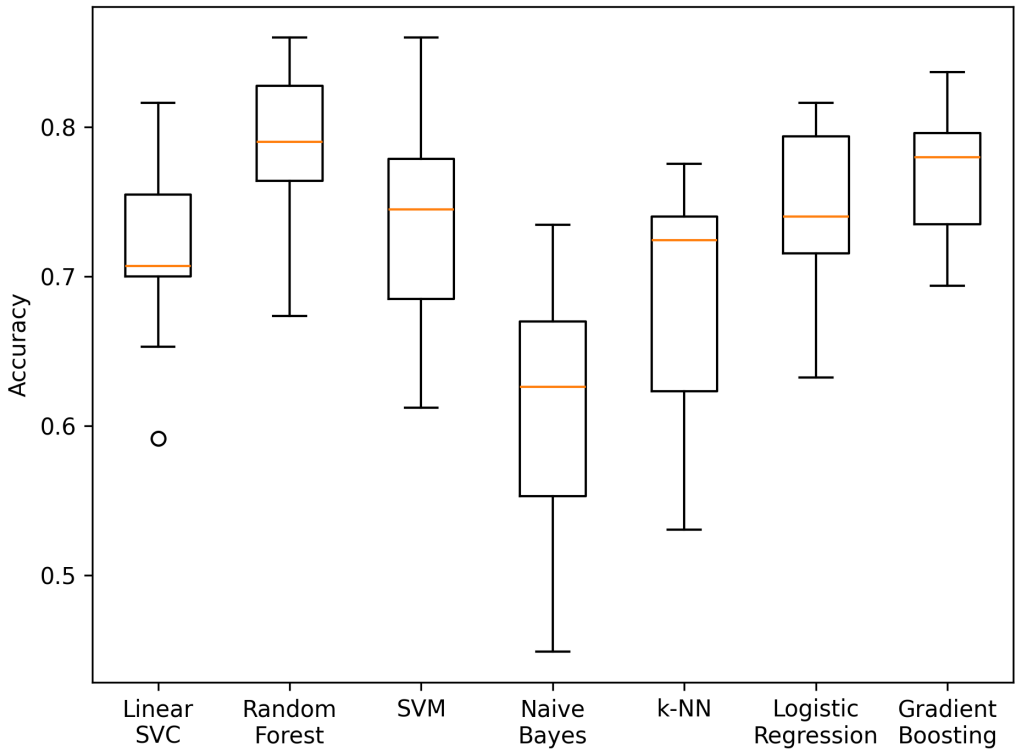

FIGURE 6 – Comparaison des modèles pour la classification des tweets claim vs ref vs contexte.

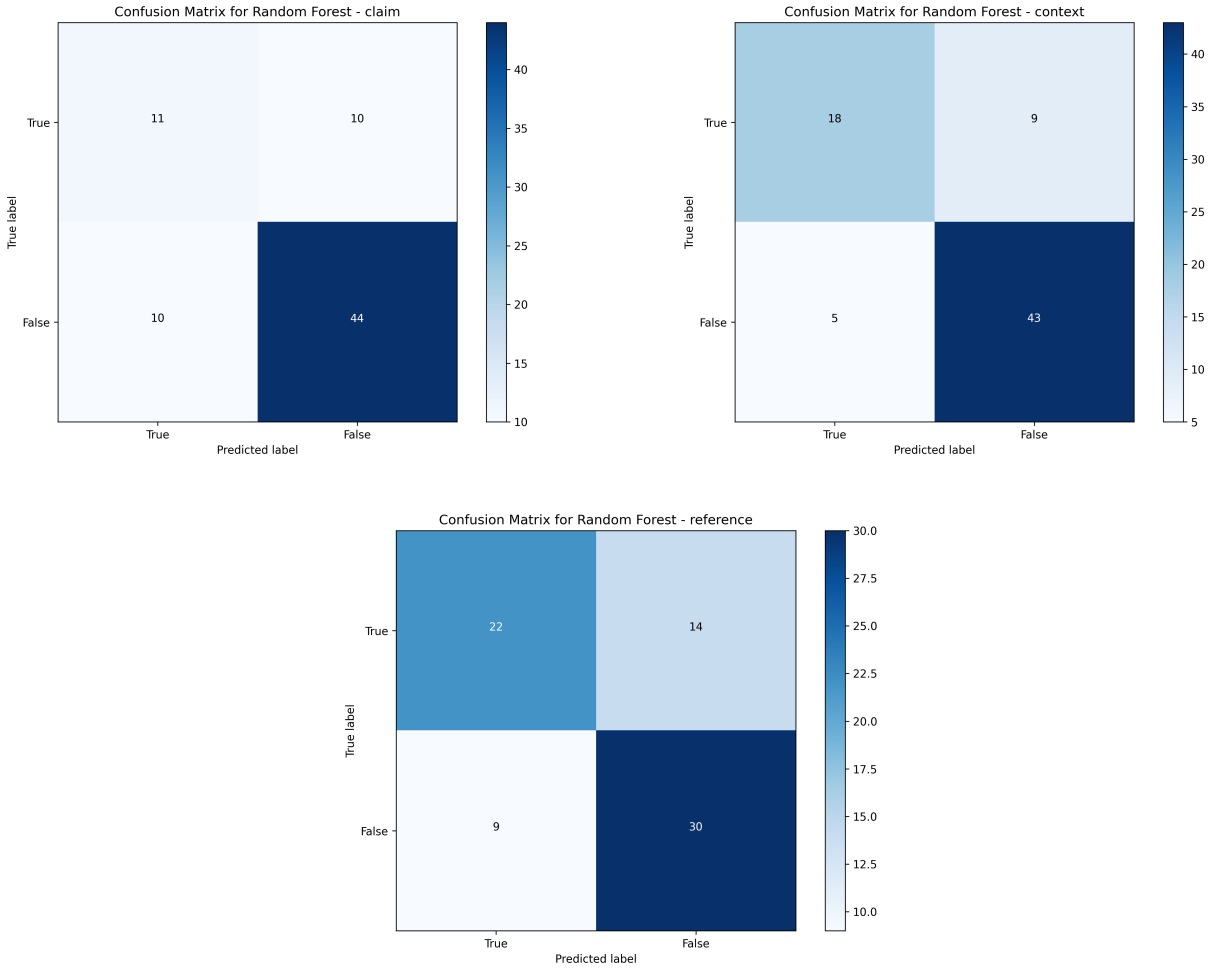

FIGURE 7 – Matrice de confusion du modèle Random Forest pour la classification des tweets claim vs ref vs contexte.

2- Discussion

3- Conclusion