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/`

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- Analyse et Évaluations des 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.

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) Tâche 1 : Classification binaire (SCI vs. NON-SCI)

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 1)	Précision (figure 2)	Rappel (figure 2)	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 du plus faible écart-type. Ces performances stables, observées à la fois sur le jeu de test et au cours de la validation croisée, sont clairement visibles dans le boxplot de la figure (figure 1). En comparaison, la régression logistique s'est également montrée très efficace, représentant une alternative pertinente, notamment dans des cas où l'on recherche un bon compromis entre précision et rappel.

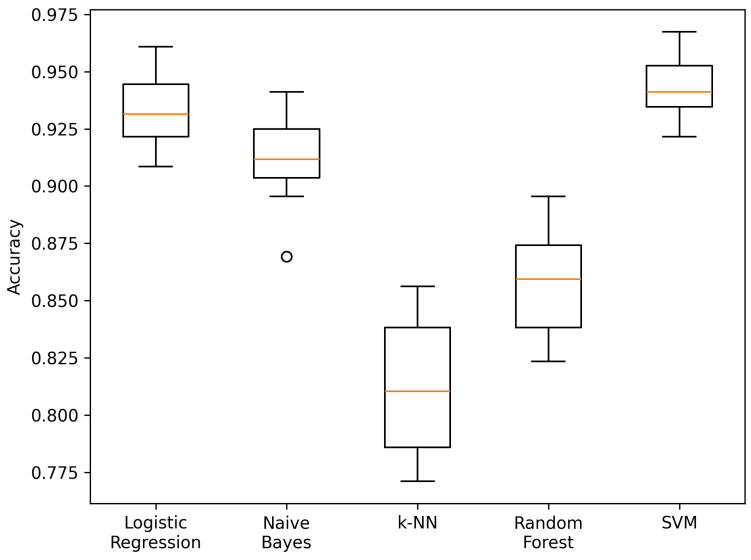

FIGURE 1 – Comparaison des modèles pour la classification des tweets scientifiques et non scientifiques.

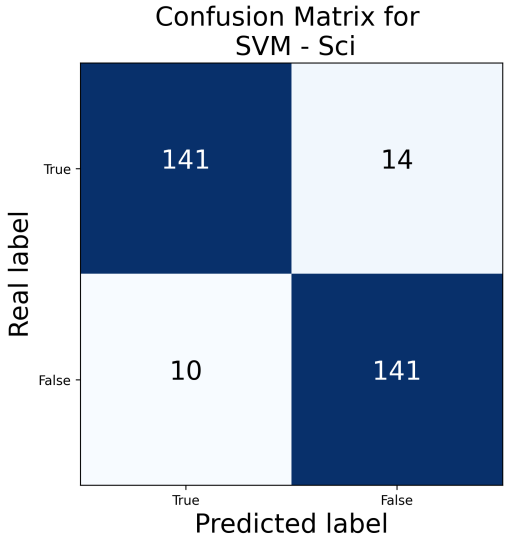

FIGURE 2 – Matrice de confusion du modèle SVM pour la classification des tweets scientifiques et non scientifiques.

D) Tâche 2 : Classification Binaire Multi-Label (CLAIM, REF vs. CONTEXT)

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) pour 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.

D).2 Analyse comparative des performances des différents modèles

Dans cette deuxième tâche, nous avons comparé plusieurs modèles de classification pour résoudre un problème de classification multi-label, où chaque tweet peut être associé à plusieurs catégories (tableau 2) L'objectif était d'évaluer l'efficacité de différentes approches d'apprentissage supervisé adaptées à ce type de problème.

Table 2 – Performances - Tâche 2

Modèle	Accuracy	Précision	Rappel	F1-score
SVM	0.9183 ± 0.0193	0.99	1.00	0.99
Naive Bayes	0.8647 ± 0.0292	0.93	0.93	0.93
k-NN	0.8876 ± 0.0183	0.99	0.99	0.99
Random Forest	0.9359 ± 0.0162	0.99	1.00	0.99
Logistic Regression	0.8961 ± 0.0232	0.98	0.97	0.97
Gradient Boosting	**0.9236 ± 0.0162**	**0.99**	**1.00**	**0.99**

Le modèle Gradient Boosting a montré des performances supérieures dans le cadre de ce problème de classification binaire, atteignant une précision globale de 0.97 et un rappel parfait de 1.00 pour la classe 'claim_or_ref'. L'optimisation des hyperparamètres à l'aide de GridSearchCV a permis d'affiner les paramètres clés, tels que le learning rate (0.1) et la profondeur maximale (5), pour obtenir un score de validation croisée de 0.9236, garantissant une excellente généralisation. En comparaison avec d'autres modèles comme le SVM et le Random Forest, qui ont également montré des performances solides, le Gradient Boosting s'est distingué par sa capacité à mieux gérer les classes déséquilibrées et à produire des résultats plus équilibrés. Bien que le k-NN ait présenté des résultats impressionnants avec des scores de précision et de rappel élevés (0.99), il a montré un léger sur-apprentissage, ce qui peut poser un problème avec de très grands ensembles de données.

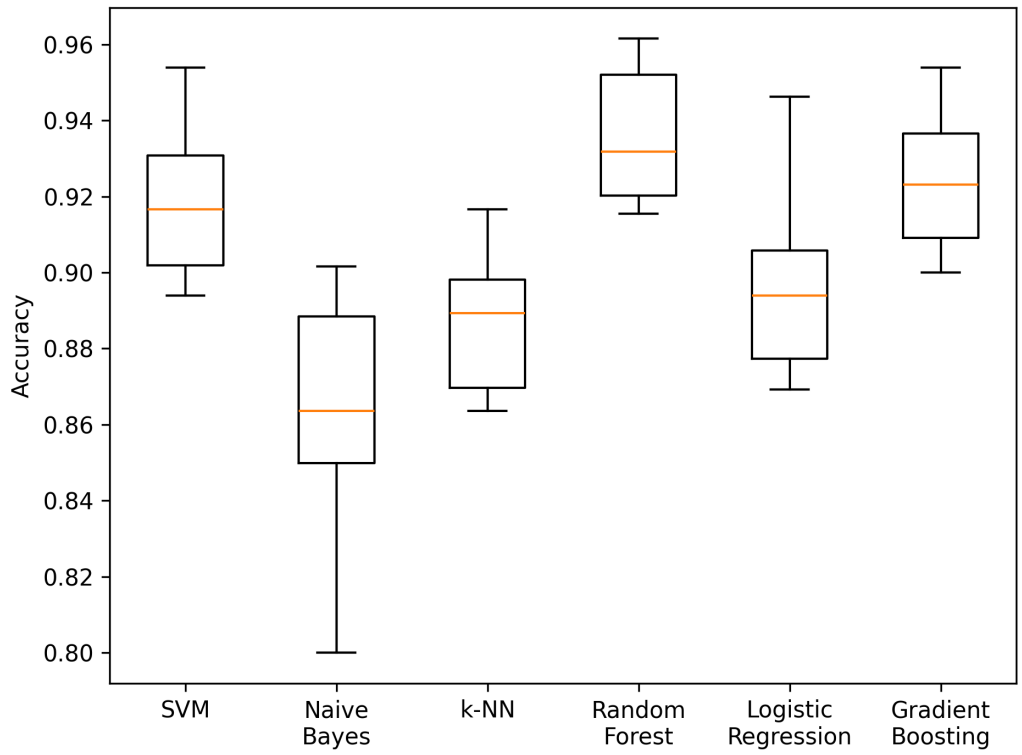

FIGURE 3 – Comparaison des modèles pour la classification des tweets claim et ref vs contexte.

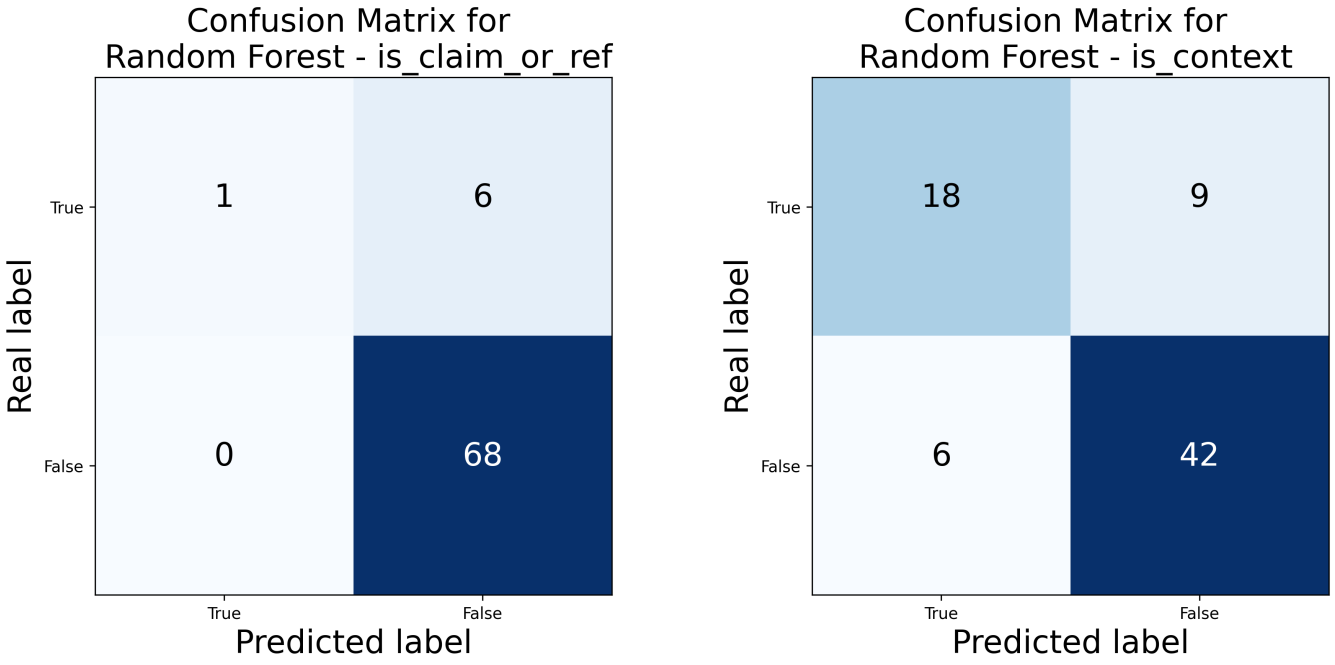

FIGURE 4 – Matrice de confusion du modèle Random Forest pour la classification des tweets claim et ref vs contexte.

E) Tâche 3 :classification multi-classe et multi-label non exclusive (CLAIM vs. REF vs. CONTEXT)

Cette dernière tâche est la plus fine du projet : elle consiste à classer chaque tweet scientifique selon trois catégories : Affirmation, Référence ou Contexte. Bien qu'il s'agisse d'une classification à trois classes, le problème reste de nature multi-label, car un même tweet peut appartenir à plusieurs catégories simultanément.

E).1 Préparation des données

Le prétraitement a suivi une démarche classique, avec une étape supplémentaire pour la classe REF, qui était moins bien prédite.
À cette fin, les URLs ont été remplacées par le token "URL", afin d'en conserver la trace (utile pour reconnaître les références) sans inclure les chaînes brutes. Les stop words ont été nettoyés à l'aide d'une liste enrichie (incluant des termes fréquents sur Twitter comme "rt", "co", "amp", "via", tout en veillant à conserver les négations, susceptibles de porter des nuances importantes. La vectorisation s'est appuyée sur un TF-IDF intégrant des n-grammes allant jusqu'à trois mots, les trigrammes étant particulièrement efficaces pour différencier entre REF (souvent formalisé), CLAIM (langage assertif) et CONTEXT (vocabulaire plus descriptif). Enfin, en raison du déséquilibre marqué entre les catégories (par exemple, très peu de REF), on a implémenté un suréchantillonnage manuel par combinaison de labels, sans détruire la nature multi-étiquette.

E).2 Analyse comparative des performances des différents modèles

Table 3 – Performances - Tâche 3

Modèle	Accuracy	Hamming Loss	F1-score	Rappel
LinearSVC	0.7171 ± 0.0625	0.2844 ± 0.0625	0.8844 ± 0.0625	0.75 ± 0.04
Random Forest	**0.7836 ± 0.0507**	0.2489 ± 0.0507	**0.9133 ± 0.0507**	0.83 ± 0.05
SVM	0.7331 ± 0.0691	0.2622 ± 0.0691	0.8992 ± 0.0691	**0.89 ± 0.07**
Naive Bayes	0.6101 ± 0.0802	0.3156 ± 0.0802	0.8601 ± 0.0802	0.82 ± 0.07
k-NN	0.6868 ± 0.0763	0.36 ± 0.0763	0.8615 ± 0.0763	0.77 ± 0.06
Logistic Regression	0.7433 ± 0.0545	0.2578 ± 0.0545	0.8933 ± 0.0545	0.79 ± 0.06
Gradient Boosting	0.7696 ± 0.0476	**0.2356 ± 0.0476**	0.9086 ± 0.0476	0.85 ± 0.05

Dans cette tâche, tous les modèles de classification présentent une précision et un F1-score relativement bons. On remarque en revanche des valeurs de Hamming Loss plus faibles pour des modèles comme Random Forest, Gradient Boosting et Logistic Regression. Ceci indique une capacité à mieux classer les données tout en minimisant les erreurs globales.
La précision est généralement élevée pour les classes claim et context, mais elle est un peu plus faible pour la classe reference, ce qui peut expliquer pourquoi ces modèles ont des performances variées. Cela montre que les modèles sont bons pour identifier les assertions et les contextes, mais ont plus de difficulté à classer correctement les références. Ceci peut être dû à une classe reference plus difficile à différencier ou moins représentée dans le jeu de données.

Cependant, le F1-score est parfois un peu plus élevé que le score d'accuracy. Les modèles cherchent à optimiser un compromis entre précision et rappel pour chaque classe.

La Hamming Loss plus élevée pour des modèles comme k-NN et Naive Bayes est probablement le reflet de leur difficulté à bien distinguer toutes les classes. Particulièrement pour la classe reference, qui semble poser plus de défis en termes de précision et de rappel. Cela explique aussi pourquoi ces modèles, bien que performants dans certaines classes, peuvent avoir des scores globaux de précision et de rappel moins élevés par rapport à d'autres. En effet, Random Forest ou Gradient Boosting, paraissent mieux gérer la classification dans l'ensemble.

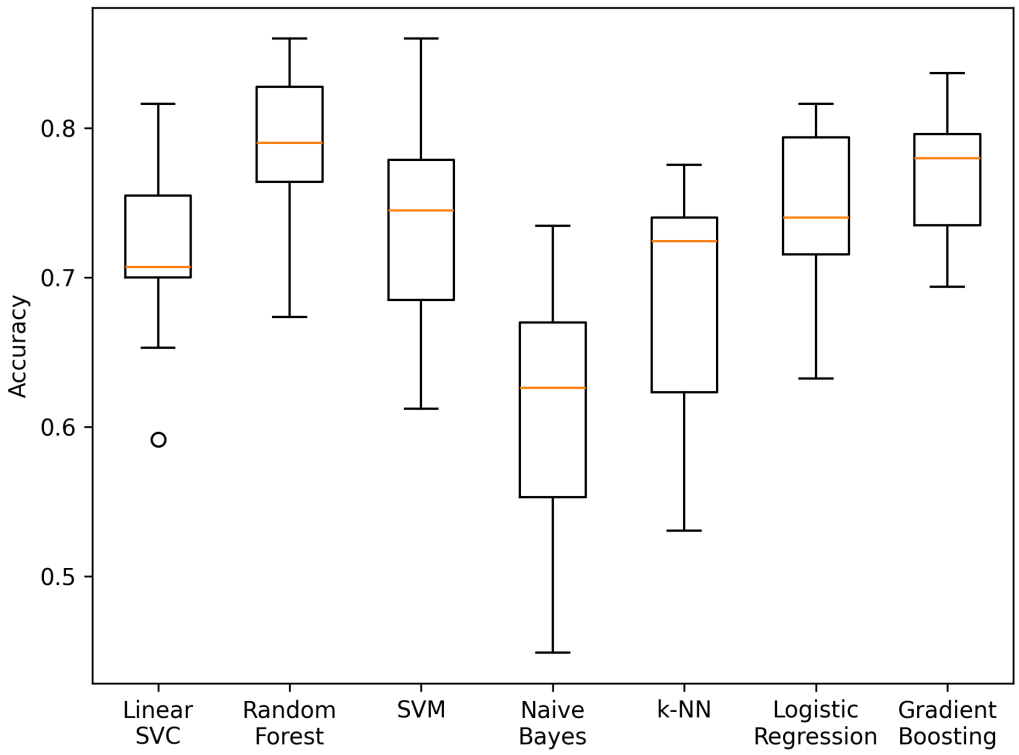

FIGURE 5 – Comparaison des modèles pour la classification des tweets claim vs ref vs contexte.

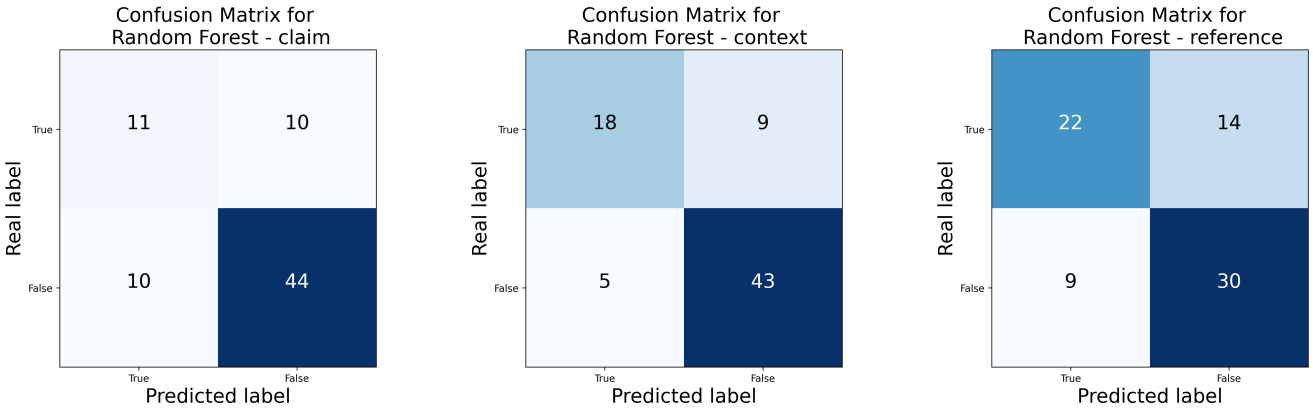

FIGURE 6 – Matrice de confusion du modèle Random Forest pour la classification des tweets claim vs ref vs contexte.

Plusieurs tweets ont été mal classés selon la matrice de confusion car ils appartiennent en réalité à plusieurs classes à la fois, ce que notre système de classification a du mal à détecter simultanément. L'accuracy, étant une métrique très stricte, pénalise fortement ce type d'ambiguïté. Par exemple, un tweet à la fois claim et reference ne sera considéré comme bien classé que s'il est entièrement attribué à la bonne classe unique. Nous présentons dans notre notebook une liste des tweets mal classés avec leur nombre pour illustrer cet impact.

2- Discussion

A) Limites des Modèles

Malgré des performances globalement satisfaisantes, chaque modèle présente certaines limites :

- **Modèles linéaires (régression logistique, SVM) :** sensibles aux features bruitées ou non pertinentes, ils peuvent peiner à capturer des relations complexes dans les données textuelles.
- **Naive Bayes :** basé sur l'indépendance conditionnelle des features, il tend à sursimplifier les relations entre les mots et peut produire des résultats biaisés dans le cas de dépendances lexicales fortes.
- **k-NN :** sensible à la dimensionnalité et aux données bruitées, ce modèle peut être affecté par des mots non informatifs ou rares présents dans les tweets.
- **Random Forest :** bien qu'efficace, il peut sur-apprendre certains patterns lorsque le déséquilibre des classes n'est pas bien traité.

B) Analyse des Features Importantes

Afin de mieux comprendre le comportement des modèles, nous avons extrait les features les plus importantes à l'aide du classifieur Random Forest. Cette démarche nous a permis d'identifier les mots les plus discriminants entre les tweets scientifiques et non-scientifiques. Ces termes ont ensuite été analysés en fonction de leur fréquence d'apparition, représentée sous forme de bar plot (cf Figure X), ce qui a facilité l'interprétation des décisions du modèle et la compréhension des signaux textuels les plus informatifs.

C) Inspection Manuelle des Erreurs

Pour les cas où les performances des modèles étaient jugées insuffisantes, une analyse manuelle des erreurs de classification a été réalisée. Nous avons examiné les tweets mal classés, en particulier les faux positifs et faux négatifs, afin d'identifier des exemples ambigus ou potentiellement mal annotés. Cette inspection a permis de mieux comprendre les limites des modèles face à des contenus complexes ou *borderlines*. Lorsque des motifs d'erreur récurrents ont été repérés, nous avons tenté d'ajuster le prétraitement des données ou d'équilibrer manuellement certaines combinaisons de labels. Enfin, des modifications ciblées du pipeline ont été mises en place pour corriger ces erreurs, tout en veillant à ne pas nuire à la capacité de généralisation des modèles sur des données nouvelles.

3- Conclusion

Cette étude a permis de comparer plusieurs modèles classiques de classification appliqués à des tweets selon leur rapport à la science. Malgré les défis liés au bruit, à l'ambiguïté sémantique et au déséquilibre partiel des classes, les performances obtenues sont encourageantes. Le modèle SVM s'est révélé le plus performant dans la majorité des tâches. Mais plusieurs pistes d'amélioration peuvent être envisagées pour les travaux futurs. Tout d'abord, l'intégration de modèles de langage pré-entraînés comme BERT ou CamemBERT pourrait permettre une meilleure prise en compte du contexte sémantique et améliorer la performance, notamment sur les classes minoritaires ou ambigües. De plus, l'exploration de techniques de data augmentation textuelle ou de rééquilibrage plus avancé pour le multi-label (comme les algorithmes de suréchantillonnage spécifiques au texte) pourrait renforcer la robustesse des modèles face au déséquilibre des combinaisons de labels. Enfin, une analyse plus fine des erreurs de classification, couplée à une révision éventuelle des annotations, permettrait d'améliorer la qualité du jeu de données, ce qui constitue une condition essentielle pour des modèles plus précis et interprétables.