**Prétraitement des données**

Un prétraitement cohérent et rigoureux est une étape essentielle pour améliorer la performance des modèles de classification, en particulier dans un contexte de tweets, caractérisé par une forte variabilité linguistique, une syntaxe bruitée, et un vocabulaire très contextuel. Nous avons adopté une stratégie de prétraitement adaptée à chacune des trois tâches, tout en maintenant une cohérence globale suivant cette ordre de taches :

-Nettoyage textuel et des stop words,

-Normalisation linguistique : Lemmatisation et tokenisation.

-Vectorisation avec TF-IDF et paramétrage des n-grammes.

-Gestion du déséquilibre.

**NOTE** : 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 ({SCI} vs. {NON-SCI}, {CLAIM, REF} vs. {CONTEXT}, {CLAIM} vs. {REF} vs. {CONTEXT}) 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.

### Partie 1 — Classification SCI vs NON-SCI ( classification binaire)

Cette tâche consiste à distinguer les tweets en lien avec la science (SCI) de ceux qui ne le sont pas (NON-SCI). La abse de donnée contient donc des tweets très hétérogènes, ce qui justifie un prétraitement rigoureux mais généraliste.

* Nous avons supprimé les mentions (@), les caractères spéciaux et converti le texte en minuscules. Cela permet de standardiser l’entrée pour éviter que des variations superficielles n’influencent la classification.
* Nous avons utilisé la liste anglaise de NLTK enrichie de mots fréquents sur Twitter ("http", "https", "rt", "co", "amp", "via"), car ces éléments sont omniprésents mais peu informatifs pour déterminer la nature scientifique du contenu.
* **Vectorisation TF-IDF (1-2-grammes)** : un choix raisonnable pour capturer à la fois des mots individuels et des associations fréquentes, tout en limitant la complexité.
* **SMOTE** : utilisé pour équilibrer la distribution entre classes SCI et NON-SCI, car ces dernières sont très déséquilibrées (moins de 10 % de SCI).

Les figures suivantes montrent la distribution des classes (SCI vs NON-SCI), ainsi que la représentation des nuages de mots et les diagrammes en barres sur lesquels nous nous sommes appuyés pour éliminer les stop words et les mentions (@).

📊 **Figures à inclure** :

* Histogramme nbr tweet scientifique et non scientifique
* Nuage de mots et barplot des mots les plus fréquents (pour les @ et cette liste : "https", "rt", "co", "amp", "via"
* Distribution des classes (SCI vs NON-SCI)

**une phrase d’analyse des figures selon le resultats,**

### 🔹 Partie 2 — Classification CLAIM\_OR\_REF vs CONTEXT (binaire, SCI uniquement)

Cette tâche vise à distinguer deux types d’usages des tweets scientifiques : ceux qui formulent ou citent une affirmation scientifique (CLAIM/REF) vs ceux qui servent de contexte (CONTEXT). Les tweets analysés sont uniquement ceux classés comme SCI dans la première tâche.

* Seuls les tweets marqués comme scientifiques (science\_related = 1) ont été conservés.
* **Prétraitement du texte** : très similaire à 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.
* **Vectorisation TF-IDF (1-3-grammes)** : on utilise ici des trigrammes pour mieux capturer des constructions typiques d'énoncés de type affirmation ou citation.
* **Pas de SMOTE classique** : Ici, le déséquilibre est partiel (tableau numeo,,) mais 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.

tableau : Répartition des tweets selon les combinaisons des labels multi-label (CLAIM/REF et CONTEXT)

| is\_claim\_or\_ref | is\_context | Nombre de tweets |
| --- | --- | --- |

|  |  |  |
| --- | --- | --- |
| 1 | 1 | 218 |

|  |  |  |
| --- | --- | --- |
| 1 | 0 | 124 |

|  |  |  |
| --- | --- | --- |
| 0 | 1 | 33 |

Legende : Cette figure met en évidence un déséquilibre partiel dans la distribution des combinaisons de labels multi-label. La combinaison {11} est majoritaire, tandis que la combinaison {01} est fortement sous-représentée. Étant donné la nature multi-label du problème (un tweet pouvant appartenir à plusieurs classes), un rééquilibrage des données s’avère nécessaire pour éviter que le modèle ne soit biaisé en faveur des classes majoritaires. Pour cela, un suréchantillonnage ciblé des combinaisons rares a été mis en place.

📊 **Figures à inclure** :

* nuage des points par class
* barplot des mots les plus fréquents (pour les @ et cette liste : "https", "rt", "co", "amp", "via"
* Distribution des classes (SCI vs NON-SCI)

qlq phrase de l’analyse des figures ou legends ,

### 🔹 Partie 3 — Classification CLAIM vs REF vs CONTEXT (multi-classes, SCI uniquement)

Cette tâche est la plus fine : chaque tweet scientifique est classé comme revendication (CLAIM), référence (REF) ou contexte (CONTEXT). C’est une classification à 3 classes, mais les tweets peuvent avoir plusieurs labels (multi-label).

* Seuls les tweets marqués comme scientifiques (science\_related = 1) ont été conservés.
* Nettoyage classique, mais **avec une étape supplémentaire** : remplacement des URLs par le token "URL", afin de conserver leur présence (utile pour REF) sans garder l’URL brute car la classe REF étais la moinne bonne predite
* **Stop words** : liste standard + mots Twitter ("rt", "co", "amp", "via"); pas de suppression de négation ici, car certaines peuvent indiquer une nuance importante.
* **Vectorisation TF-IDF (1-3-grammes)** : les trigrammes sont particulièrement utiles pour différencier entre REF (souvent formalisé), CLAIM (langage assertif) et CONTEXT (vocabulaire plus descriptif).
* **Suréchantillonnage manuel multi-label** : comme les classes sont très déséquilibrées (par exemple, très peu de REF), on a implémenté un suréchantillonnage par combinaison de labels, sans détruire la nature multi-étiquette.

**Figures à inclure** :

* nuage des points [CLAIM, REF, CONTEXT]
* Histogramme de fréquence des classes
* barplot frequnec des url + "rt", "co", "amp", "via"

soit legend de chque figure ou une anlyde generale des figure ca depond.