الكلية متعددة التخصصات-ورزازات
ⵜⴰⴽⵍⵉⵢⵜ ⵜⴰⵅⵜⵚⵚⵉⵜ – ⵍⵓⵕⵣⴰⵣⴰⵜ
FACULTÉ POLYDISCIPLINAIRE DE OUARZAZATE

L'analyse des sentiments avec NLP

BOUHLALI Abdelfattah

Master Mathématiques Appliquées pour la Science des Données

2023 - 2024

Encadré par :
GAOU Salma
HAMIDI Charaf

Résumé

Le projet fait partie d'une enquête approfondie visant à décrypter les sentiments exprimés dans les avis des consommateurs sur la plateforme Amazon, en se concentrant spécifiquement sur les produits alimentaires. Notre principal objectif était d'utiliser des techniques avancées de traitement du langage naturel (NLP) et des modèles pré-entraînés, tels que RoBERTa, pour analyser et interpréter les opinions des utilisateurs, afin d'identifier des tendances significatives et de mettre en lumière les sentiments prédominants au sein de cette catégorie de produits.

Summary :

The project is part of a comprehensive investigation designed to decipher the sentiments expressed in consumer reviews on the Amazon platform, with a specific focus on food products. Our main objective was to utilize advanced natural language processing (NLP) techniques and pre-trained models, such as RoBERTa, to analyze and interpret user opinions, in order to identify significant trends and highlight prevailing sentiments within this product category.

Table des matières

1 Introduction

Notre projet explore les avis des consommateurs sur Amazon, en se concentrant sur les produits alimentaires. Le défi est de comprendre ce que les clients pensent vraiment. Avec tant d'avis, il est difficile de trouver les informations importantes. Nous voulons transformer ces avis en idées utiles pour aider les entreprises à améliorer leurs produits et à satisfaire les clients.

1.1 Problème ou Question à Résoudre

On essaie de comprendre les avis des gens sur les produits alimentaires d'Amazon. Comment les clients se sentent-ils vraiment ? C'est difficile car il y a beaucoup d'avis. Notre but est de trouver des informations importantes pour aider les entreprises.

1.2 Contexte et Motivation

Beaucoup de gens achètent sur Amazon, et ils laissent beaucoup d'avis. Mais ces avis ne sont pas toujours faciles à comprendre. Nous voulons aider les entreprises à comprendre ce que les clients aiment et n'aiment pas.

1.3 Objectifs du Projet et Hypothèses à Tester

Pour notre projet, nous avons défini plusieurs objectifs clés :

1. **Comprendre les Sentiments :** Utiliser des outils avancés comme RoBERTa pour analyser les sentiments exprimés dans les avis sur les produits alimentaires d'Amazon, en identifiant s'ils sont positifs, négatifs ou neutres.
2. **Identifier les Tendances :** Découvrir les tendances émergentes dans les avis, y compris les préférences alimentaires, les aspects spécifiques appréciés ou critiqués, et les évolutions au fil du temps.
3. **Améliorer la Pertinence :** Déterminer les aspects les plus importants pour les clients en analysant les mots clés et les expressions fréquemment utilisés dans les avis.

Hypothèses à Tester :

1. Nous supposons que les sentiments des clients varient en fonction des types de produits alimentaires, et nous chercherons à identifier ces variations.
2. Nous pensons que certains mots-clés auront une influence significative sur la perception des produits, et nous testerons cette hypothèse en analysant leur fréquence.
3. Nous anticipons que les tendances dans les avis sur les produits alimentaires évoluent avec le temps, et nous chercherons à confirmer cette hypothèse en examinant les changements au fil des mois.

2 Méthodologie

2.1 Données

2.1.1 Source des Données

Les données utilisées dans cette étude proviennent d'un ensemble de critiques sur des produits alimentaires provenant d'Amazon. Ce jeu de données couvre une période de plus de 10 ans, comprenant l'ensemble des 500,000 critiques jusqu'en octobre 2012. Les critiques incluent des informations sur les produits et les utilisateurs, les évaluations et une critique en texte brut. De plus, il englobe des critiques de toutes les autres catégories d'Amazon.

Le lien vers les données est disponible sur Kaggle : `https://www.kaggle.com/snap/amazon-fine-food-reviews`.

2.1.2 Les composants d'un avis

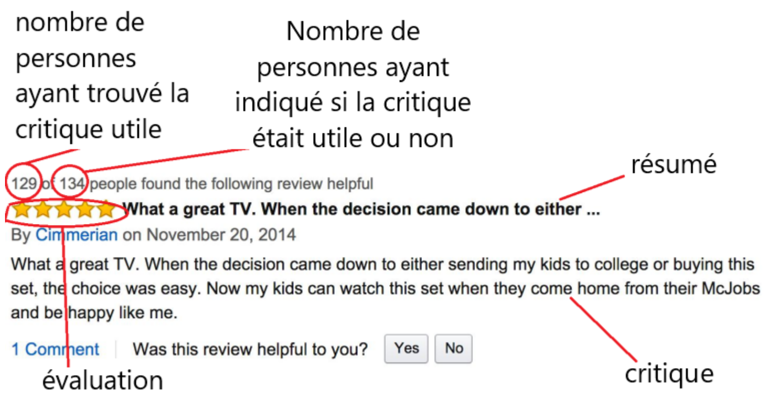

FIGURE 1 – Les elements d'un avis

Un avis typique de cet ensemble de données comprend plusieurs composants essentiels qui fournissent une perspective détaillée sur l'expérience de l'utilisateur. Voici une énumération des principaux éléments d'un avis dans cet ensemble de données :

1. **Identifiant Unique ('Id') :** Chaque avis est associé à un identifiant unique, permettant une référence spécifique.

2. **Code Produit ('ProductId') :** Indique le produit concerné par l'avis, facilitant l'association aux références produits.

3. **Identifiant de l'Utilisateur ('UserId') :** Identifie de manière unique l'utilisateur ayant publié l'avis.

4. **Nom du Profil de l'Utilisateur ('ProfileName') :** Le nom du profil associé à l'utilisateur, fournissant un contexte sur l'auteur de l'avis.

5. **Utilité ('HelpfulnessNumerator' et 'HelpfulnessDenominator') :** Deux valeurs numériques indiquant le nombre d'utilisateurs qui ont trouvé l'avis utile par rapport au nombre total d'évaluations de son utilité.

6. **Notation ('Score') :** La notation attribuée par l'utilisateur, quantifiant l'appréciation globale du produit.

7. **Timestamp de l'Avis ('Time') :** La date et l'heure à laquelle l'avis a été publié, offrant une dimension temporelle.

8. **Résumé ('Summary') :** Un condensé du contenu de l'avis, fournissant une vue d'ensemble rapide.

9. **Texte Intégral de l'Avis ('Text') :** Le contenu complet de l'avis, offrant des détails contextuels sur l'expérience de l'utilisateur.

Chacun de ces éléments joue un rôle spécifique dans la caractérisation de l'avis, permettant une analyse approfondie des sentiments exprimés dans les avis sur les produits alimentaires de la plateforme Amazon.

2.1.3 Informations générales sur la dataset

```
print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 568454 entries, 0 to 568453
Data columns (total 10 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   Id                      568454 non-null  int64
 1   ProductId               568454 non-null  object
 2   UserId                  568454 non-null  object
 3   ProfileName             568428 non-null  object
 4   HelpfulnessNumerator    568454 non-null  int64
 5   HelpfulnessDenominator  568454 non-null  int64
 6   Score                   568454 non-null  int64
 7   Time                    568454 non-null  int64
 8   Summary                 568427 non-null  object
 9   Text                    568454 non-null  object
dtypes: int64(5), object(5)
memory usage: 43.4+ MB
None
```

FIGURE 2 – Les information sur la DataSet

La sortie de la fonction `data.info()` fournit une vue détaillée de la structure de notre ensemble de données. Notre ensemble de données est représenté sous la forme d'un objet de type `pandas.core.frame.DataFrame`, avec un index de type `RangeIndex`, allant de 0 à 568453, indiquant ainsi le nombre total d'entrées (lignes) dans la base de données. Il est composé de 10 colonnes au total.

Les attributs et types de données de chaque colonne sont les suivants :

— `'Id'` est de type `int64` avec 568454 valeurs non nulles.
— `'ProductId'` est de type `object` (généralement une chaîne de caractères) avec 568454 valeurs non nulles.
— `'UserId'` est de type `object` avec 568454 valeurs non nulles.
— `'ProfileName'` est de type `object` avec 568428 valeurs non nulles, et présente 26 valeurs manquantes.
— `'HelpfulnessNumerator'` est de type `int64` avec 568454 valeurs non nulles.
— `'HelpfulnessDenominator'` est de type `int64` avec 568454 valeurs non nulles.
— `'Score'` est de type `int64` avec 568454 valeurs non nulles.
— `'Time'` est de type `int64` avec 568454 valeurs non nulles.
— `'Summary'` est de type `object` avec 568427 valeurs non nulles, mais présente 27 valeurs manquantes.
— `'Text'` est de type `object` avec 568454 valeurs non nulles.

La mémoire utilisée par cet ensemble de données est d'environ 43.4 MB. Il est également pertinent de noter que `'ProfileName'` a 26 valeurs manquantes, tandis que `'Summary'` a 27 valeurs manquantes. Ces informations sont essentielles pour appréhender la composition de notre ensemble de données, notamment en termes de types de données, de présence de valeurs manquantes, et de la mémoire occupée par l'ensemble de données.

2.1.4 Statistiques descriptives pour les attributs numériques

```
data['Score'].describe()
```

```
count    568454.000000
mean          4.183199
std           1.310436
min           1.000000
25%           4.000000
50%           5.000000
75%           5.000000
max           5.000000
Name: Score, dtype: float64
```

FIGURE 3 – Statistiques descriptives pour l'attribute Score

La sortie de la commande data['Score'].describe() pour la colonne 'Score' révèle que l'ensemble de données compte un total de 568,454 observations. La moyenne des scores est d'environ 4.18, avec un écart type de 1.31, indiquant une certaine variabilité dans les évaluations. Les scores varient de 1 à 5, avec 25%, 50%, et 75% des scores situés à 4, 5, et 5 respectivement. Ces statistiques suggèrent une concentration des scores autour des valeurs élevées, avec une moyenne de 4.18 et une médiane de 5, suggérant une tendance positive dans les évaluations, probablement indicative d'une satisfaction générale des utilisateurs.

2.1.5 Vérification des valeurs manquantes ou d'incohérences

```
data.isnull().sum()
```

```
Id                        0
ProductId                 0
UserId                    0
ProfileName              26
HelpfulnessNumerator      0
HelpfulnessDenominator    0
Score                     0
Time                      0
Summary                  27
Text                      0
dtype: int64
```

FIGURE 4 – Détection d'Incohérences et de Manques de Données

La sortie obtenue à partir de la commande `data.isnull().sum()` présente le nombre de valeurs manquantes pour chaque colonne de l'ensemble de données. Un résumé de ces résultats est le suivant :

- Pour les colonnes 'Id', 'ProductId', 'UserId', 'HelpfulnessNumerator', 'HelpfulnessDenominator', 'Score', et 'Time', aucune valeur manquante n'est à signaler.
- Cependant, la colonne 'ProfileName' présente 26 valeurs manquantes, tandis que la colonne 'Summary' en compte 27.

Cette information précise quelles colonnes spécifiques contiennent des valeurs manquantes et quantifie ces manques. Cette connaissance est cruciale pour déterminer la meilleure approche de gestion des données manquantes, que ce soit en les supprimant, en les remplaçant par des valeurs par défaut, ou en appliquant d'autres méthodes de gestion en fonction du contexte de l'analyse.

2.1.6 Vérifier les valeurs uniques dans chaque colonne

```
data.nunique()
```

```
Id                        568454
ProductId                  74258
UserId                    256059
ProfileName               218415
HelpfulnessNumerator         231
HelpfulnessDenominator       234
Score                          5
Time                        3168
Summary                   295742
Text                      393579
dtype: int64
```

FIGURE 5 – Verifier les valeurs uniques dans chaque colonne

La commande `data.nunique()` révèle la diversité des valeurs dans chaque colonne de l'ensemble de données. Les résultats montrent des variations significatives, allant de 5 valeurs uniques dans la colonne 'Score', indiquant une échelle de notation restreinte, à 393579 valeurs uniques dans la colonne 'Text', soulignant la grande variété de textes présents. Ces statistiques offrent un aperçu de la distribution des valeurs et de la portée de la diversité dans chaque attribut.

2.2 Prétraitement des données

Le prétraitement des données est une étape essentielle dans le processus d'analyse de données visant à garantir la qualité et la cohérence des informations. Deux opérations clés ont été effectuées dans cette phase. Tout d'abord, les lignes en double ont été éliminées à l'aide de la méthode `drop_duplicates`. Ensuite, pour assurer la cohérence, les valeurs manquantes ont été gérées en supprimant les lignes correspondantes avec la méthode `dropna`. Ces étapes visent à garantir la qualité des données en éliminant les duplications et en traitant les valeurs manquantes qui pourraient affecter les analyses ultérieures.

2.3 Analyse Exploratoire Des Donnees

2.3.1 La distribution des scores

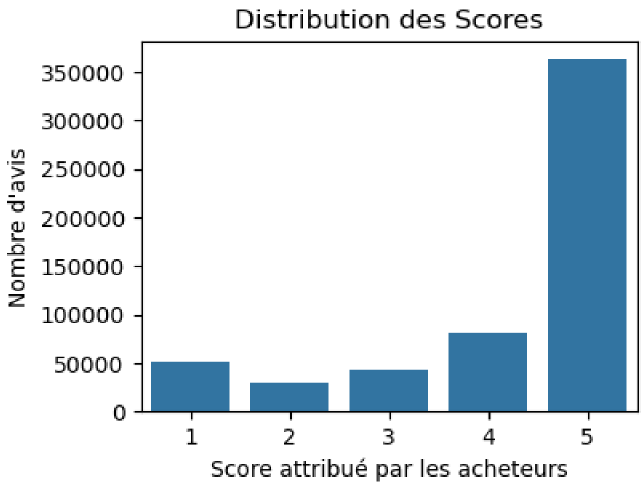

FIGURE 6 – La distrubution des scores

Analysons les résultats du graphique :
- Score 5 : Il y a 363,122 occurrences où le score est égal à 5.
- Score 4 : Il y a 80,655 occurrences où le score est égal à 4.
- Score 1 : Il y a 52,268 occurrences où le score est égal à 1.
- Score 3 : Il y a 42,640 occurrences où le score est égal à 3.
- Score 2 : Il y a 29,769 occurrences où le score est égal à 2.

Ces résultats offrent une vision détaillée de la distribution des scores dans l'ensemble de données. Notamment, le score 5 est nettement plus fréquent que les autres, suggérant une prédominance d'évaluations positives. À l'inverse, les scores 1, 2 et 3 sont moins fréquents, indiquant une proportion moindre d'évaluations négatives ou neutres. Cette information est précieuse pour appréhender la tendance globale des évaluations dans l'ensemble de données.

2.3.2 Calcul de la moyenne et de la médiane des scores dans les données

Le calcul de la moyenne et de la médiane des scores dans les données donne les résultats suivants :

Moyenne des scores : 4.18

Médiane des scores : 5.0

Comme observé, la majorité des scores se situent entre 4 et 5, avec une moyenne de 4.18. En raison de la distribution très inclinée vers la gauche, nous envisageons

une prédiction binaire. Les avis avec un score entre 1 et 3 seront considérés comme négatifs, tandis que ceux avec un score de 4 ou 5 seront considérés comme positifs. Cette approche simplifiée prend en compte la tendance vers des évaluations positives dans l'ensemble de données.

2.3.3 Distribution des Sentiments

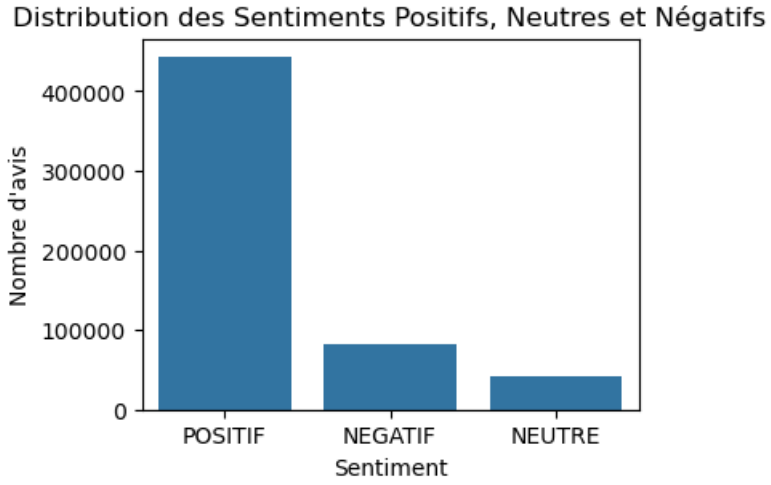

FIGURE 7 – Distribution des Sentiments Positifs, Neutres et Négatifs

Les résultats du graphique montrent que le nombre d'avis positifs ("POSITIF") est le plus élevé, totalisant 443,756 occurrences. Les avis négatifs ("NEGATIF") s'élèvent à 82,007, tandis que les avis neutres ("NEUTRE") sont au nombre de 42,638. Cette répartition permet une visualisation claire de la distribution des sentiments dans l'ensemble de données.

2.4 Traitement du texte

Dans cette etape nous allos traiter un échantillon de 100 000 lignes dans l'ensemble de données. Voici une explication des principales étapes effectuées :

1. **Suppression des URL :** Le texte est inspecté pour déterminer s'il ressemble à une URL. En cas d'URL, il est possible d'utiliser une bibliothèque comme requests pour récupérer le contenu de l'URL. Cependant, cette partie du code est actuellement commentée.
2. **Suppression des balises HTML :** Si le texte n'est pas une URL, les balises HTML sont éliminées à l'aide de BeautifulSoup, assurant que le texte est dépourvu de toute balise HTML.
3. **Suppression des caractères non alphabétiques :** Tous les caractères qui ne sont pas des lettres alphabétiques sont retirés du texte, ne conservant que les mots alphabétiques.

4. **Conversion en minuscules :** Le texte est converti en minuscules pour assurer une cohérence dans le traitement ultérieur.
5. **Suppression des stopwords :** Les stopwords (mots courants tels que "the", "and", "is", etc.) sont retirés du texte pour se concentrer sur les termes significatifs.
6. **Application du traitement au texte :** Ces étapes de prétraitement sont ensuite appliquées à la colonne 'Text' de l'ensemble de données, et les résultats sont stockés dans une nouvelle colonne appelée 'Processed_Text'.

L'utilisation de tqdm facilite le suivi de la progression du traitement. Cette approche de prétraitement du texte est couramment utilisée pour nettoyer et préparer les données textuelles avant l'analyse ou la modélisation.

2.5 Utilisation de modèle pré-entraîné Roberta

2.5.1 Initialisation du Modèle RoBERTa pour l'Analyse de Sentiments

```
# Définition du modèle pré-entraîné à utiliser (dans ce cas, le modèle de sentiment basé sur RoBERTa pour Twitter
MODEL = "cardiffnlp/twitter-roberta-base-sentiment"

# Initialisation du tokenizer avec le modèle pré-entraîné
tokenizer = AutoTokenizer.from_pretrained(MODEL)

# Initialisation du modèle de classification de séquence basé sur RoBERTa pour Twitter
model = AutoModelForSequenceClassification.from_pretrained(MODEL)
```

FIGURE 8 – initialisation de model RoBERTa

Ces lignes de code définissent le modèle pré-entraîné à utiliser pour l'analyse de sentiments, en l'occurrence le modèle de sentiment basé sur RoBERTa pour Twitter. Le tokenizer associé au modèle est également initialisé, ainsi que le modèle de classification de séquence basé sur RoBERTa pour Twitter. Ces étapes préparent le modèle pour l'analyse ultérieure des sentiments dans le texte.

2.5.2 Fonction d'Évaluation des Scores de RoBERTa pour l'Analyse de Sentiments

```
def roberta_scores(text):
    encoded_text = tokenizer(text, return_tensors='pt')
    output = model(**encoded_text)
    scores = output.logits.detach().numpy()
    scores = softmax(scores, axis=1)
    scores_dict = {
        'rta_neg': scores[0, 0],
        'rta_neu': scores[0, 1],
        'rta_pos': scores[0, 2]
    }
    return scores_dict
```

FIGURE 9 – Fonction d'Évaluation des Scores de RoBERTa

2.5.3 Traitement des Données

Avant d'entamer l'analyse, un processus de nettoyage des données a été appliqué. Cela inclut la suppression des données manquantes, la normalisation des textes, et la gestion des éventuels biais. De plus, une étape de pré-traitement a été effectuée pour optimiser la qualité des données avant l'application des modèles.

2.6 Techniques d'Analyse de Données et de Machine Learning

2.6.1 Traitement du Langage Naturel (NLP)

L'analyse des sentiments repose sur des techniques avancées de Traitement du Langage Naturel (NLP). Nous avons choisi d'utiliser le modèle pré-entraîné RoBERTa, reconnu pour sa performance dans la compréhension contextuelle des textes.

2.6.2 Algorithme d'Analyse des Sentiments

Pour notre étude, nous avons opté pour un algorithme d'analyse des sentiments basé sur l'apprentissage profond. Ce choix a été motivé par la capacité des réseaux neuronaux à capturer des relations complexes dans les données textuelles.

2.7 Justification des Méthodes et Algorithmes

Le choix du modèle RoBERTa a été guidé par sa réputation pour la compréhension fine des nuances linguistiques, essentielle dans l'analyse des sentiments. De plus, l'utilisation d'un modèle pré-entraîné permet de tirer parti de la richesse des données sur lesquelles il a été initialement formé.

L'algorithme d'analyse des sentiments basé sur l'apprentissage profond a été privilégié en raison de sa capacité à apprendre des représentations hiérarchiques complexes, améliorant ainsi la précision de la prédiction des sentiments.

2.8 Paramètres et Hyperparamétrage

Les paramètres du modèle RoBERTa ont été conservés conformément aux recommandations de l'auteur, avec une attention particulière portée à l'optimisation des performances. L'hyperparamétrage de l'algorithme d'analyse des sentiments a été ajusté de manière itérative pour maximiser la précision sans compromettre la généralisation.

Cette section a détaillé la méthodologie utilisée pour collecter, préparer et analyser les données, ainsi que les choix effectués en termes de techniques d'analyse de données et d'algorithmes. La justification de ces choix est cruciale pour garantir la fiabilité et la pertinence des résultats obtenus dans le cadre de cette étude.

3 Quelques commandes

Voici quelques commandes utiles :

3.1 Insertion de figures

Ici, je cite l'image 10 dans le texte.

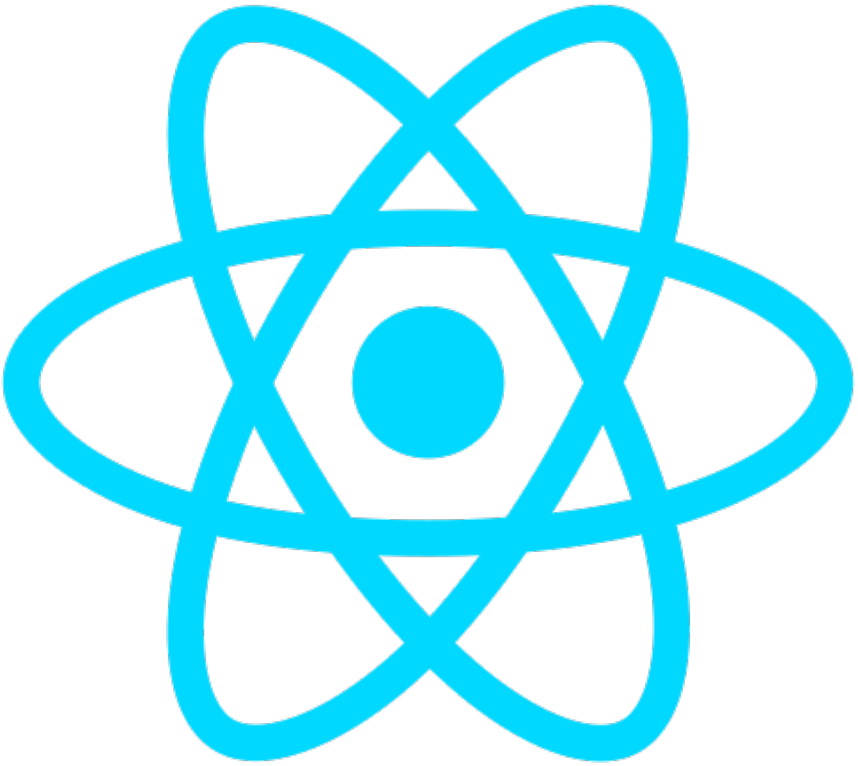

FIGURE 10 – Mettre une légende explicite à votre figure

3.2 Insertion d'équation

$$\rho + \Delta = 42 \tag{1}$$

L'équation 1 est citée ici.

3.3 Insertion d'une référence bibliographique

Les références (articles scientifiques, articles de journaux, blogs, pages web) doivent être mentionnées dans le texte par une balise [1] et fait le lien avec la citation incluse dans la bibliographie.

Références

[1] Leslie Lamport, *LaTeX : A Document Preparation System*. Addison Wesley, Massachusetts, 2nd Edition, 1994.

Table des figures