# RAPPORT D'EXPLORATION ET DE PRE-PROCESSING – Projet Trustpilot

Auteur : Pierre Poulouin, Almira Rodrigues

Formation : Data Scientist – Datascientest

Projet : Analyse d'avis clients – Trustpilot (RingConn, OuraRing, Circular)

## 1. Introduction au projet

Le projet s'inscrit dans le cadre de mon alternance au Puy du Fou et vise à analyser des avis clients provenant de Trustpilot afin d'évaluer automatiquement la satisfaction client.

### Contexte

Contexte métier : Le projet illustre un cas concret d'application du Machine Learning dans l'analyse d'avis clients. Il s'insère dans le contexte d'un métier de Data Scientist, où l'objectif est d'extraire de la valeur des textes libres pour mieux comprendre la perception des utilisateurs.

Contexte technique : Le projet repose sur un pipeline complet de Machine Learning : scraping des données, nettoyage et fusion, vectorisation TF-IDF, puis modélisation. L'environnement de travail est Python, avec les bibliothèques pandas, nltk, scikit-learn et matplotlib.

Contexte économique : L'analyse automatique des avis clients permet d'identifier rapidement les axes d'amélioration produits ou services sans lecture manuelle. Cela représente un gain de temps et un outil d'aide à la décision marketing.

Contexte scientifique : Le projet se situe dans le domaine du NLP (Natural Language Processing) et de la classification supervisée. Il applique des méthodes statistiques et vectorielles éprouvées pour transformer le texte en représentation numérique exploitable.

### Objectifs

L'objectif principal est de construire un modèle capable de prédire la note associée à un avis client à partir de son texte, puis d'en déduire un sentiment global positif ou négatif.

Objectifs secondaires : nettoyage, préparation, vectorisation TF-IDF, construction d'un pipeline reproductible et évaluation via F1, ROC-AUC.

Niveau d'expertise : Je travaille seul sur le projet, avec un niveau intermédiaire en Machine Learning et une volonté d'approfondir les techniques de NLP.

Interactions métier : Des échanges avec des professionnels du marketing ont confirmé l'intérêt du passage en binaire ('satisfait ou non') et la valeur métier de ce modèle.

## 2. Compréhension et manipulation des données

Les données proviennent du scraping d'avis publics sur Trustpilot pour trois marques : RingConn, OuraRing et Circular.

Elles sont publiques, issues des pages officielles Trustpilot.

Fichier final : trustpilot_dataset_final_cleaned.csv. Volumétrie : environ 1000 à 1500 avis.

Variables clés : Title, Content, Rating, Author, Date, Country, ReviewsCount.

Variable cible : Rating (1 à 5). Variable explicative principale : CleanText (texte fusionné et nettoyé).

Limitations : taille modeste du dataset, ambiguïtés linguistiques, avis parfois ironiques ou neutres.

## 3. Pre-processing et Feature Engineering

Le prétraitement a consisté à normaliser et nettoyer les textes :

1. Fusion des colonnes Title et Content.

2. Suppression des caractères spéciaux, emojis et HTML.

3. Passage en minuscules et suppression des stopwords.

4. Lemmatisation via nltk.WordNetLemmatizer.

5. Création de la colonne CleanText finale.

Ces étapes garantissent une représentation homogène du langage et évitent la redondance ('liked', 'likes', 'liking' → 'like').

Feature engineering prévu : ajout futur de variables simples comme la longueur, le nombre d'exclamations, le ratio de majuscules et la subjectivité.

Le texte est vectorisé via TF-IDF, qui normalise déjà les fréquences des mots : aucune standardisation supplémentaire n'est nécessaire.

## 4. Visualisations et Statistiques

Analyse exploratoire :

- Répartition des notes : les notes 4 et 5 sont majoritaires (~70 %), les notes 1 et 2 représentent ~20 %.

- Longueur des avis : les avis négatifs sont plus courts, les positifs plus détaillés.

- Nuages de mots : les mots fréquents pour les avis 5 sont ‘amazing’, ‘perfect’, ‘recommend’; pour les avis 1 : ‘disappointed’, ‘worst’, ‘refund’.

- Corrélation observée entre la polarité du vocabulaire et la note.

Ces observations confirment que le texte reflète bien la satisfaction client et justifient l’approche de classification supervisée.

## 5. Conclusion du rapport d’exploration

L’exploration et le prétraitement ont permis d’obtenir un dataset propre, structuré et prêt pour la modélisation.

Les principales difficultés concernaient la variabilité linguistique et les doublons entre marques, corrigés par un nettoyage rigoureux.

Cette phase a confirmé que le texte seul suffisait à expliquer la note, ouvrant la voie à une modélisation fiable et interprétable.