Projet débuté en Jan. 2023

Classification automatique de biens de consommation

-

Entreprise Place de Marché

Open Classrooms parcours Data Science - projet 6

Camille Besançon

PROBLÉMATIQUE

- Entreprise “Place de Marché” : Souhaite lancer une plateforme d’e-commerce entre particuliers.
- Plusieurs problèmes :
  - Peu d’articles (plateforme encore récente)
  - Catégorie des articles définie par les vendeurs : Peu fiable
  - **L’expérience utilisateur (acheteurs / vendeurs) est pénalisée**

Une solution proposée serait d’automatiser la classification des articles en différentes catégories

OBJECTIF ET APPROCHE

Objectif : Présenter une étude de faisabilité d'un moteur de classification se basant sur des informations textuelles ou sur des images

- Classification basée sur la description des produits
  - Extraction des features selon plusieurs approches (Bag of Words, 'sentence embedding')
- Classification basée sur les images des produits
  - Extraction de features avec la méthode SIFT et par un algorithme de type CNN Transfer learning
- Après l'extraction de features : Etapes de réduction de dimension, clustering et évaluation de la qualité de la classification (Score ARI : Adjusted Rand Index - Visualisation : t-SNE)

OBJECTIF ET APPROCHE : ARI

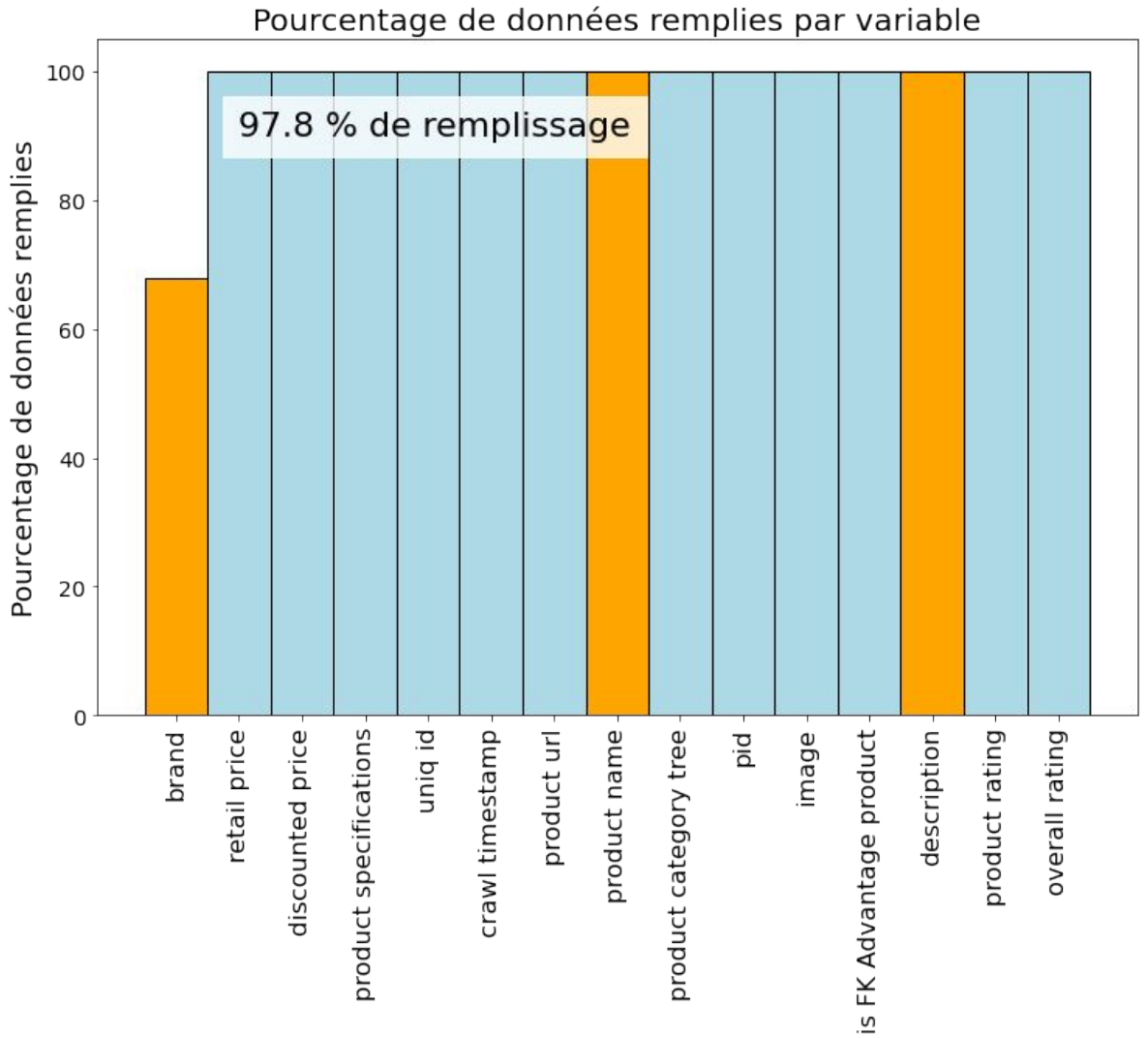
ARI : Indice de Rand ajusté, défini par la proportion de paires d'observations qui sont regroupés de la même façon dans les deux répartitions (réelle et prédite)

Exemple : Deux points qui sont dans un même cluster ou deux points qui sont dans des clusters différents sont considérés comme "groupés de la même façon"

- ARI = 1 signifie que tous les points sont classés de la même façon. Résultat "parfait"
- ARI = 0 signifie que les points sont classés de façon aléatoire
- Cette métrique nous permettra d'évaluer la qualité de nos classifications.

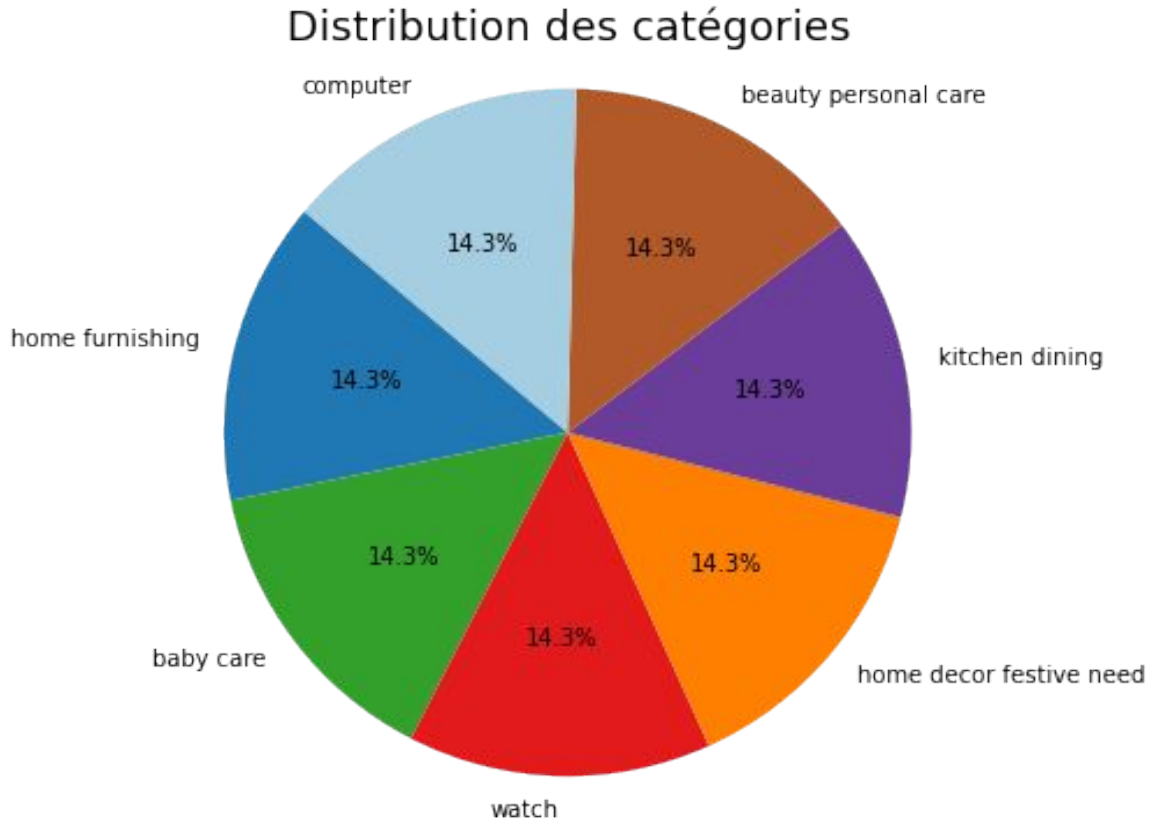
DONNÉES FOURNIES

- Données texte et images sur 1050 produits
- Données texte : Peu de données manquantes
  - Marques des produits incomplètes : pas utilisables en l'état (axe à travailler ?)
  - Description et nom des produits complets
- Chaque produit a une image associée



DONNÉES FOURNIES

- Données texte et images sur 1050 produits
- Données texte : Peu de données manquantes
  - Marques des produits incomplètes : pas utilisables en l'état (axe à travailler ?)
  - Description et nom des produits complets
- Chaque produit a une image associée
- 7 Catégories : 150 produits / catégories



TRAITEMENT DES DONNÉES TEXTE

- Textes : Beaucoup de mots peu significatifs (pronoms, conjonction de coordination, …), ponctuation, caractères spéciaux (internet : emojis)
- Des étapes de nettoyage sont nécessaires.
  - Passage en minuscule
  - Suppression des “mots” composés de chiffres
  - ‘Tokenization’ : Séparation des phrases en tokens individuels (mots)
    - Filtre des Stopwords : Liste de mots ayant peu de signification (exemple : by, when, how, to, for, and, of, …)
  - ‘Lemmatization’ : Transformation des mots / tokens en leur racine
  - Filtrage des mots peu significatifs :
    - Dans notre cas les noms et verbes sont les plus importants. Adverbes et adjectifs supprimés
- Données à haute dimensions : Basées sur le vocabulaire. Nécessité de réduire ces dimensions / sélectionner les plus pertinentes

TRAITEMENT DES DONNÉES TEXTE

- Exemple de traitement du texte :

'Key Features of Lock&Lock Kitchen - 5.5 L Polypropylene Multi-purpose Storage Container Airtight Pack of 6,Lock&Lock Kitchen - 5.5 L Polypropylene Multi-purpose Storage Container (Pack of 6, Clear) Price: Rs. 2,145 Flexible Silicone Seal Ensures Precision Lid Fitting.Large Locking Hinges For Easy Open And Closing Movement.Fridge And Freezer Safe,Specifications of Lock&Lock Kitchen - 5.5 L Polypropylene Multi-purpose Storage Container (Pack of 6, Clear) General Brand Lock&Lock Model Number HPL816X3,HPL806X2,HPL836 Disposable No Model Name Kitchen Material Polypropylene Airtight Yes Capacity 5.5 L Container Type Multi-purpose Storage Container Color Clear In the box Sales Package 6 CONTAINER Pack of 6 Warranty Covered in Warranty Manufacturing Defects Warranty Summary 1 year warranty on manufacturing defects. Warranty Service Type Customer Care Not Covered in Warranty Accidental Damages'

'feature, lock, lock, kitchen, polypropylene, multi, storage, container, pack, lock, lock, kitchen, polypropylene, multi, storage, container, pack, price, silicone, seal, precision, hinge, closing, movement, fridge, freezer, specification, lock, lock, kitchen, polypropylene, multi, storage, container, pack, brand, lock, lock, model, number, hpl806x2, hpl836, model, name, material, polypropylene, capacity, container, type, multi, storage, container, color, box, sale, package, container, pack, warranty, manufacturing, warranty, year, warranty, manufacturing, warranty, service, type, customer, care, damage'

TRAITEMENT DES DONNÉES TEXTE

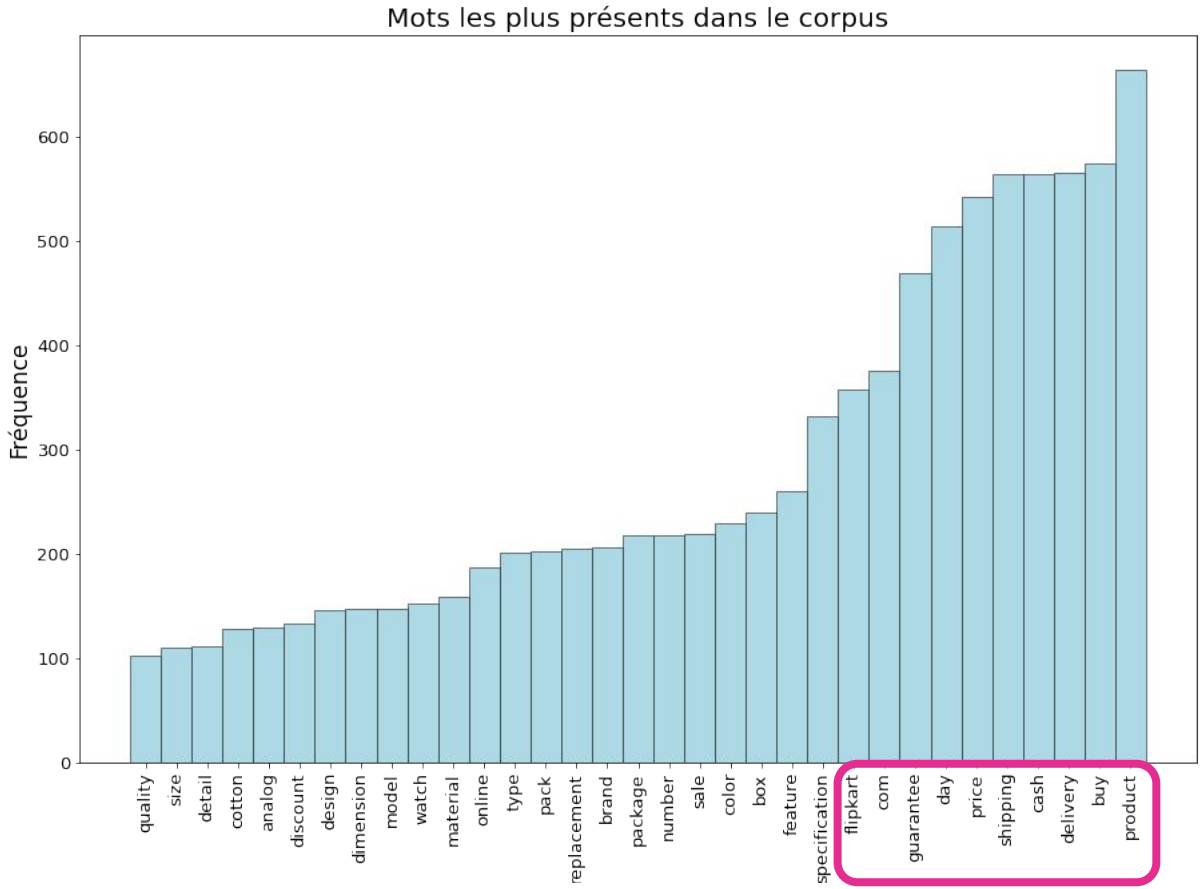
- Exemple de traitement du texte : Certains mots apparaissent plusieurs fois ou sont uniques dans le corpus

```
'feature, lock, lock, kitchen, polypropylene, multi, storage, container, pack, lock, lock, kitchen, polypropylene,
multi, storage, container, pack, price, silicone, seal, precision, hinge, closing, movement, fridge, freezer, spec
ification, lock, lock, kitchen, polypropylene, multi, storage, container, pack, brand, lock, lock, model, number,
hpl806x2, hpl836, model, name, material, polypropylene, capacity, container, type, multi, storage, container, colo
r, box, sale, package, container, pack, warranty, manufacturing, warranty, year, warranty, manufacturing, warrant
y, service, type, customer, care, damage'
```

Certains de ces mots peuvent être significatifs. Mais pas tous : Mots uniques ou trop présents peu discriminants.

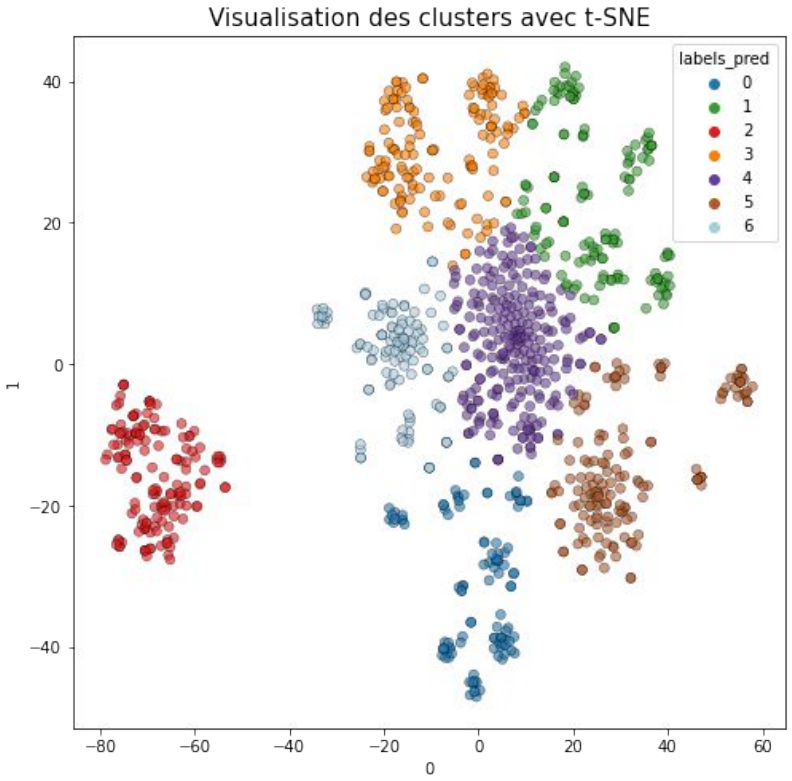
TRAITEMENT DES DONNÉES TEXTE

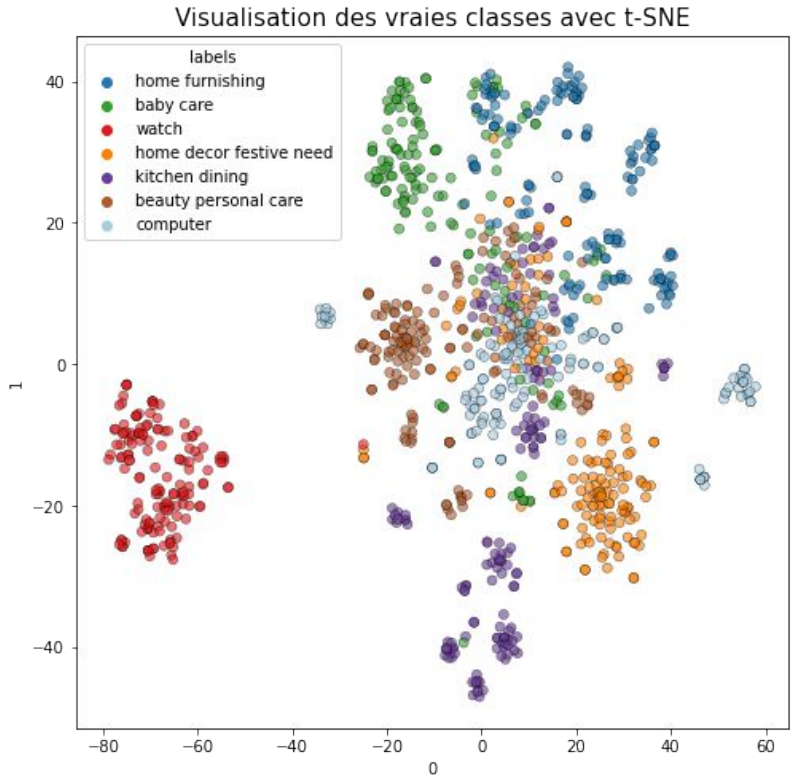
- Exemple de traitement du texte : Certains mots apparaissent plusieurs fois ou sont uniques dans le corpus



ANALYSE PAR COMPTAGE SIMPLE (CountVectorizer)

- Construction du vocabulaire à partir du corpus de texte puis, pour chaque description, comptage du nombre d'occurrence de chaque mot
- Réduction de dimension : PCA, t-SNE, clustering, mesure de l'ARI

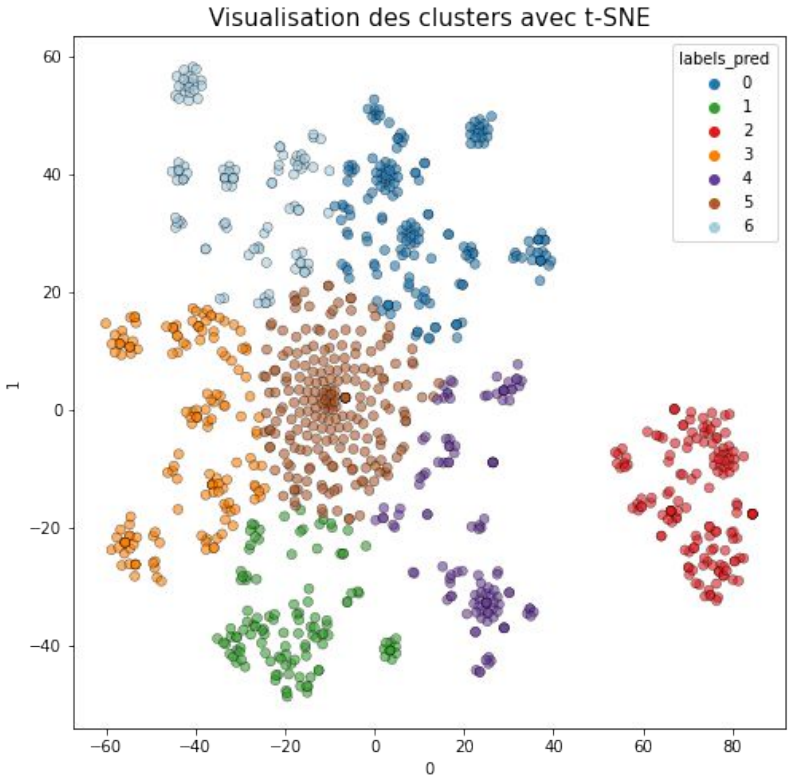


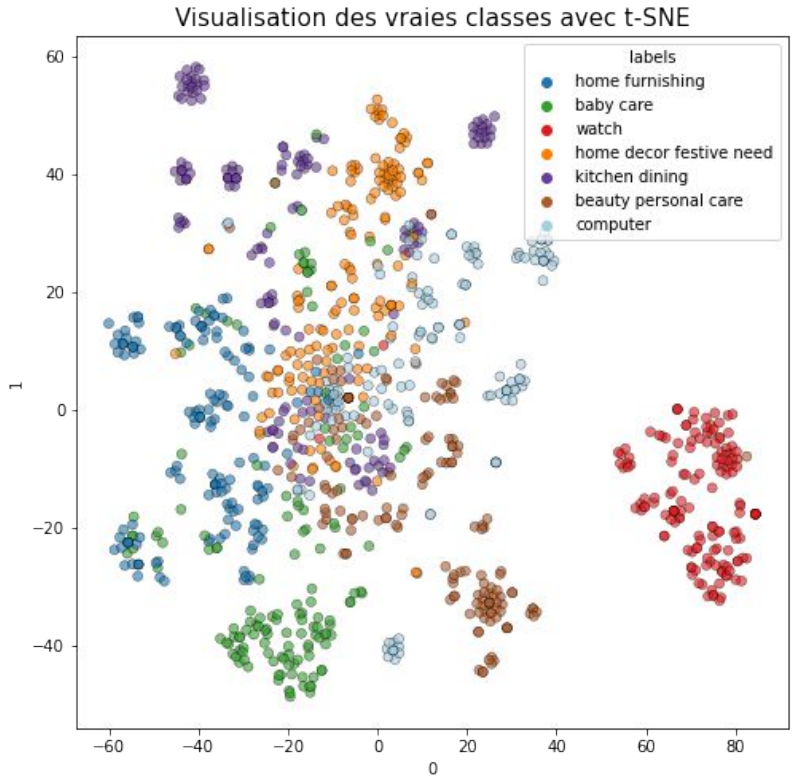


ARI : 0.46

ANALYSE BASÉE SUR LA FRÉQUENCE (TF-IDF)

- Term-Frequency * Inverse Document Frequency
  - TF : Fréquence d'apparition d'un mot **dans un document donné**
  - IDF : Fréquence **inverse** d'un mot **dans le corpus** (Si un mot est rare, cette valeur est élevée)
  - Scores élevés aux mots apparaissant souvent dans peu de documents, donc discriminants

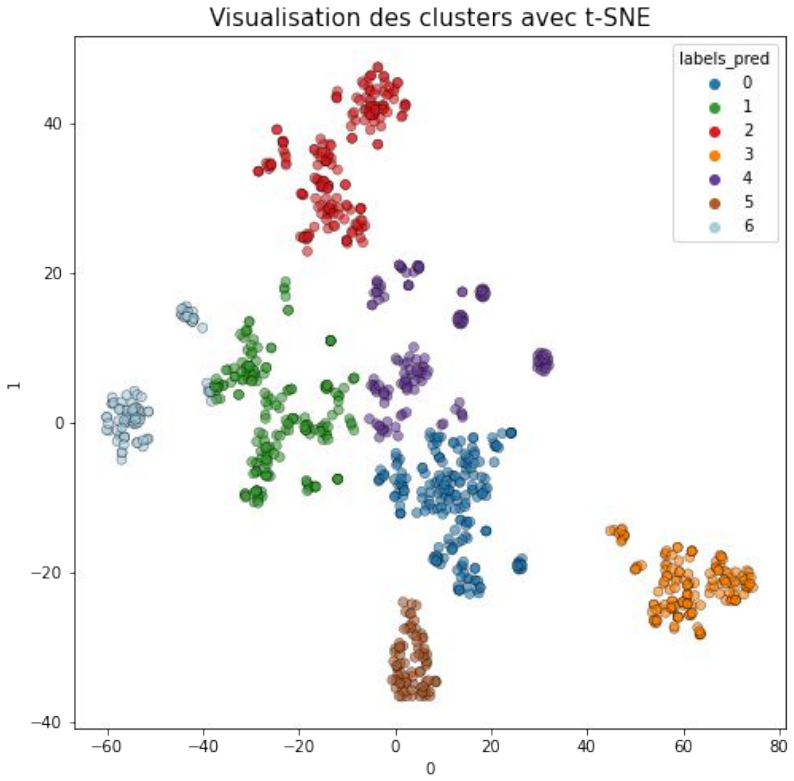


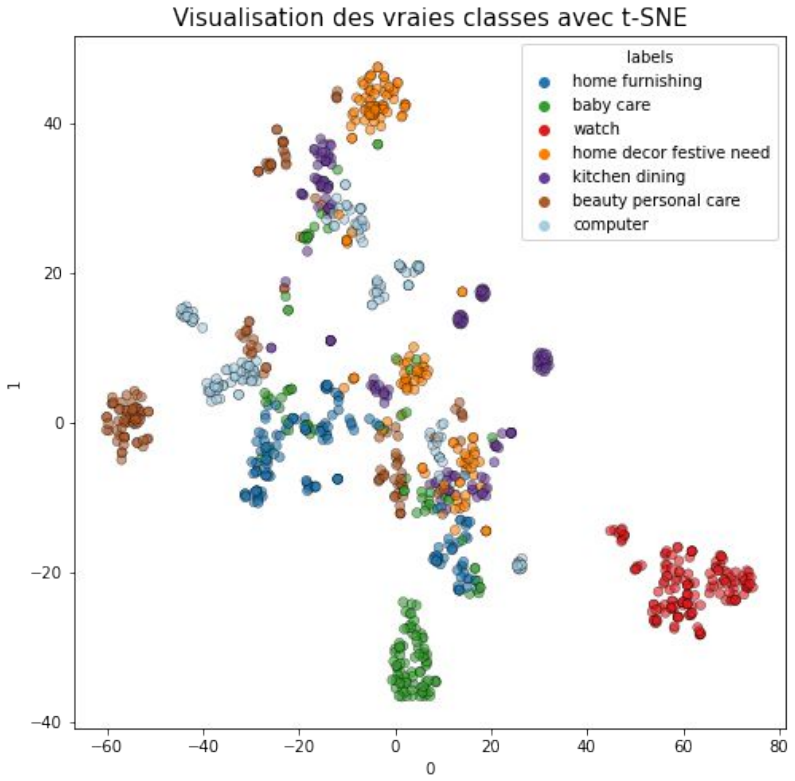


ARI : 0.43

ANALYSE AVEC Word2Vec

- Méthode d'embedding - Prise en compte **du contexte des mots** (Skip-gram) et transformation en vecteurs numériques

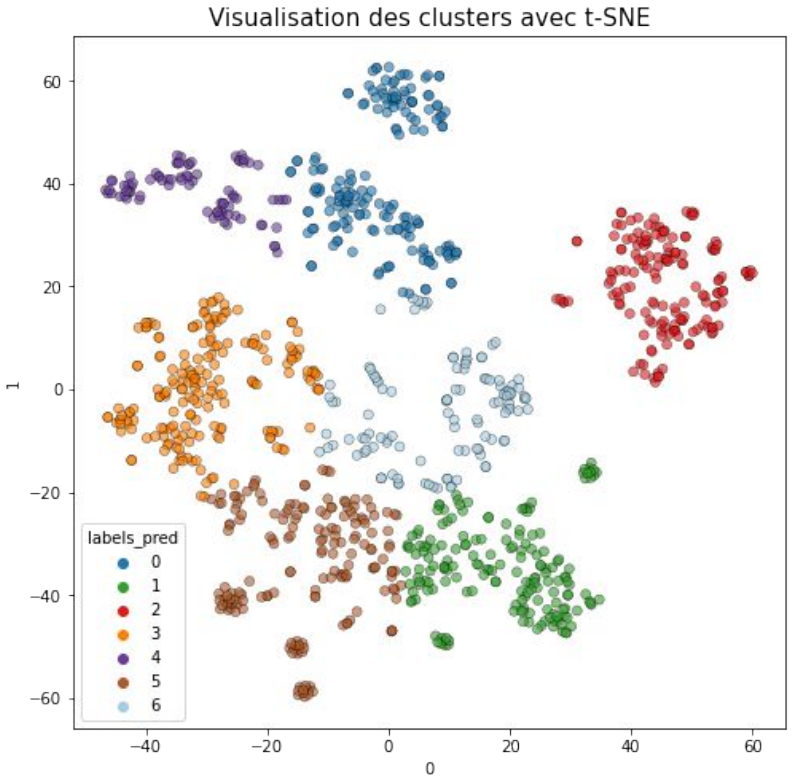


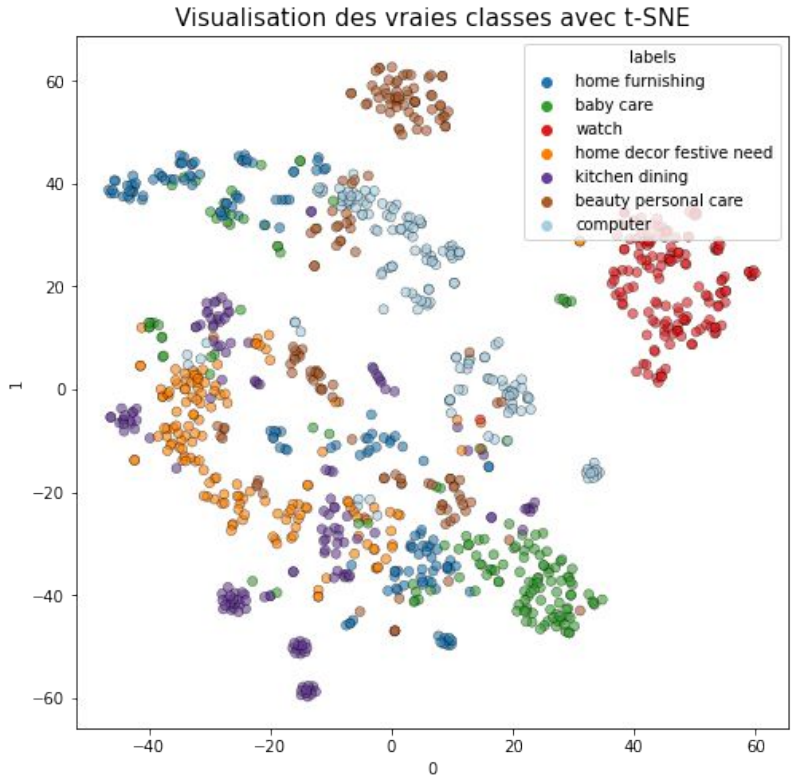


ARI : 0.31

ANALYSE AVEC Bert (*Bidirectional Encoder Representations from Transformers)*

- Modèle de langage développé par Google en 2018

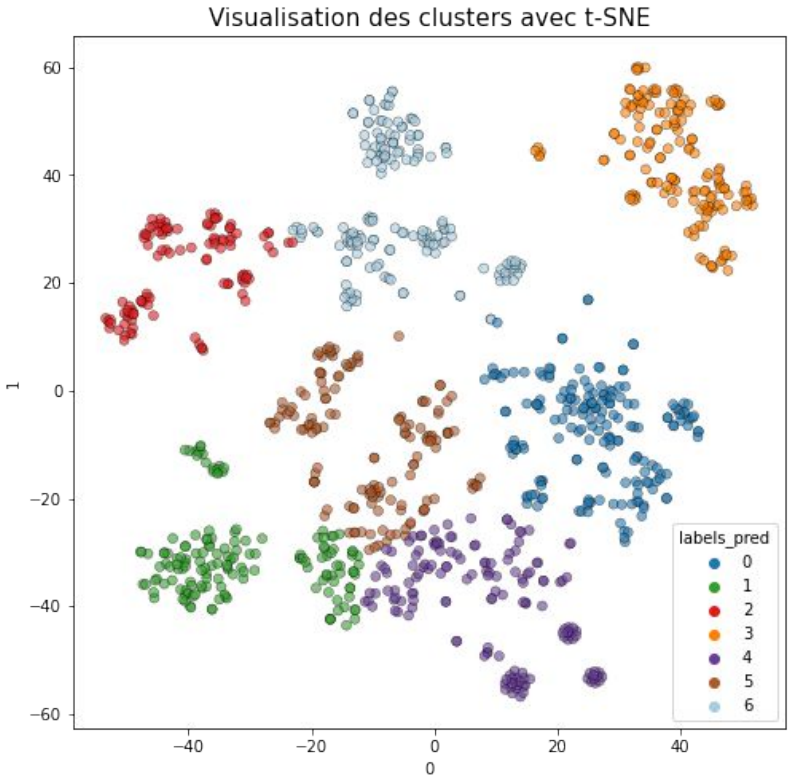


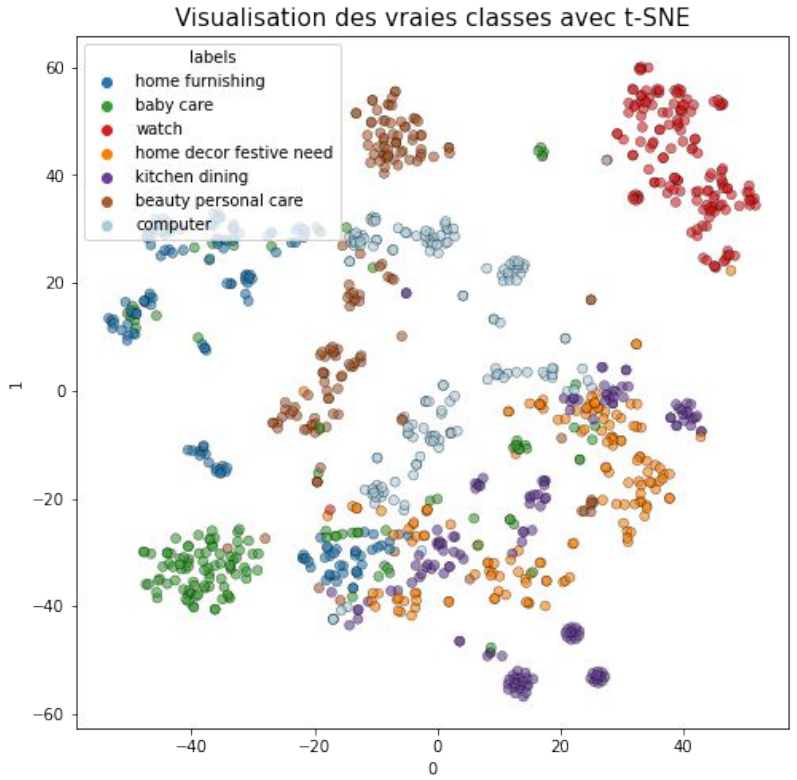


ARI : 0.36

ANALYSE AVEC USE (Universal Sentence Encoder)

- Modèle de langage pré-entraîné, conçu pour être “universel”





ARI : 0.4

CONCLUSIONS - CLASSIFICATION DES TEXTES

Méthode	ARI
CountVectorizer	0.46
TF-IDF	0.43
Word2Vec	0.31
BERT	0.36
USE	0.40

- Les tests de plusieurs méthodes montrent que certaines catégories (montres par ex.) sont facilement détectées.
- Méthodes rapides comme le comptage ou TF-IDF sont efficaces
  - Simples à mettre en place et peu coûteuses
- Limites : En l'état, jeu de données réduit (150 produits / catégories)
- Ajout de la marque des produits ?

La classification automatique à partir des données textuelles semble possible. Avec plus de données et obtenant d'autres informations comme la marque, il semble possible de différencier efficacement les produits.

TRAITEMENT DES IMAGES

- Création des descripteurs avant PCA, t-SNE et clustering
  - Nuance de gris et égalization de l'histogramme
  - Détection d'une liste de descripteurs pour chaque image -> Liste de descripteurs pour l'ensemble du jeu de données
- Clustering :
  - Pour chaque image, prédiction du cluster auxquels ses descripteurs appartiennent
  - Pour chaque cluster, comptage du nombre de descripteurs de l'image qui lui appartiennent (histogramme)
- Les features de l'image seront cet histogramme final
- Réduction de dimensions via PCA (99% de variance conservée) pour garder les feature les plus significatives
- Visualisation avec t-SNE et calcul de l'ARI

TRAITEMENT DES IMAGES (exemple SIFT)

ORIGINAL PICTURE

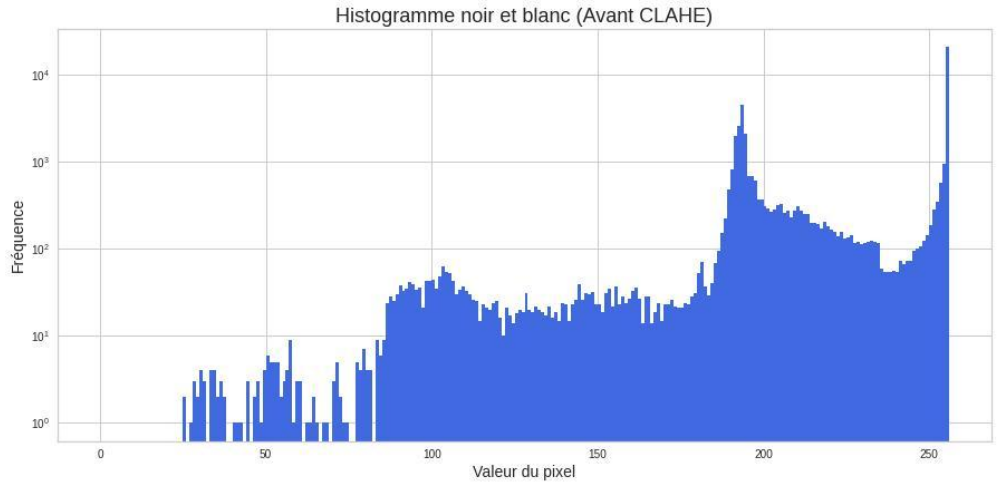
GRAYSCALE + RESIZE

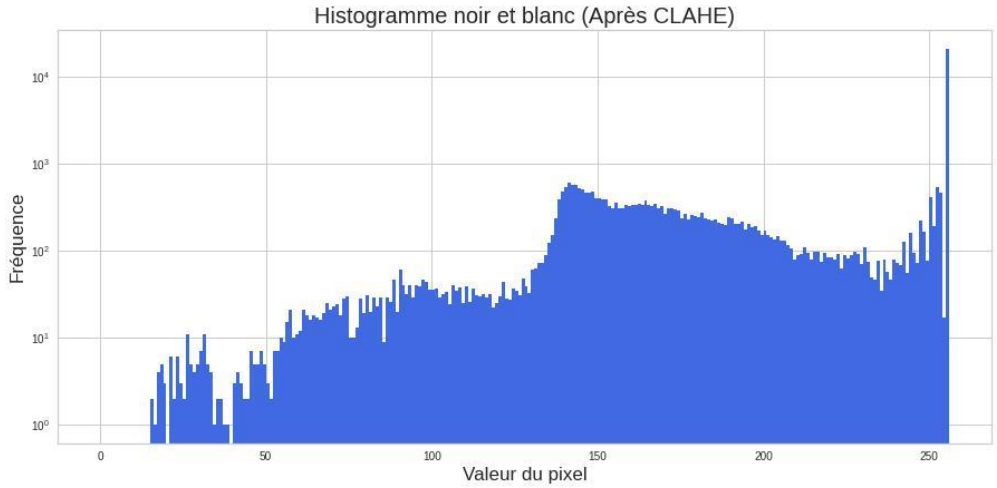
EQUALIZE

REMOVE WHITE BACKGROUND

DENOISING

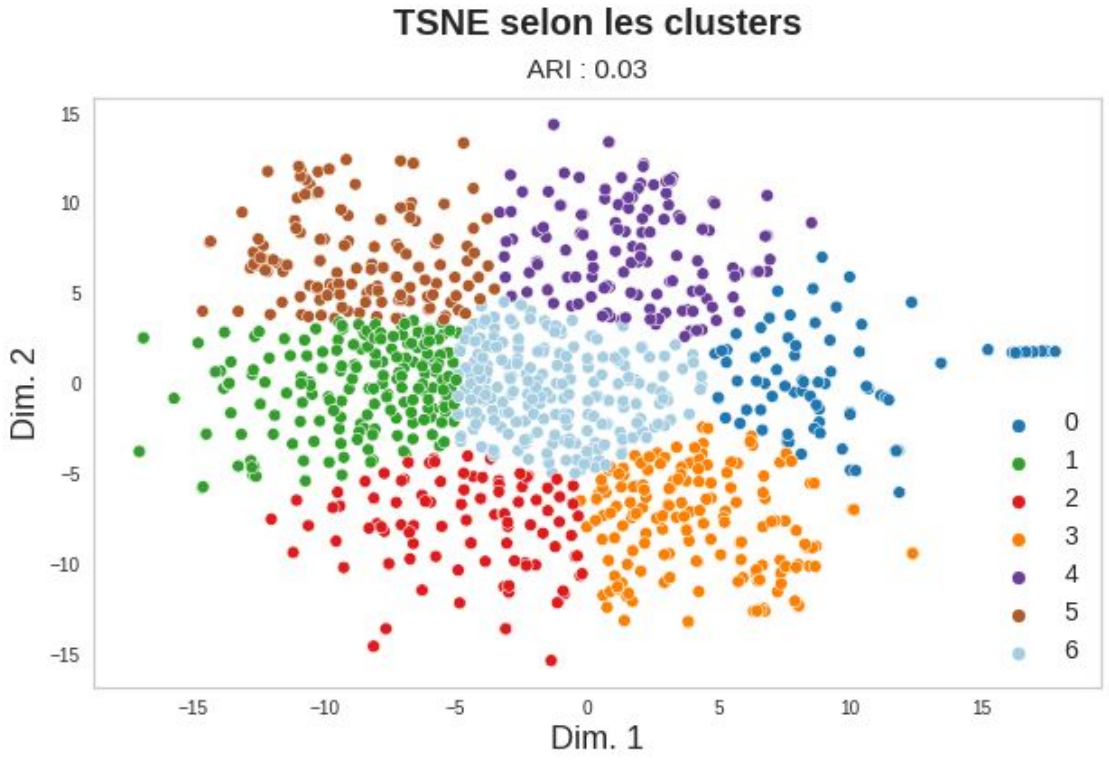
SIFT FEATURES

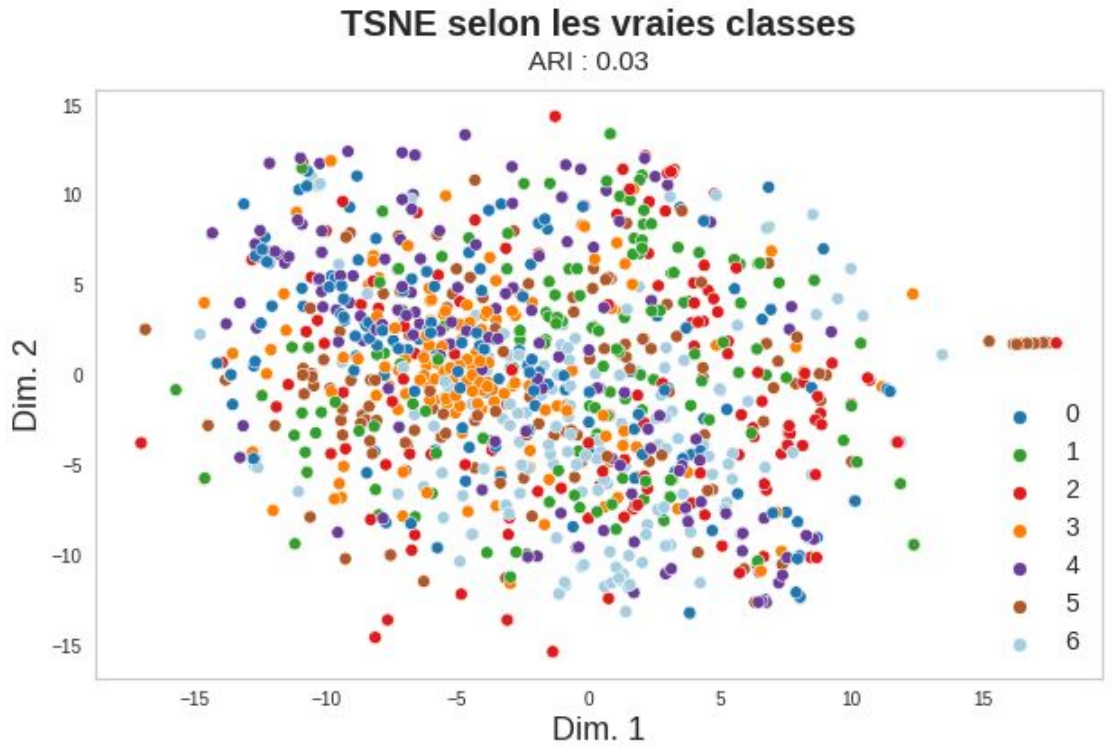




ANALYSE AVEC SIFT (Scale-Invariant Feature Transform)

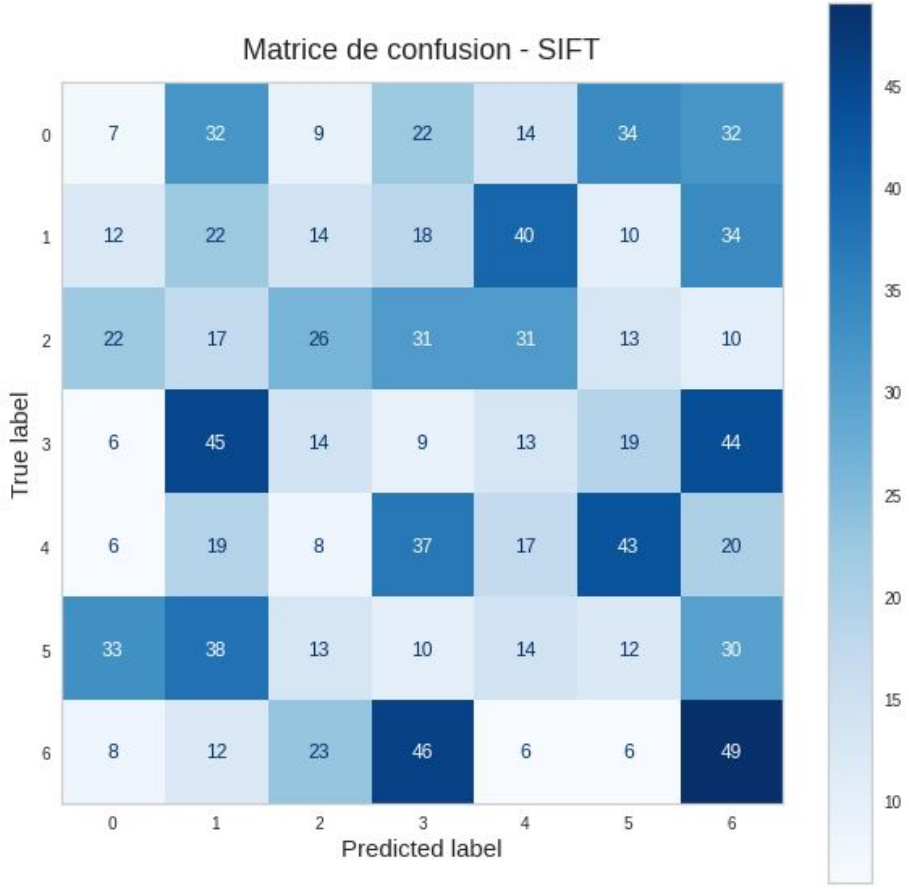
- Algorithme d'analyse d'image très populaire (mais sous licence)
- ARI : 0.03 - Classification impossible





ANALYSE AVEC SIFT (Scale-Invariant Feature Transform)

- Algorithme d’analyse d’image très populaire (mais sous licence)
- ARI : 0.03 - Classification impossible

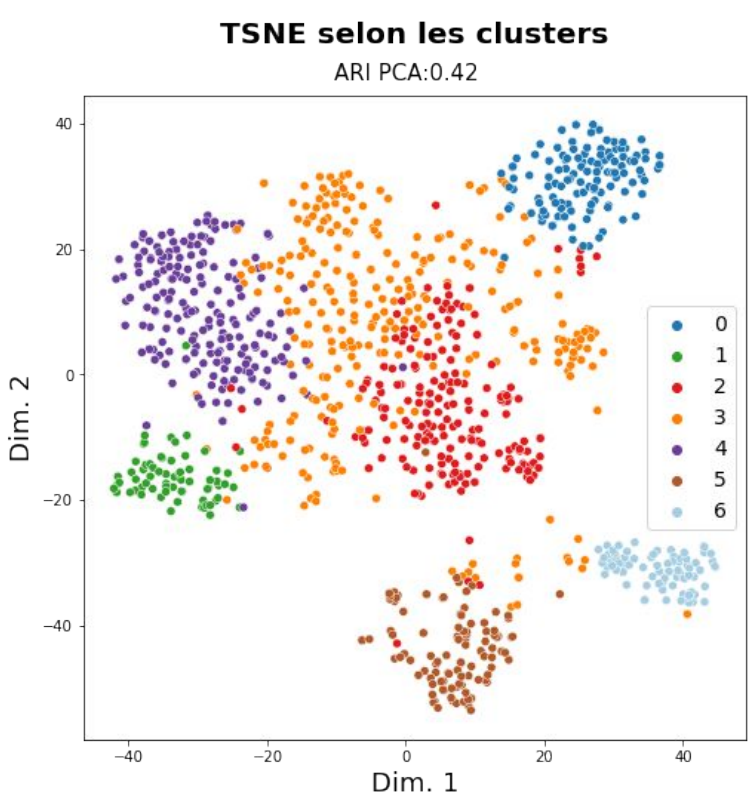


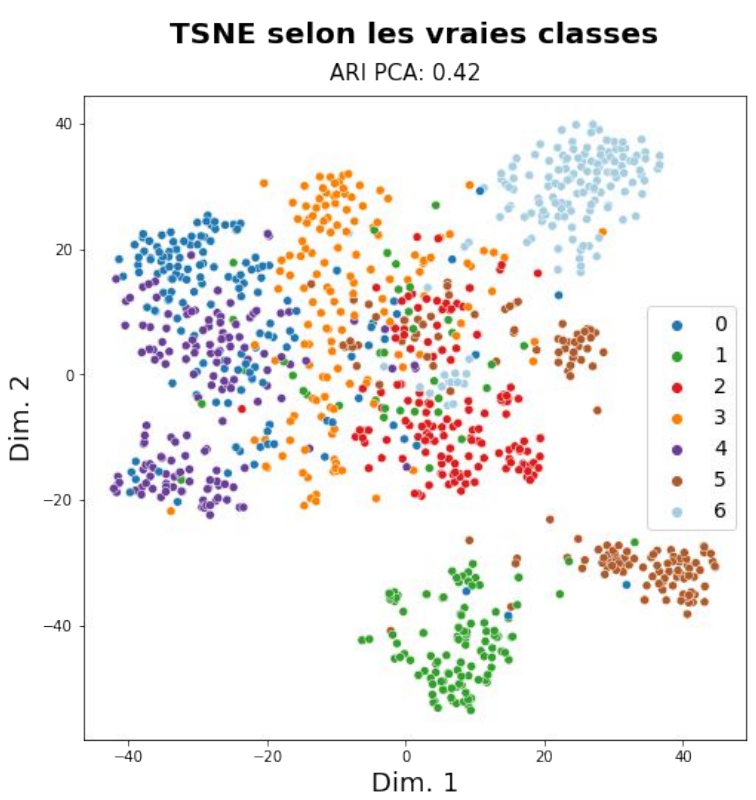
Analyse par CNN transfer learning

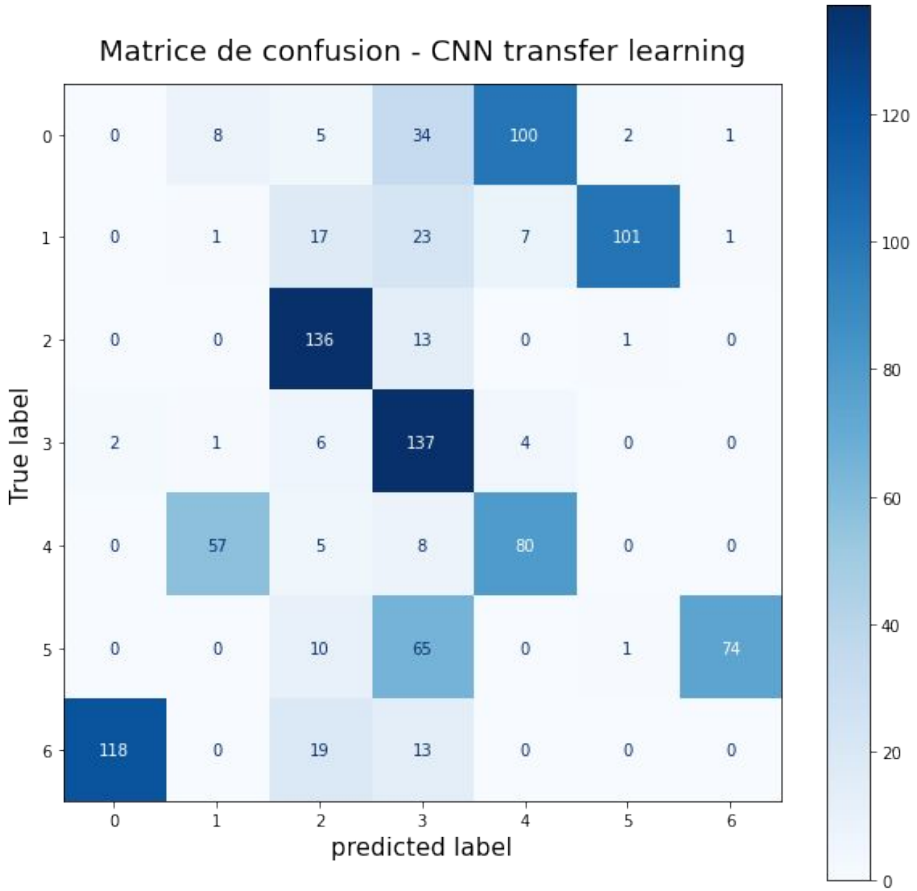
- CNN : Convolutional Neural Network (Réseau de neurones convolutifs)
  - Utilisation d'un modèle pré-entrainé : VGG-16
  - Fonctionnement par couches. Les couches "supérieures" permettent d'extraire les caractéristiques générales des images, les couches plus profonde sont consacrées aux détails plus fins
  - En sélectionnant certaines couches du modèle déjà entraîné sur de nombreuses données, il est possible d'extraire les features de nos images et tester la qualité de la classification.
- Préparation des images différente
  - On conserve la couleur : pas de nuances de gris ni d'égalisation de l'histogramme
  - Redimensionnement des images + réduction du bruit

Analyse par CNN transfer learning

- ARI = 0.42 - Classification possible grâce au modèle pré-entraîné







CONCLUSIONS GÉNÉRALES

- Classification à partir des descriptions + noms de produits possible
  - Méthode peu coûteuse à mettre en place
  - Avec plus de données, possibilité d'améliorer les résultats.
  - Travail sur les noms de marques pour améliorer la classification ?

- Classification à partir des images possible également
  - Nécessité d'utiliser des modèles complexes (réseaux de neurones)
  - Mais : Modèles pré-entraînés semblent suffisant
  - Améliorer les résultats avec d'autres traitements d'image : Détection des bords, masques, …

- Possibilité d'une approche mixte intégrant le texte et les images ?

Classification automatique de biens de consommation

-

Entreprise Place de Marché

Merci pour votre attention

Open Classrooms parcours Data Science - projet 6

Camille Besançon