Étude de faisabilité d’un moteur de classification automatique

Présenté par *Gabriel Chehade*

Introduction

Contexte :

Place de marché souhaite automatiser l'attribution de la catégorie des articles mis en ligne par les vendeurs.

Objectif :

Réaliser une étude de faisabilité d'un moteur de classification des articles.

Méthodes :

- Effectuer un prétraitement des données.
- Extraire les features (texte et image)
- Faire une réduction de dimension T-SNE pour visualiser les données en 2D.
- Faire un clustering sur les données de dimension réduite.
- Effectuer un calcul de similarité entre catégories réelles et clusters

Présentation du jeu de données

- 1050 lignes x 15 colonnes (chaque ligne représente 1 article)
- Variables importantes : *product_name* et *description* pour les données textuelles, *image* pour importer les images, et *product_category_tree* pour les extraire les catégories.

product_name	description	image	product_category_tree
Elegance Polyester Multicolor Abstract Eyelet ...	Key Features of Elegance Polyester Multicolor ...	55b85ea15a1536d46b7190ad6fff8ce7.jpg	["Home Furnishing >> Curtains & Accessories >>...
Sathiyas Cotton Bath Towel	Specifications of Sathiyas Cotton Bath Towel (...	7b72c92c2f6c40268628ec5f14c6d590.jpg	["Baby Care >> Baby Bath & Skin >> Baby Bath T...
Eurospa Cotton Terry Face Towel Set	Key Features of Eurospa Cotton Terry Face Towe...	64d5d4a258243731dc7bbb1eef49ad74.jpg	["Baby Care >> Baby Bath & Skin >> Baby Bath T...
SANTOSH ROYAL FASHION Cotton Printed King size...	Key Features of SANTOSH ROYAL FASHION Cotton P...	d4684dcdc759dd9cdf41504698d737d8.jpg	["Home Furnishing >> Bed Linen >> Bedsheets >>...
Jaipur Print Cotton Floral King sized Double B...	Key Features of Jaipur Print Cotton Floral Kin...	6325b6870c54cd47be6ebfbffa620ec7.jpg	["Home Furnishing >> Bed Linen >> Bedsheets >>...

Extraction des catégories

	product_category_tree
0	["Home Furnishing >> Curtains & Accessories >>...
1	["Baby Care >> Baby Bath & Skin >> Baby Bath T...
2	["Baby Care >> Baby Bath & Skin >> Baby Bath T...
3	["Home Furnishing >> Bed Linen >> Bedsheets >>...
4	["Home Furnishing >> Bed Linen >> Bedsheets >>...

	category
0	Home Furnishing
1	Baby Care
2	Baby Care
3	Home Furnishing
4	Home Furnishing

```
In [14]:  1 df_text['category'].value_counts()
```

executed in 5ms, finished 11:32:32 2022-10-23

```
Out[14]: Home Furnishing              150
         Baby Care                    150
         Watches                      150
         Home Decor & Festive Needs   150
         Kitchen & Dining             150
         Beauty and Personal Care     150
         Computers                    150
         Name: category, dtype: int64
```

- 7 catégories
- 150 échantillons par catégorie (jeu de données équilibré)

Analyse des données textuelles

Preprocessing

- Document = nom produit + description
- Mise en minuscule du texte
- Tokenization
- Suppression des mots uniques
- Suppression des mots courts (< 3 lettres)
- Tokens alphabétiques uniquement
- Suppression des mots fréquents communs à toutes les catégories
- Création de 2 corpus : un sur lequel on a appliqué le *stemming*, un autre sur lequel on a appliqué la *lemmatization*.

Extraction des Features

Extraction des features texte avec :

- deux approches de type “bag-of-words”, **comptage simple** de mots et **Tf-idf** ;
- une approche de type word/sentence embedding classique avec **Word2Vec** ;
- une approche de type word/sentence embedding avec **BERT** ;
- une approche de type word/sentence embedding avec **USE**.

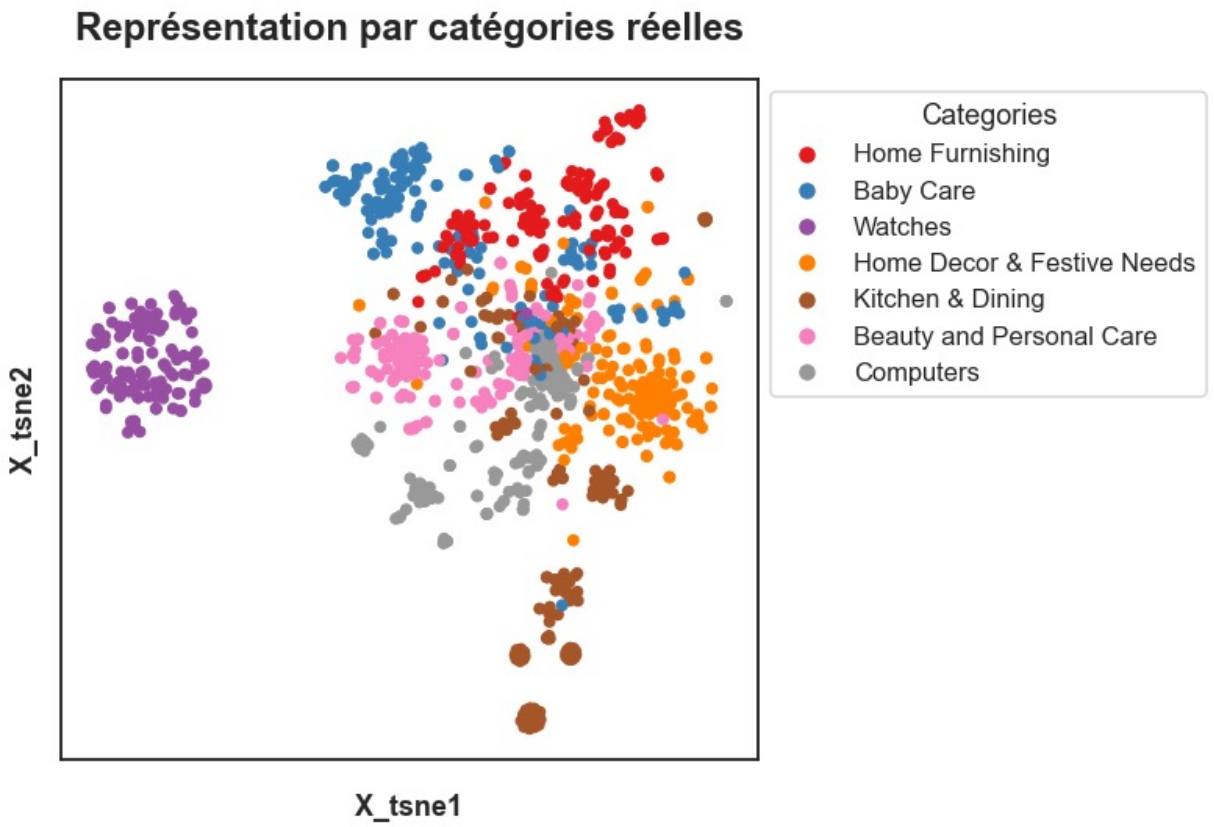
Remarque : Pour les modèles de Deep Learning (BERT et USE) aucun prétraitement n’a été effectué.

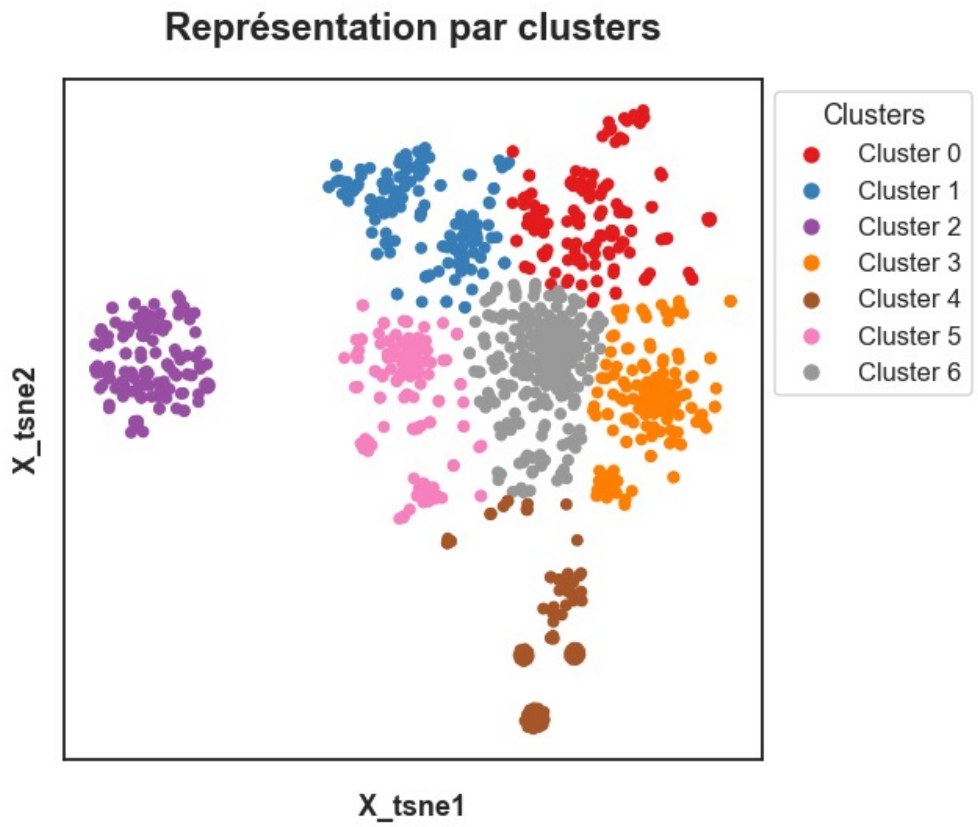
Cluestering et mesure de similarité

La mesure de similarité s'effectue de la manière suivante :

1. Extraction des features
2. Réduction dimensionnelle dans un espace 2D par T-SNE
3. Clustering sur les données projetées (KMeans)
4. Mesure de l'ARI entre les catégories réelles et les clusters

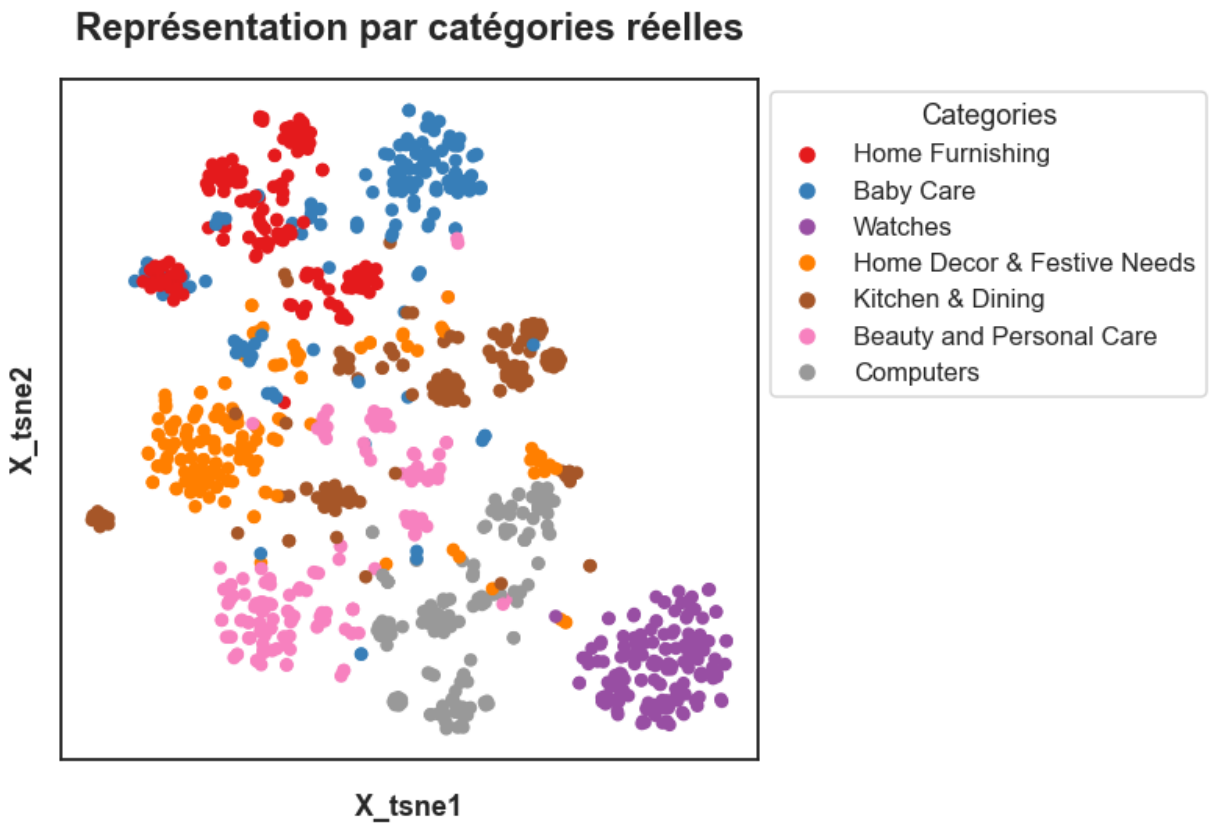
Comptage simple

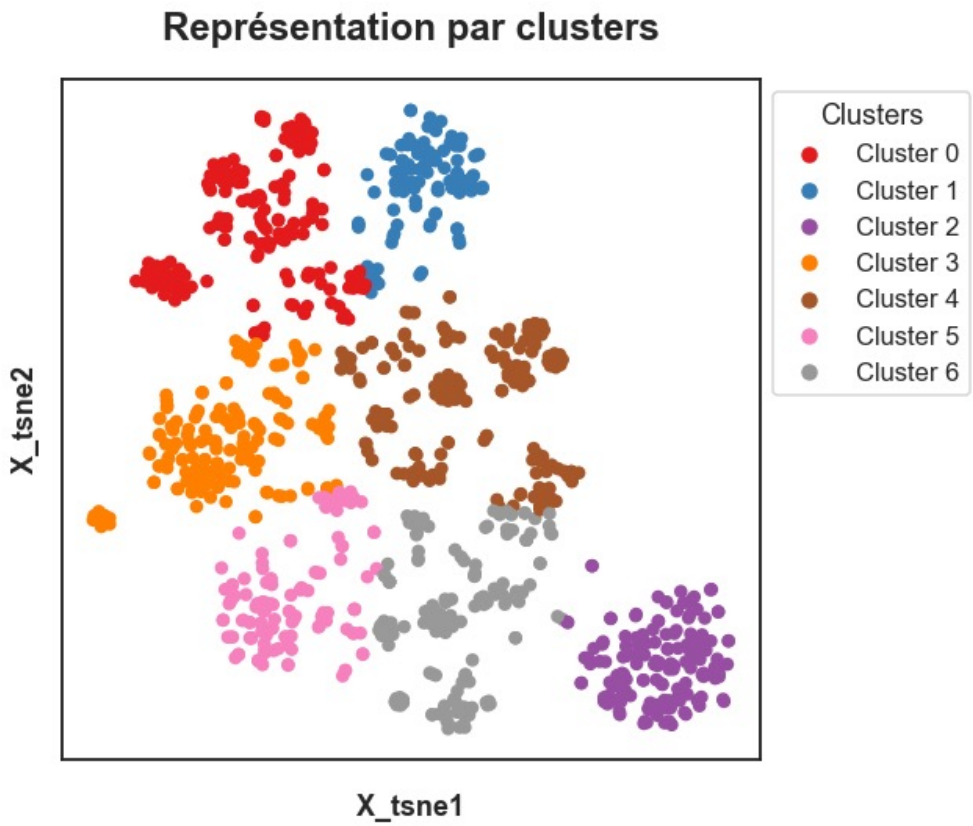




$$ARI = 0.45$$

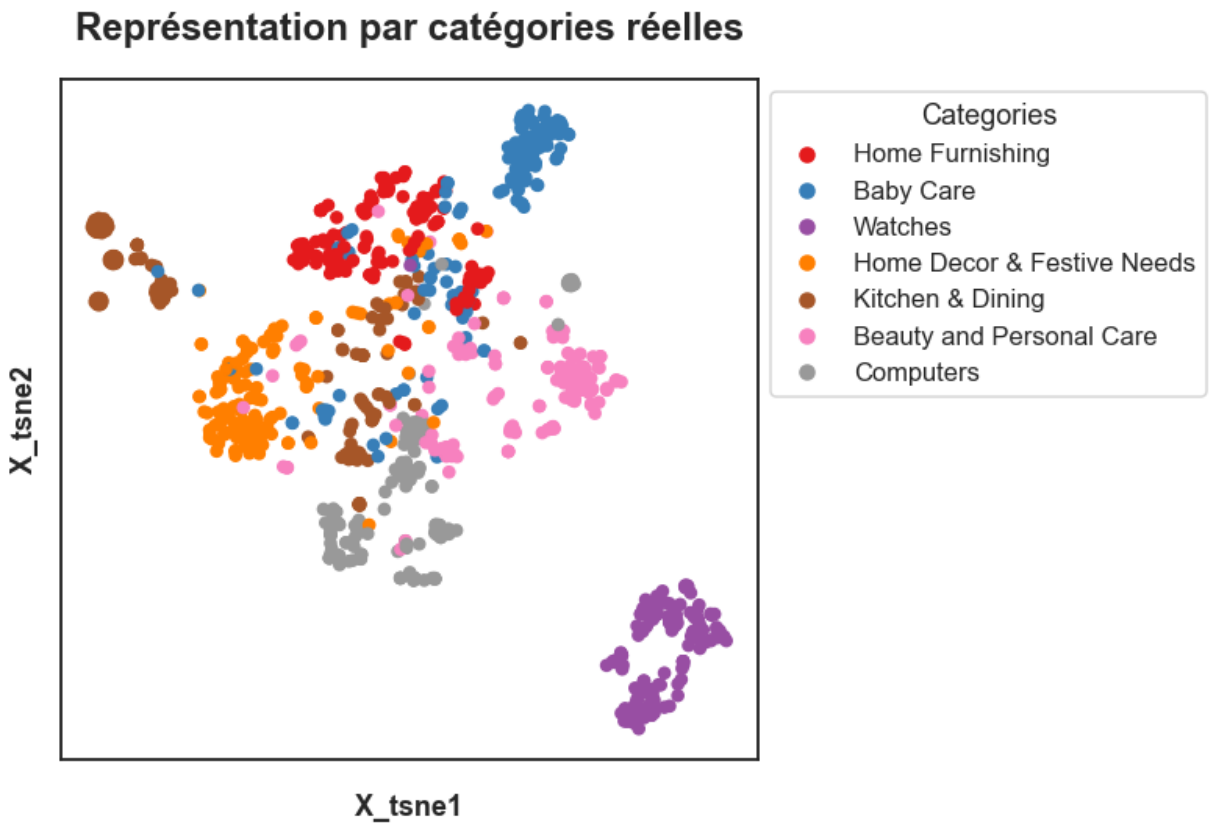
TF-IDF

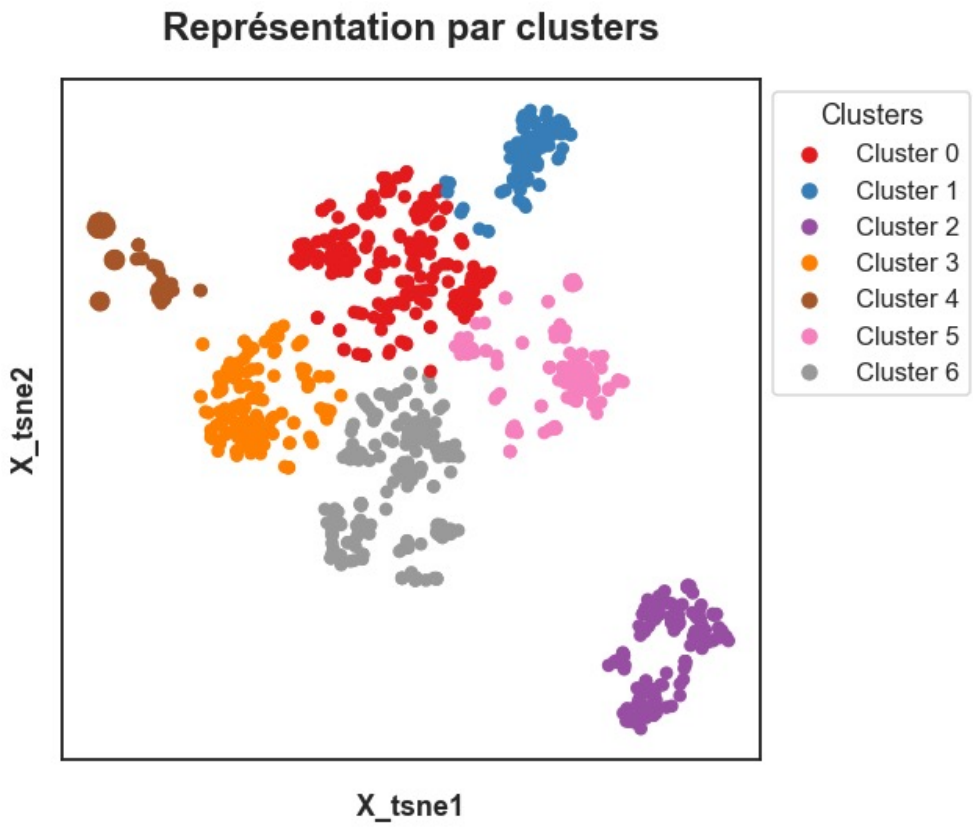




$$ARI = 0.60$$

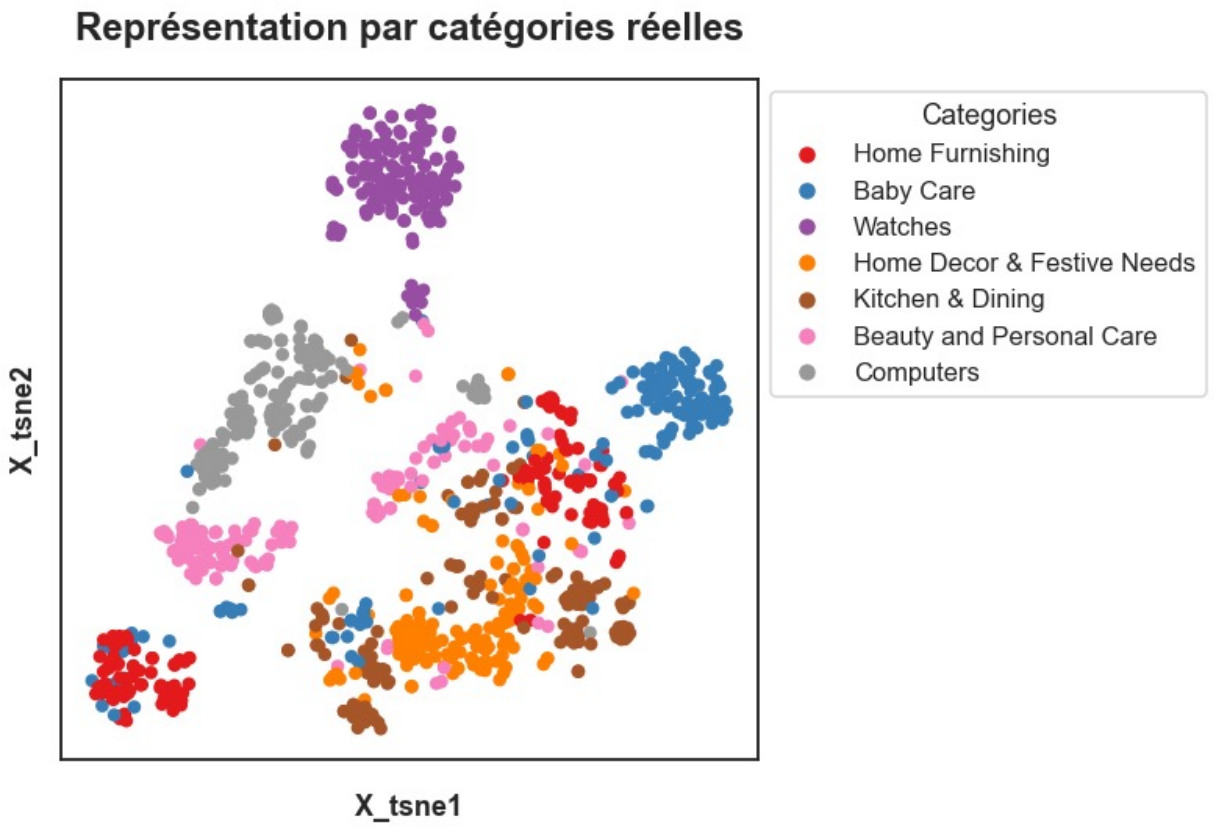
Word2Vec

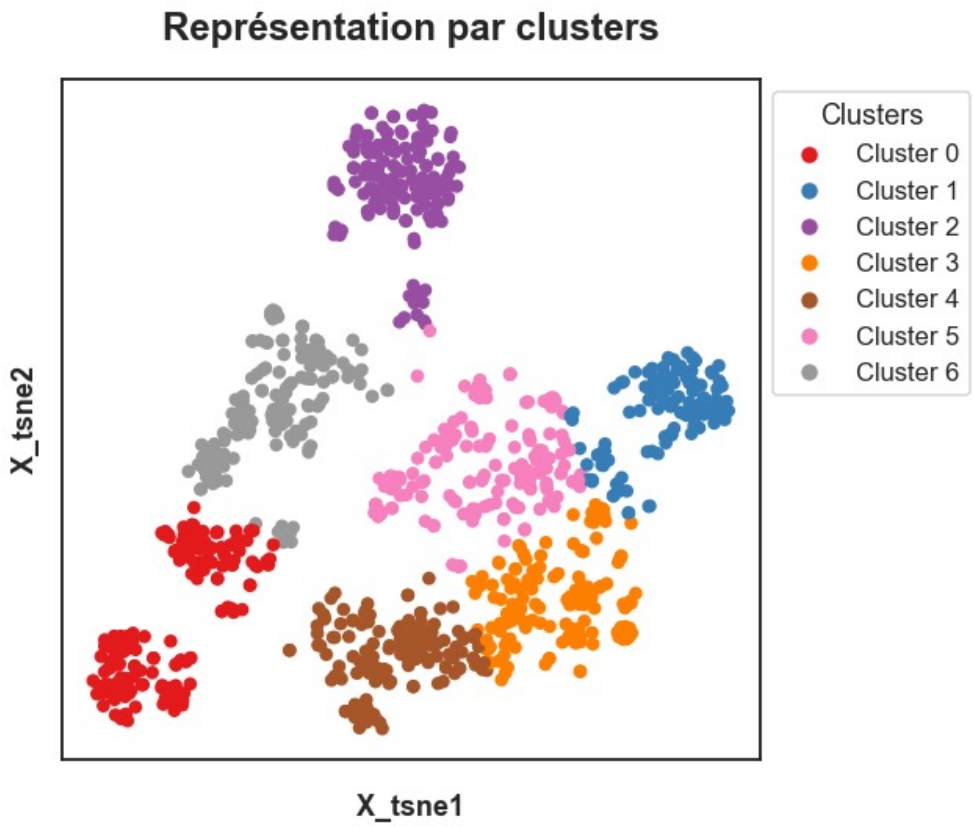




$$ARI = 0.59$$

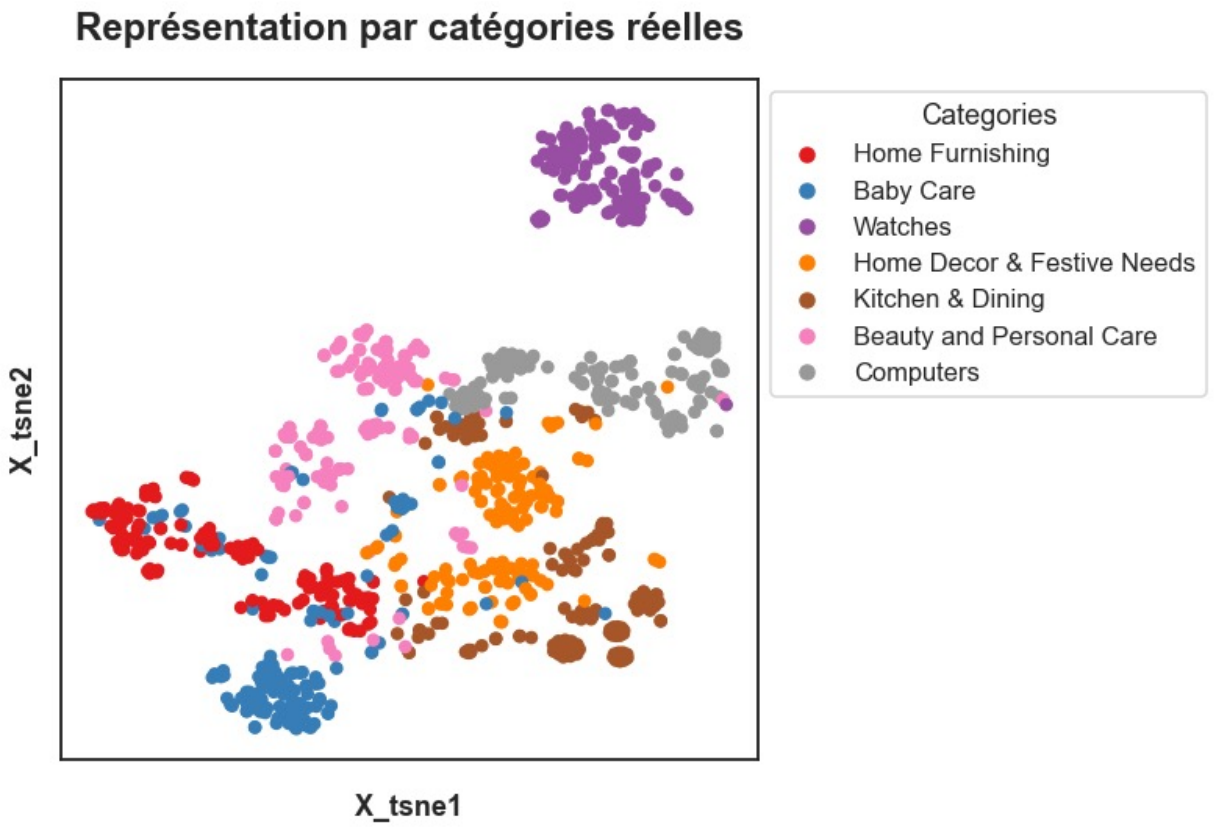
BERT

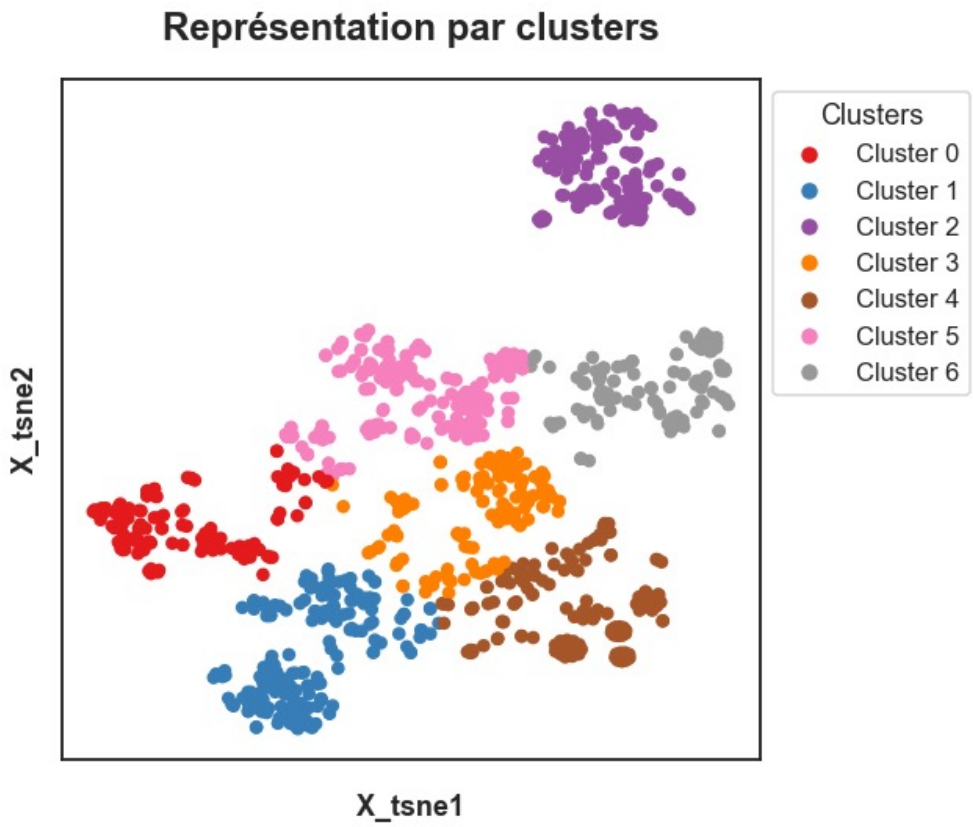

Représentation par catégories réelles
X_tsne2
X_tsne1
Categories
Home Furnishing
Baby Care
Watches
Home Decor & Festive Needs
Kitchen & Dining
Beauty and Personal Care
Computers


Représentation par clusters
X_tsne2
X_tsne1
Clusters
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6

$$ARI = 0.43$$

USE


Représentation par catégories réelles
X_tsne2
X_tsne1
Categories
Home Furnishing
Baby Care
Watches
Home Decor & Festive Needs
Kitchen & Dining
Beauty and Personal Care
Computers


Représentation par clusters
X_tsne2
X_tsne1
Clusters
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6

$$ARI = 0.52$$

Synthèse des résultats

	Comptage	TF-IDF	Word2Vec	BERT	USE
ARI	0.45	0.60	0.59	0.42	0.52
Temps d'exécution	5 s	5 s	11 s	64 s	5 s
Accuracy sur le test set	92.4%	85.7%	85.7%	91.4%	86.7%

Obtenu avec un modèle de **régression logistique**

Analyse des images

Extraction des Features

Extraction des features image avec :

- un algorithme **SIFT** ;
- un algorithme de type CNN Transfer Learning.

Pour le Transfer Learning j'ai utilisé le **VGG-16** fourni par Keras et pré-entrainé sur ImageNet.

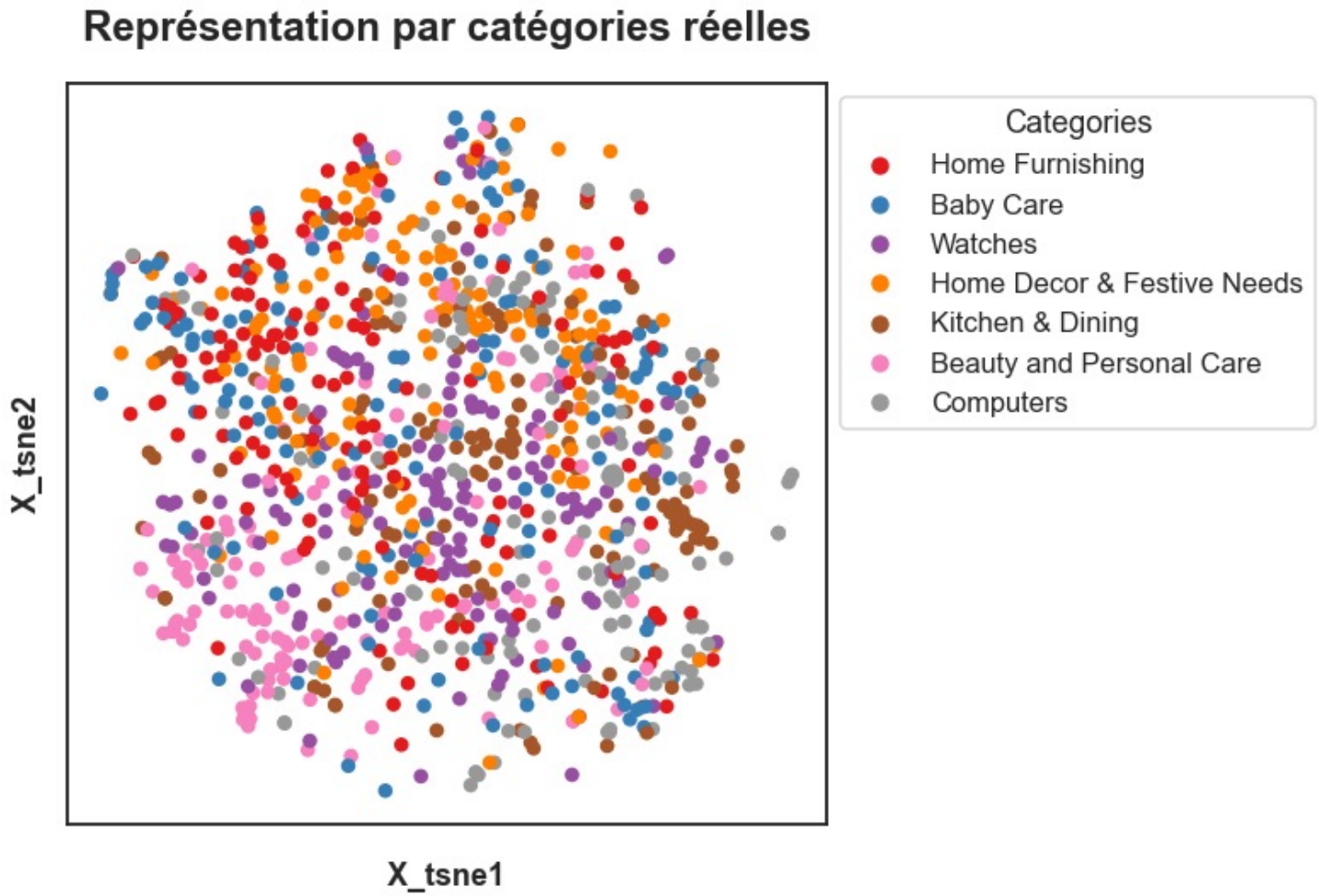
Extraction des features avec SIFT

- Création des descripteurs :
  - *sift_keypoints_all* : matrice dont chaque ligne représente un descripteur.
- Estimation du nombre de clusters k :

$$k = \mathrm{Int}\left(\sqrt{nombre\ de\ descripteurs}\right) = 719$$

- Constructions des histogrammes (nb de descripteurs par cluster) → matrice (1050, 719)
- Réduction dimensionnelle par ACP : 719 → 495 features (99% de la variance)
- Réduction 2D avec T-SNE

Extraction des features avec SIFT

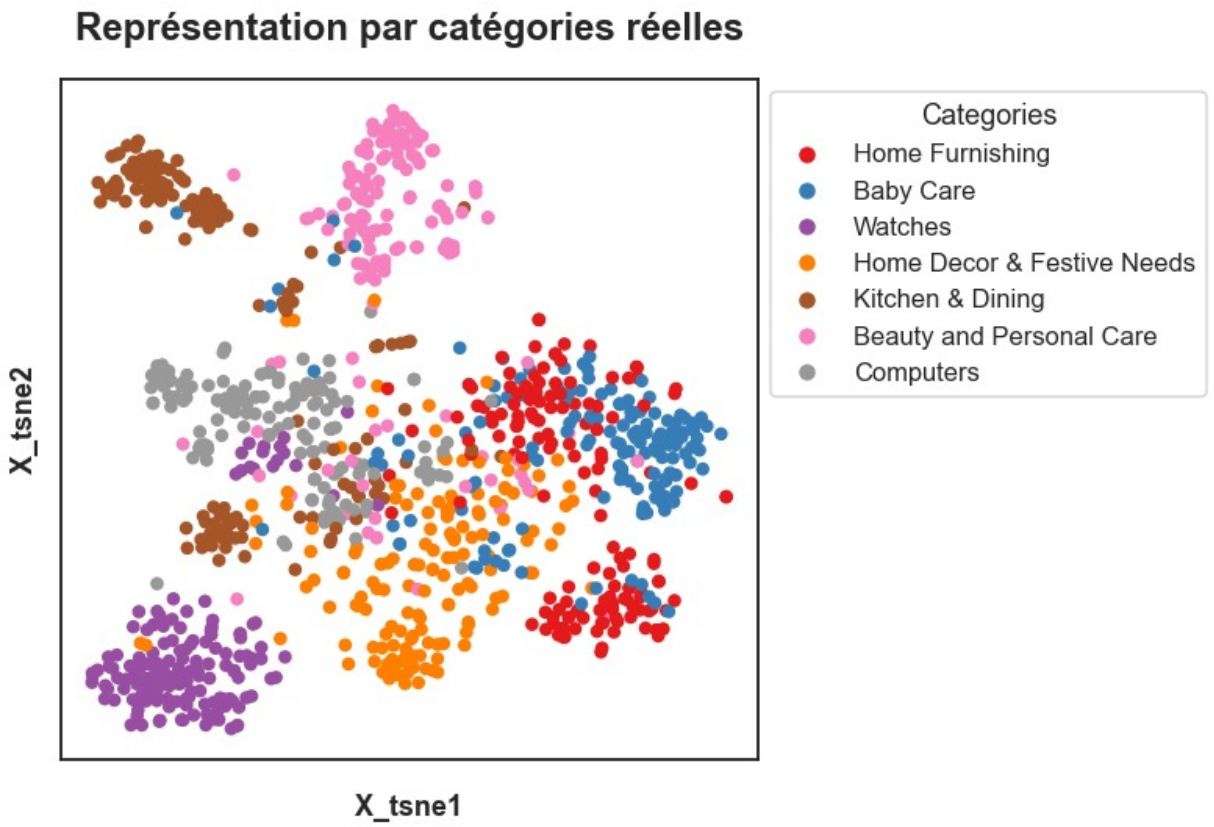


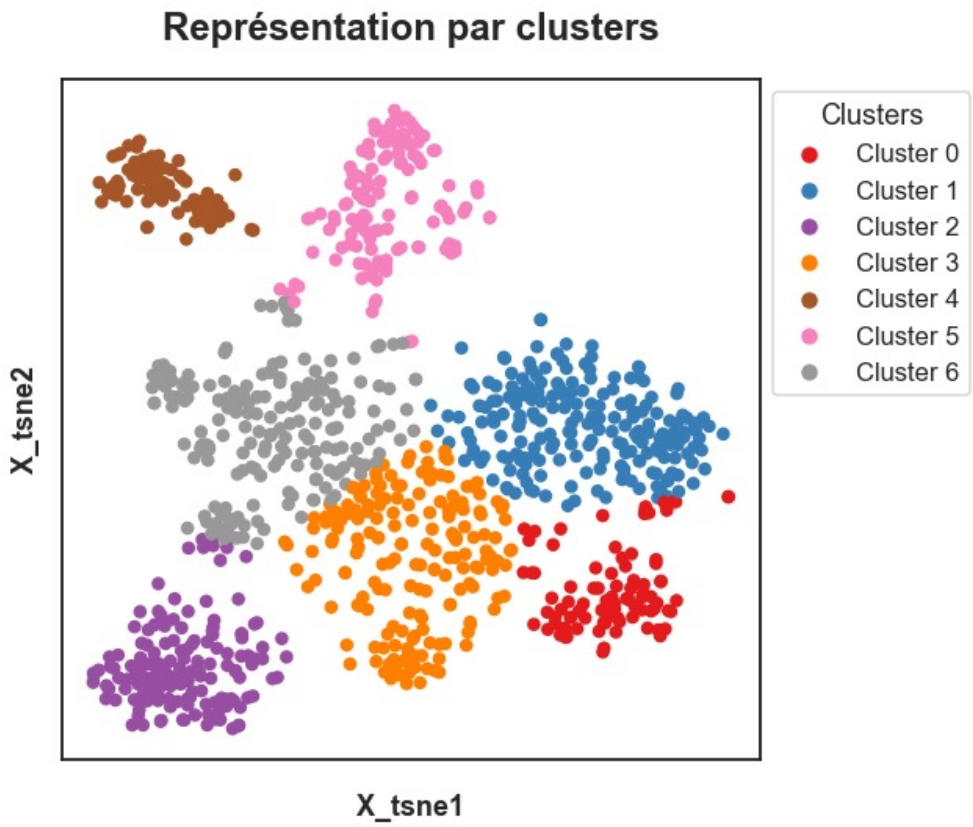
Les catégories ne sont pas du tout séparées (ARI < 0.1)

CNN Transfer Learning

- Utilisation du modèle VGG-16 privé de sa dernière couche (couche qui sert à la classification)
- Chargement des images au format (224, 224, 3)
- Conversion en tableau numpy
- Pré-traitement des images
- Extraction des features avec model.predict → X.shape = (1050, 4096)
- Réduction dimensionnelle par ACP : 4096 → 803 features (99% de la variance)
- Réduction 2D avec T-SNE

CNN Transfer Learning





$$ARI = 0.45$$

accuracy = 85.7%

Conclusion

- Le modèle le plus adapté pour extraire les features texte est le bag-of-words.
- Le modèle le plus adapté pour extraire les features image est le CNN Transfer Learning.
- Les résultats d'un modèle simple d'apprentissage supervisé (régression logistique) et non optimisé ont été très bons, à la fois sur les données textuelles (accuracy > 92%) et les images (accuracy > 85%).
- **Il est possible de créer un moteur de classification automatique des articles.**

MERCI POUR VOTRE ATTENTION