Classifiez automatiquement des biens de consommation

Projet 6

Sommaire

- Rappel de la problématique
- Présentation du jeu de données
- Faisabilité de classification par les descriprions textulles
- Faisabilité de classification automatique d'images
- Conclusion sur la faisabilité du moteur de classification

Rappel de la problématique

L’entreprise "Place de marché” souhaite lancer une marketplace e-commerce

La mission: réaliser une première étude de faisabilité d'un moteur de classification d'articles, basé sur une image et une description, pour l'automatisation de l'attribution de la catégorie de l'article

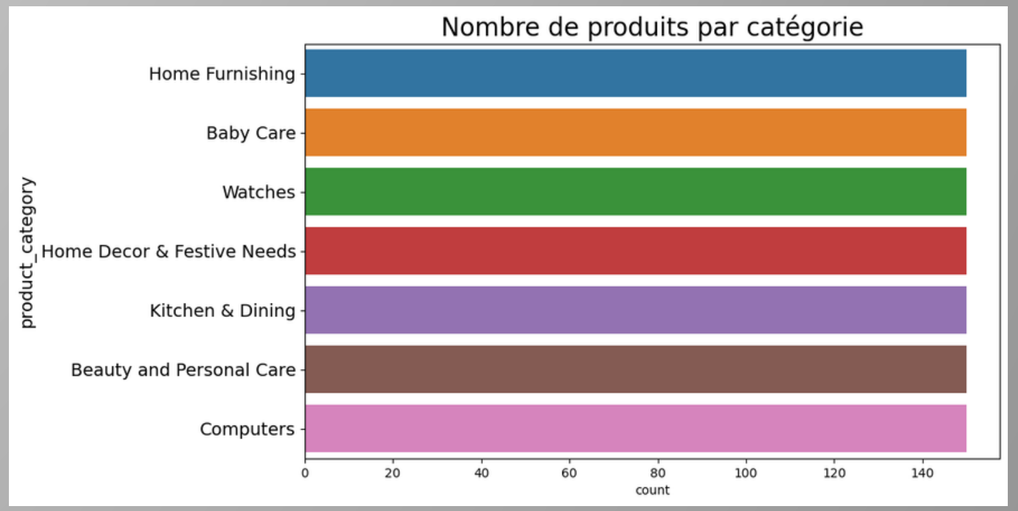
La méthode :

- analyser le jeu de données
- réaliser un prétraitement des descriptions des produits et des images,
- faire une réduction de dimension, puis un clustering.

Présentation du jeu de données

- flipkart_com-ecommerce_sample_1050.csv
- Sample d'un jeu de données de Flipkart.com, leader Indien de l'e-Commerce
- dimensions: (1050, 15)
- 341 valeurs manquantes:

	type de variable	valeurs uniques	valeurs manquantes
uniq_id	object	1050	0
crawl_timestamp	object	149	0
product_url	object	1050	0
product_name	object	1050	0
product_category_tree	object	642	0
pid	object	1050	0
retail_price	float64	354	1
discounted_price	float64	424	1
image	object	1050	0
is_FK_Advantage_product	bool	2	0
description	object	1050	0
product_rating	object	27	0
overall_rating	object	27	0
brand	object	490	338
product_specifications	object	984	1



Faisabilité de classification par le texte

La démarche

01

PRÉTRAITEMENT DE TEXTE:

- nettoyage (ponctuation, chiffres stopwords, majuscules…),
- lemmatization…

02

BAG OF WORDS

countVectorizer, Tf-idf

02

Text Embedding

Word2Vec, BERT, USE

03

CLUSTERING

- Reduction de dimensions PCA, T-SNE
- Clustering k-means
- Adjusted Rand Score (ARI)
- Matrice de confusion

Faisabilité de classification par le texte

preprocessing

Traitements efféctués:

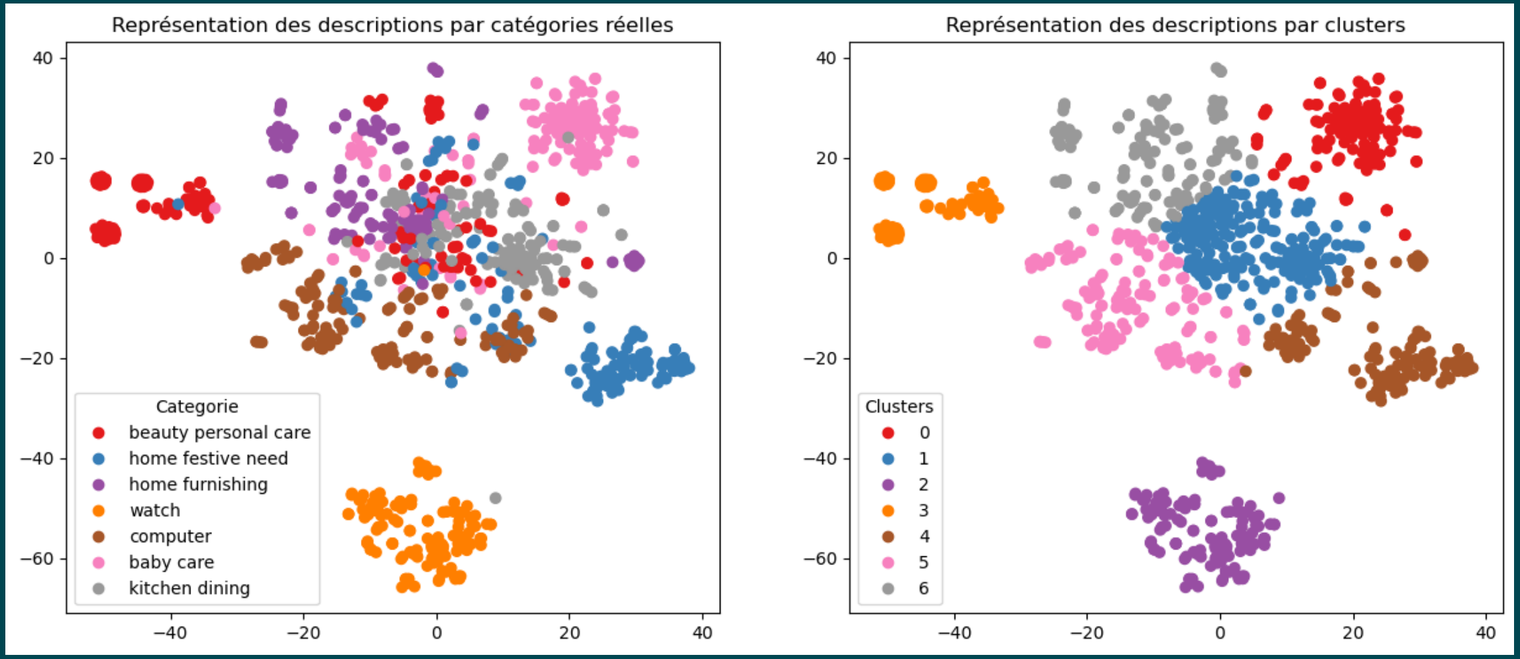
- *la ponctuation*
- *les chiffres*
- *tokenisation*
- *les stopwords*
- *lemmatization*
- *enlever les majuscules*

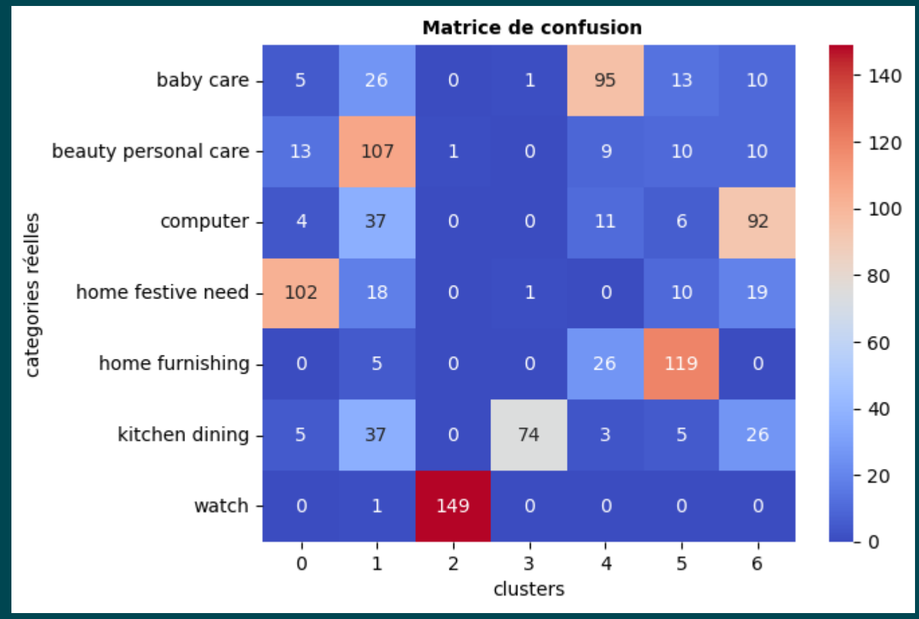
```
**Exemple de nettoyage de texte**
Phrase de base : Sathiyas Cotton Bath Towel Specifications of Sathiy
as Cotton Bath Towel (3 Bath Towel, Red, Yellow, Blue) Bath Towel Fe
atures Machine Washable Yes Material Cotton Design Self Design Gener
al Brand Sathiyas Type Bath Towel GSM 500 Model Name Sathiyas cotton
bath towel Ideal For Men, Women, Boys, Girls Model ID asvtwl322 Colo
r Red, Yellow, Blue Size Mediam Dimensions Length 30 inch Width 60 i
nch In the Box Number of Contents in Sales Package 3 Sales Package 3
Bath Towel
-------
-------
Phrase nettoyée : cotton bath towel specification cotton bath towel
bath towel red yellow blue bath towel feature machine washable yes m
aterial cotton design self design general brand type bath towel mode
l name cotton bath towel ideal men woman boy girl model id color red
yellow blue size dimension length inch width inch box number content
sale package sale package bath towel
```

Faisabilité de classification par le texte: BOW-CountVectorizer

ARI : 0.4515

CountVectorizer extrait les termes les plus fréquents d'une collection de textes et convertit ces termes en vecteurs

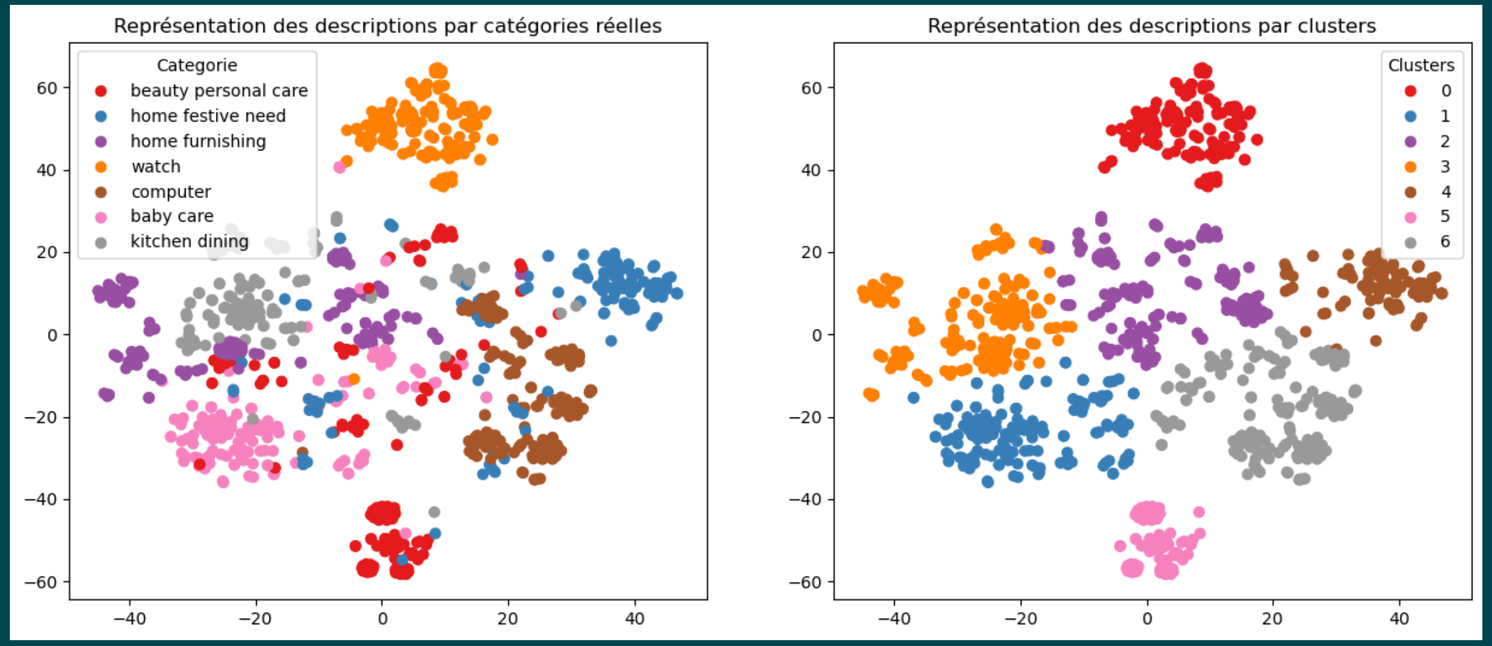


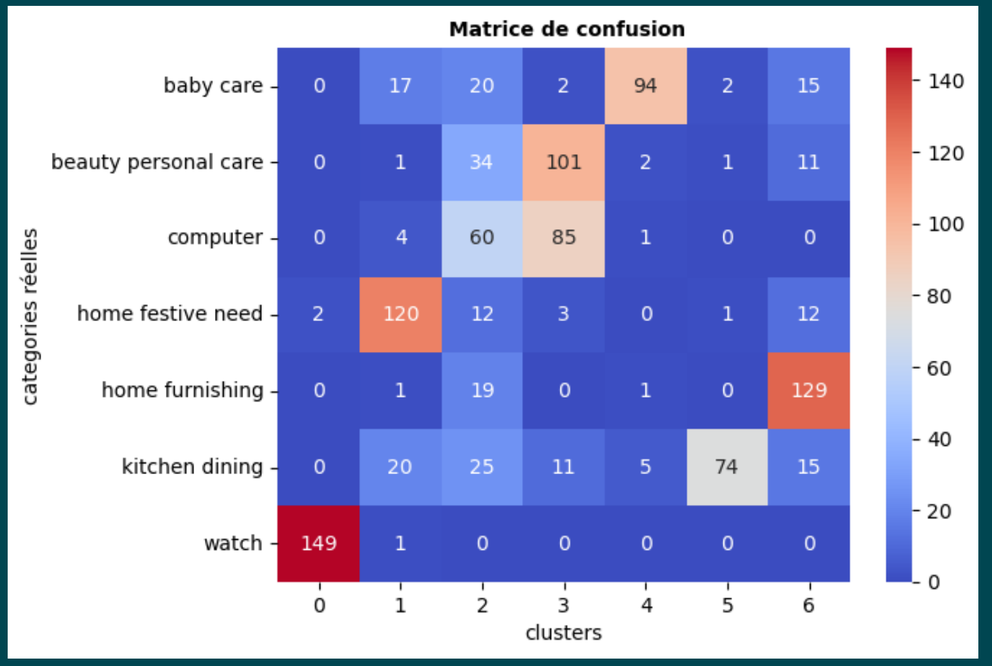


Faisabilité de classification par le texte: BOW-Tf-idf

Tf-idf: mesure la pertinence d'un mot en se basant sur sa rareté dans un ensemble de pages.

ARI : 0.4907

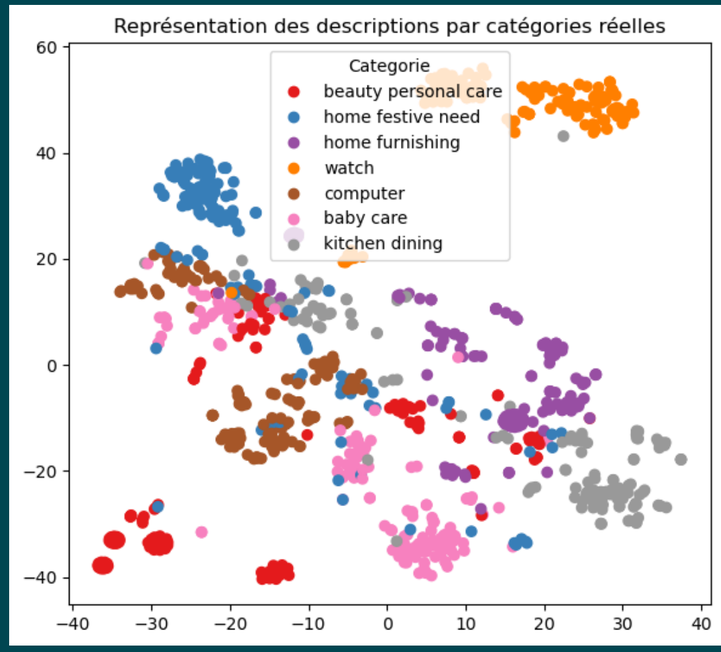


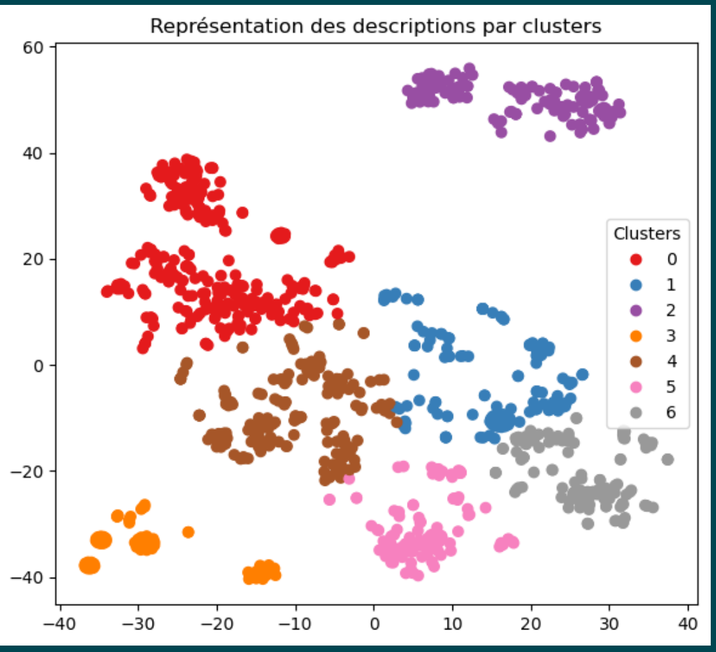


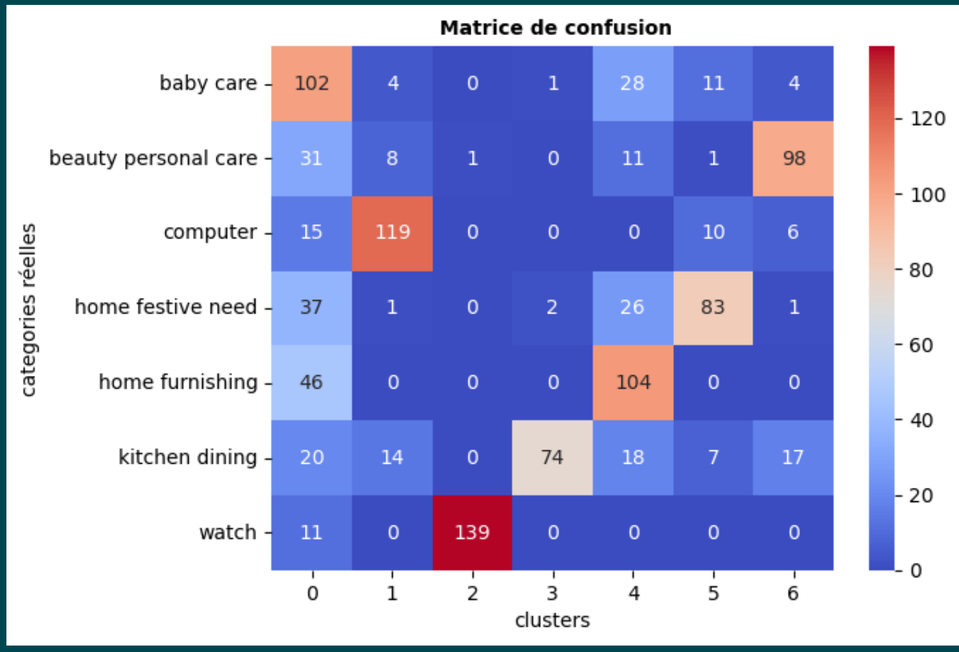
Faisabilité de classification par le texte: Word2Vec

Le « word2vec » capture efficacement les propriétés sémantiques et arithmétiques d'un mot.

ARI : 0.4113





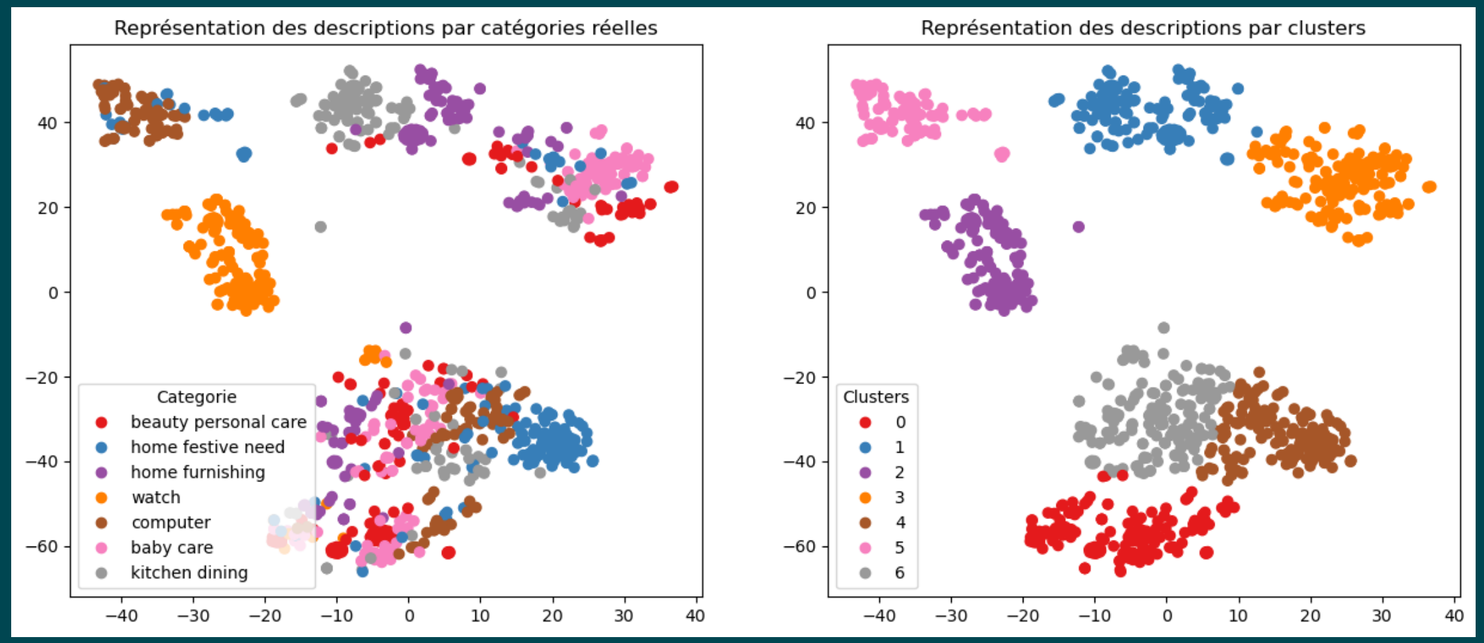


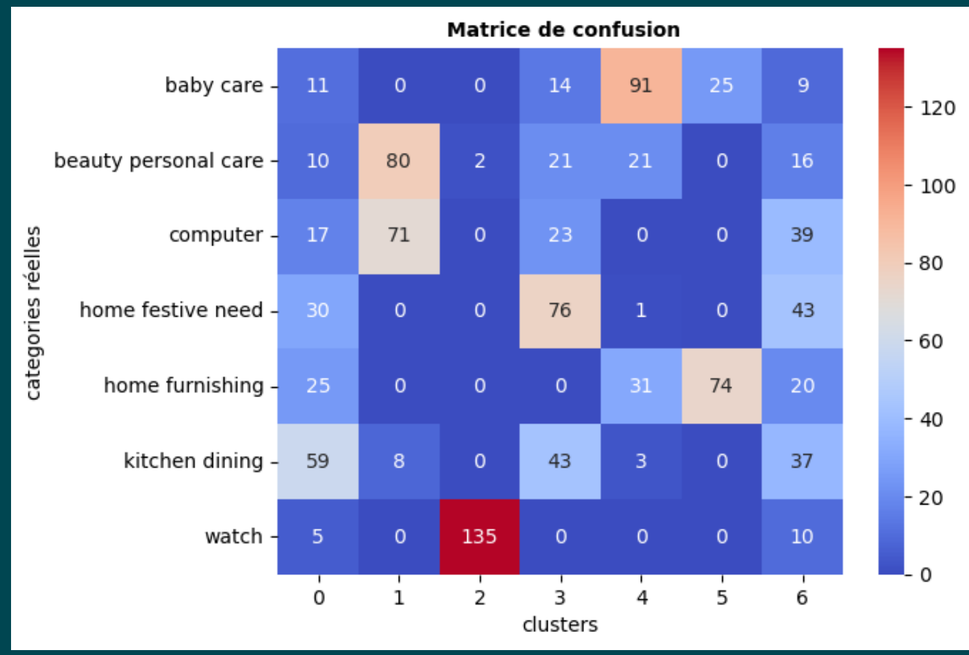
Faisabilité de classification par le texte:

BERT-Bidirectional Encoder Representations from Transformers

BERT masque aléatoirement des mots dans la phrase, puis il essaie de les prédire. Il utilise l'architecture des transformers, un encodeur pour lire le texte et un décodeur pour predire.

ARI : 0.3062



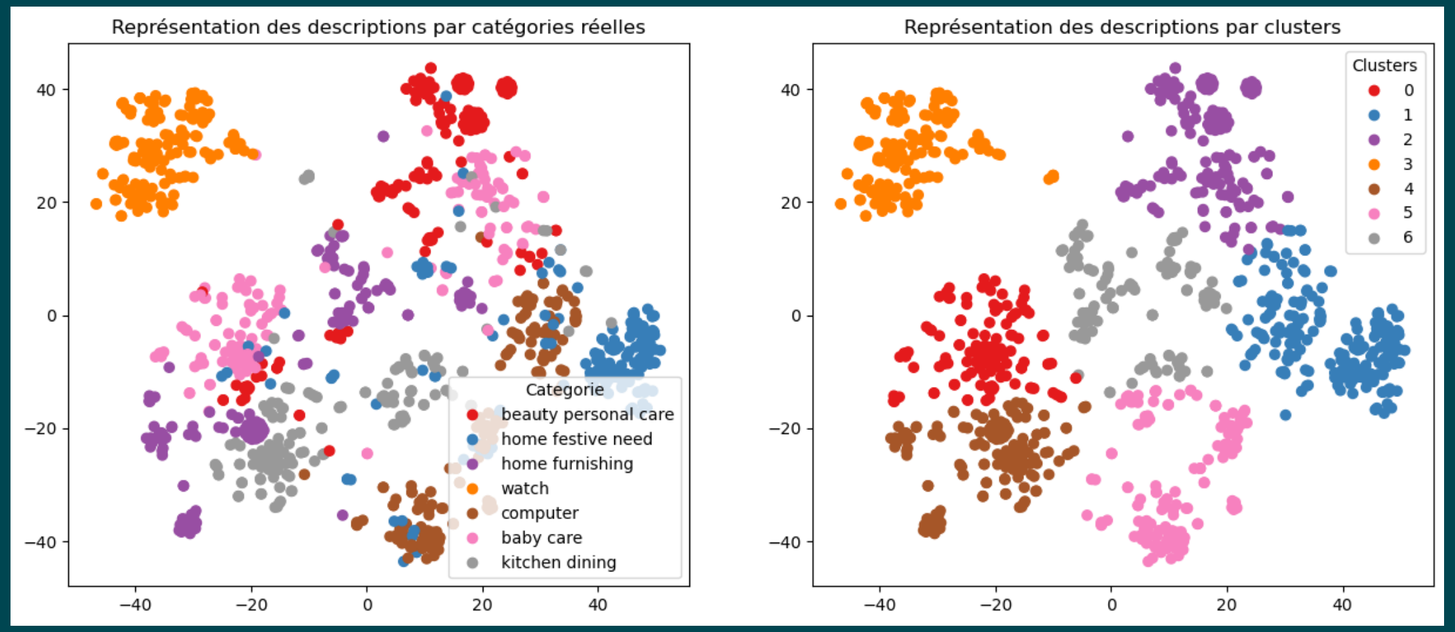


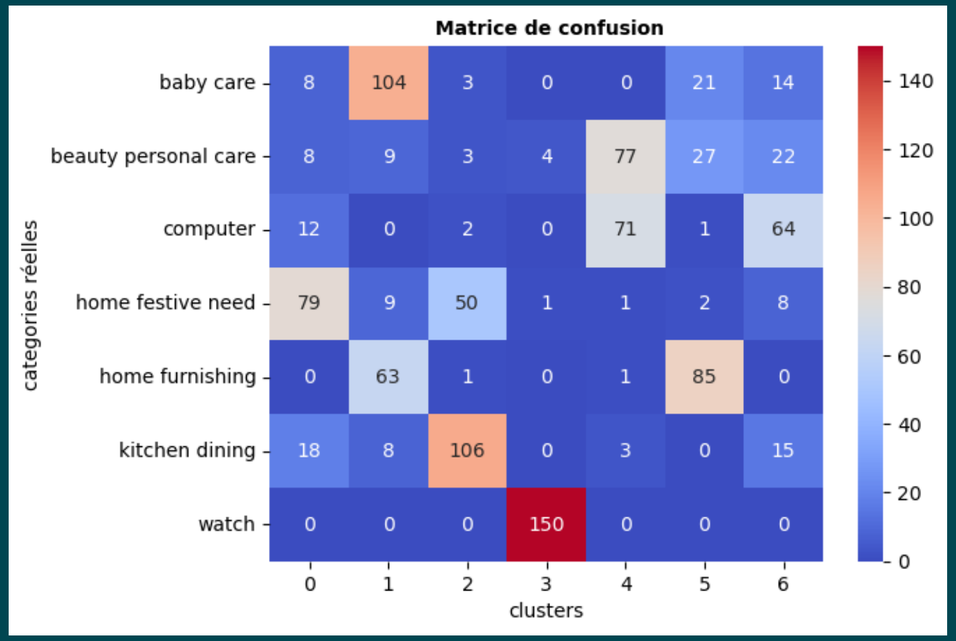
Faisabilité de classification par le texte:

USE - Universal Sentence Encoder

Les modèles USE effectuent une moyenne pondérée des mots contenus dans une phrase. La pondération est basé sur le concept d'Attention, permettant d'identifier l'importance des mots dans un contexte en fonction de leur position et de leur identité.

ARI : 0.4337





Faisabilité de classification automatique d’images

La démarche

SIFT
Scale Invariant FeatureTransform

- Prétraitement d'images(niveaux de gris, equalization, filtrage bruit)
- l’extraction des descripteurs SIFT
- Création des clusters de descripteurs
- Réduction de dimension(ACP, T-SNE)
- Création de clusters à partir du T-SNE
- ARI: similarité de la catégorisation (catégorie réelle / cluster k-means)

CNN
Convolutional Neural Network

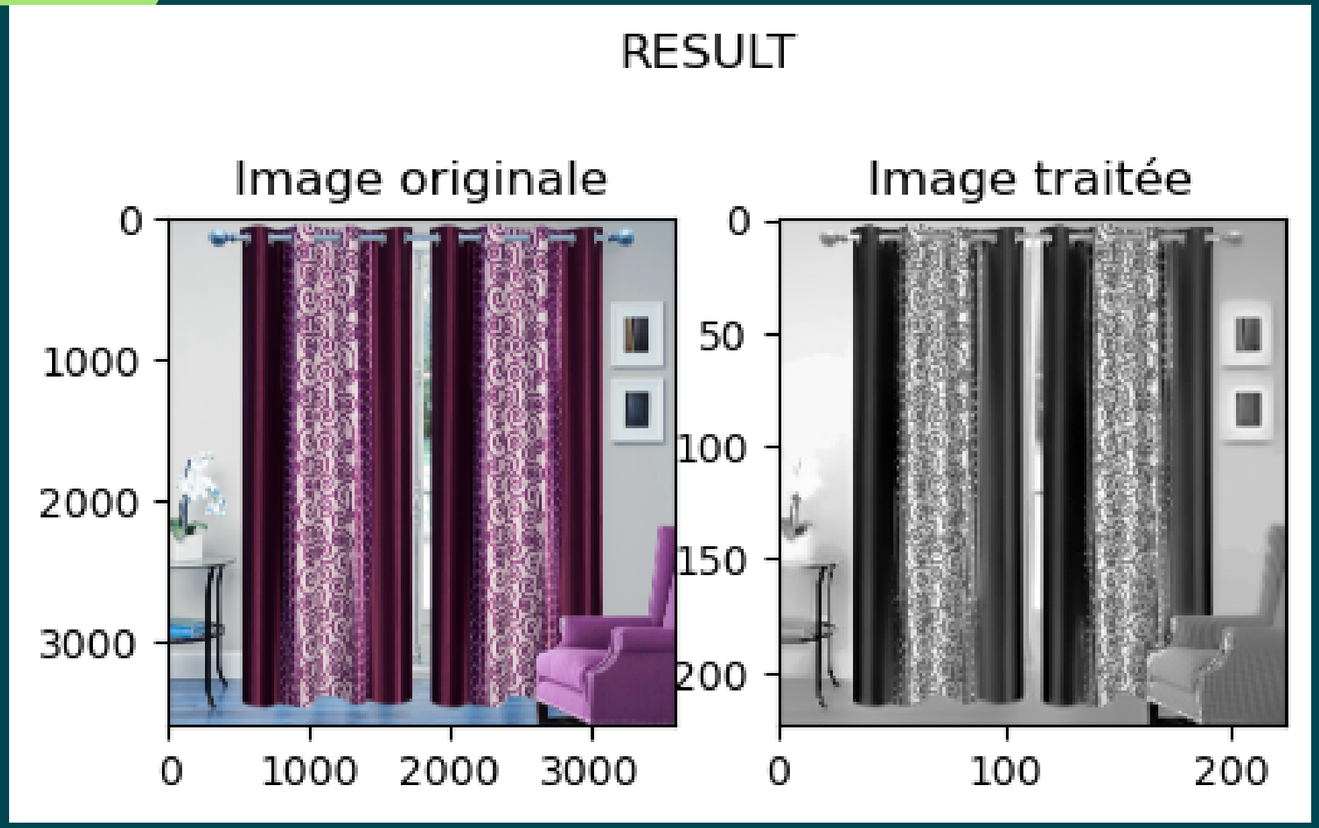
- Transfert learning model VGG-16
- Création des features des images
- Réduction de dimension(ACP, T-SNE)
- Création de clusters à partir du T-SNE
- ARI: similarité de la catégorisation (catégorie réelle / cluster k-means)

Faisabilité de classification automatique d’images

preprocessing

Traitements effectués:

- *Passage en gris*
- *Redimensionnement*
- *Élimination du bruit*
- *Égalisation d’histogrammes*

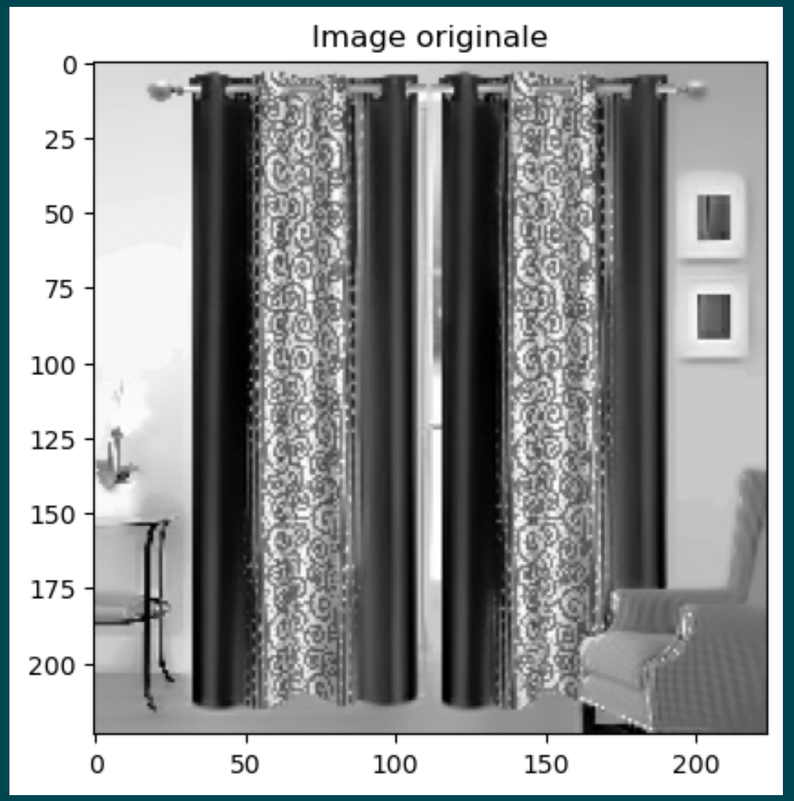


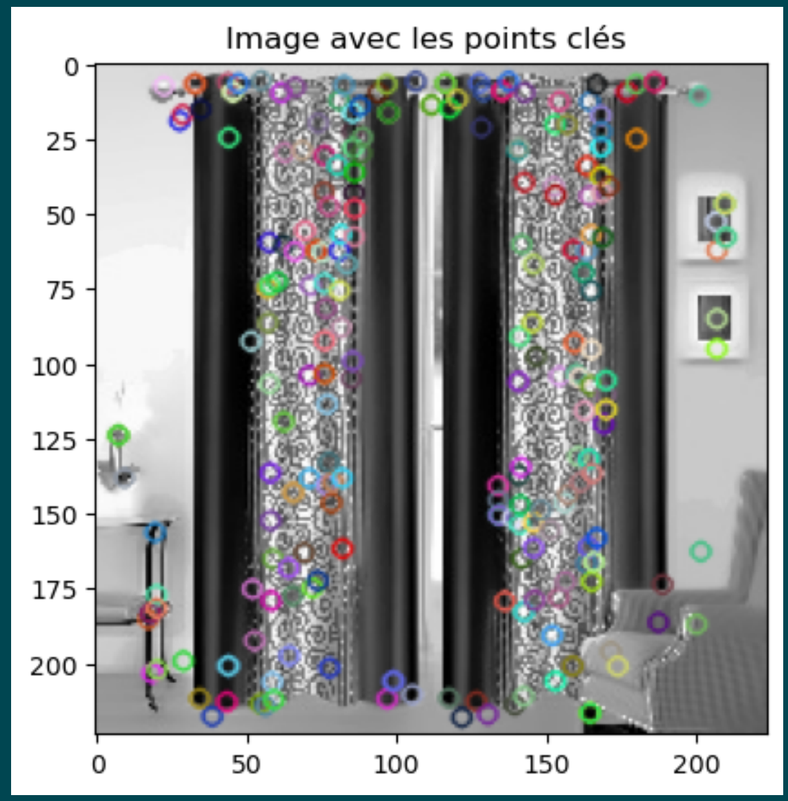
Faisabilité de classification automatique d’images

Description des features avec SIFT

L'algorithme SIFT se divise en plusieurs étapes:

- la localisation des points d'intérêt
- la création des descripteurs





Les vecteurs descripteurs, traduisent numériquement chacun des points-clés.

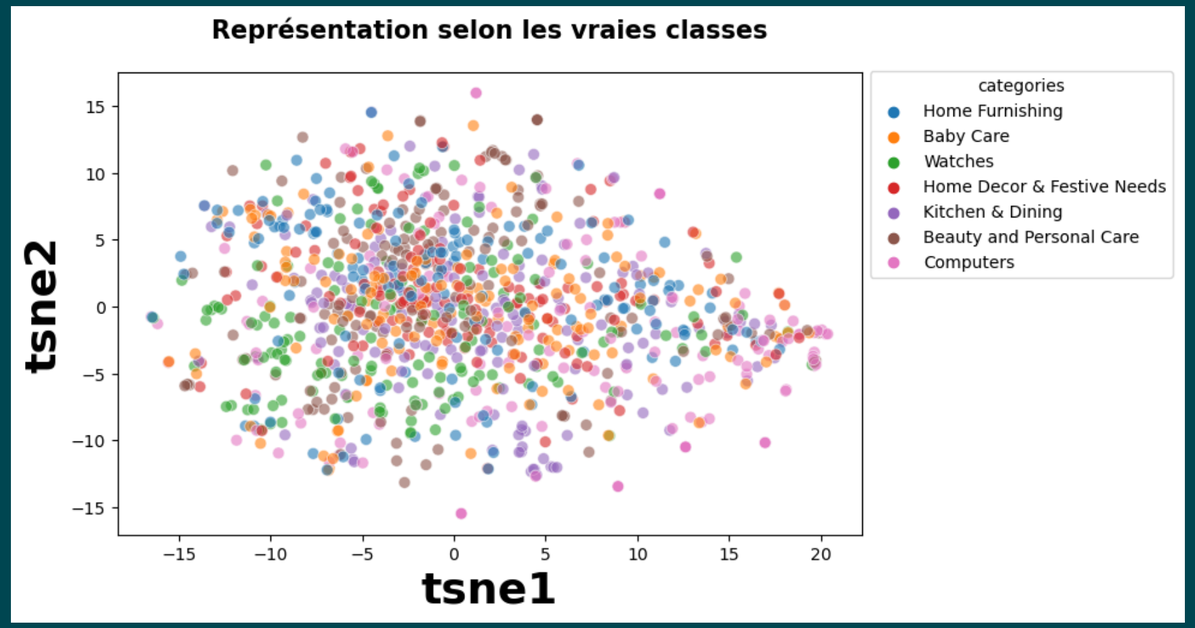
```
Descripteurs :  (246, 128)

[[ 0.  0.  0. ...  2.  3.  4.]
 [ 0.  0.  1. ...  0.  0.  0.]
 [ 0. 12. 15. ...  1.  1. 24.]
 ...
 [ 0.  2. 27. ...  0.  0.  3.]
 [ 0.  0.  0. ...  1.  0.  0.]
 [ 0.  5. 51. ...  0.  0.  2.]]
```

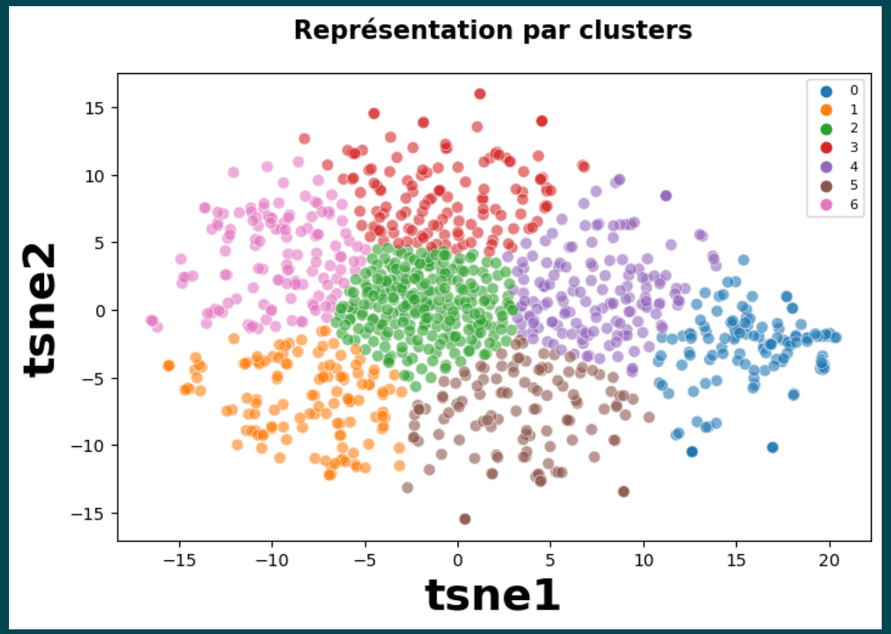
Faisabilité de classification automatique d'images

SIFT

```
Dimensions dataset avant réduction PCA :  (1050, 435)
Dimensions dataset après réduction PCA :  (1050, 366)
```

ARI : 0.032

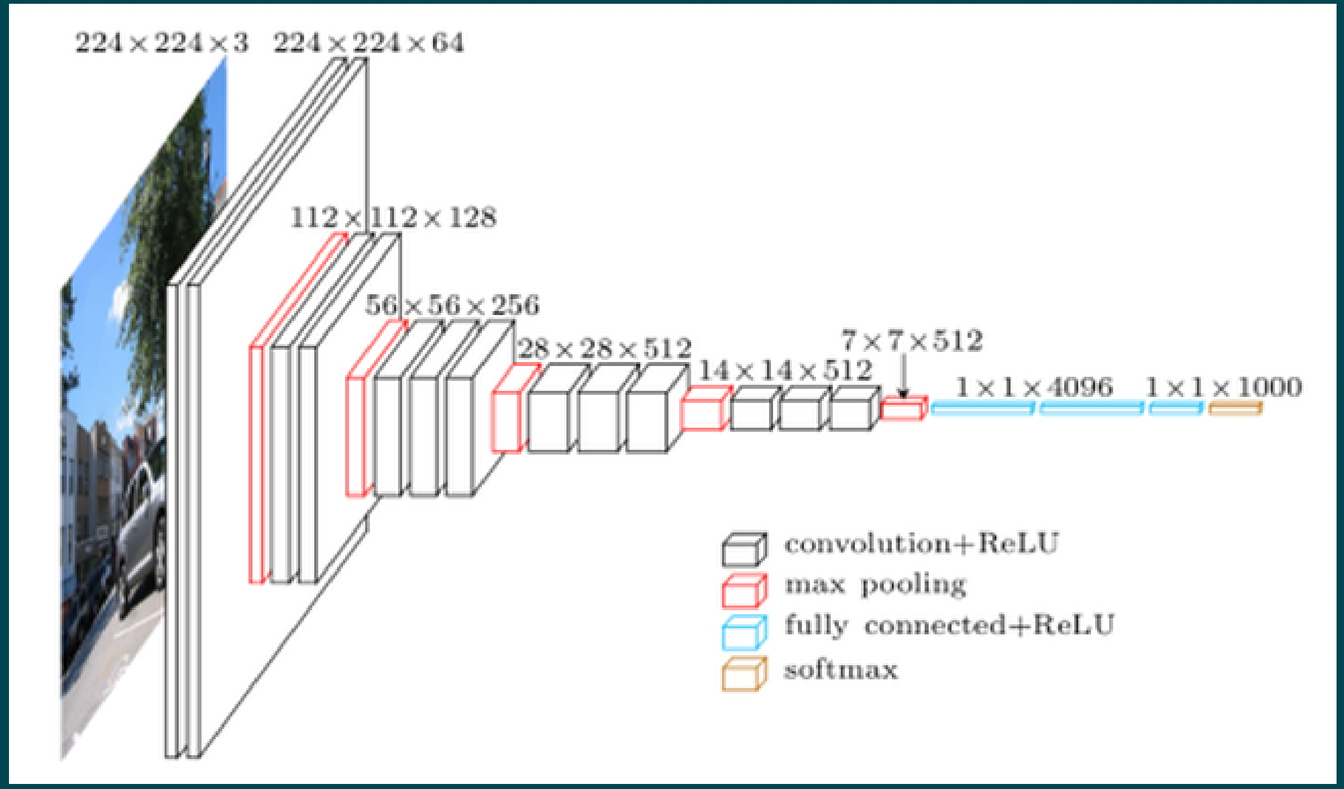




Faisabilité de classification automatique d’images

VGG-16

Représentation 3D de l'architecture de VGG-16



VGG-16:

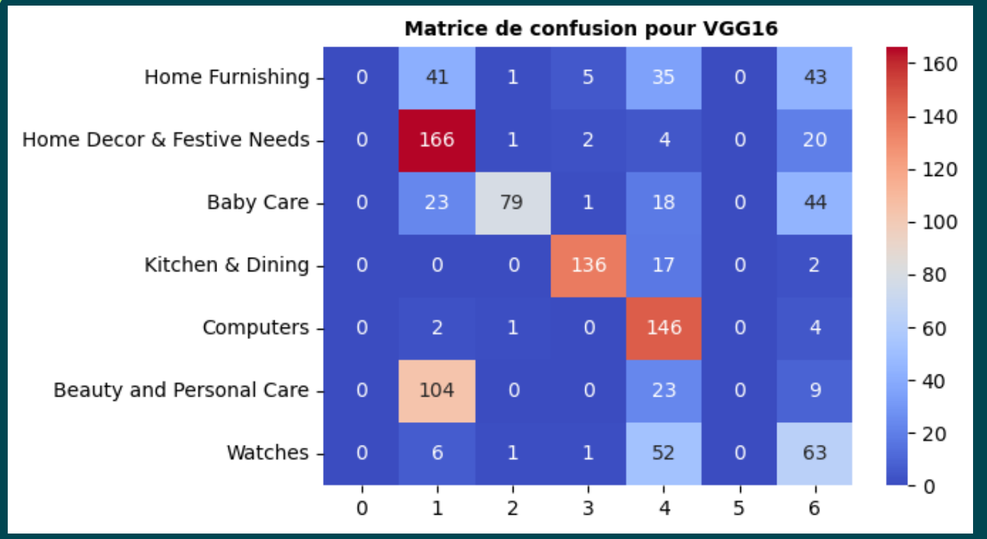
- 13 couches de convolution
- 3 fully-connected
- Image en couleurs 224x224 px en entrée
- vecteur de taille 1000 en sortie

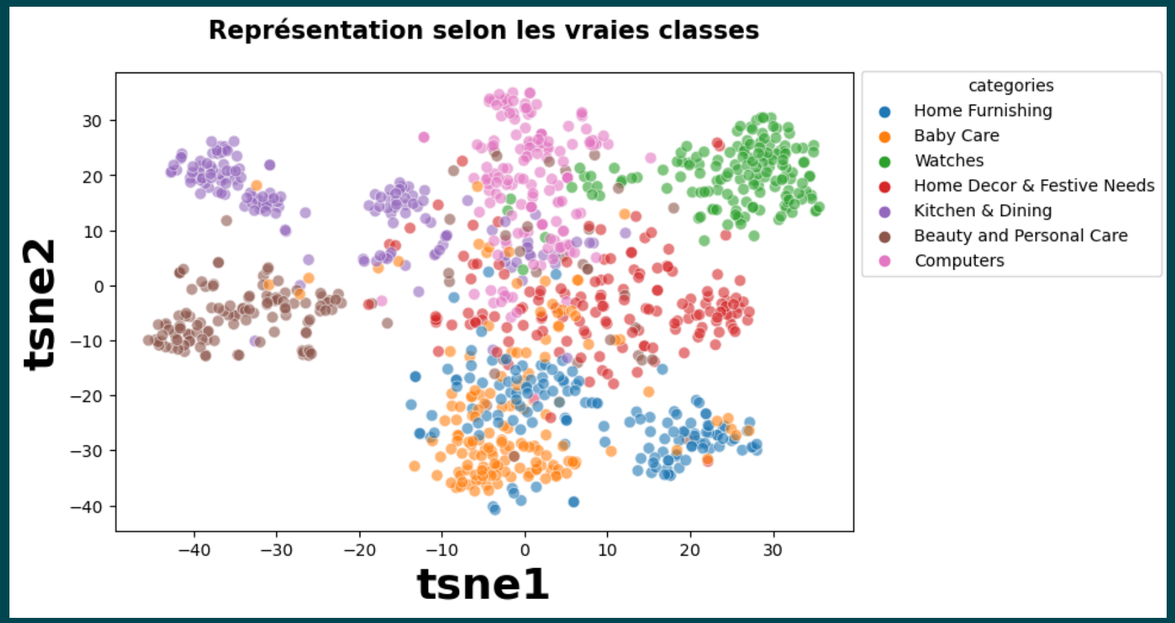
Faisabilité de classification automatique d'images

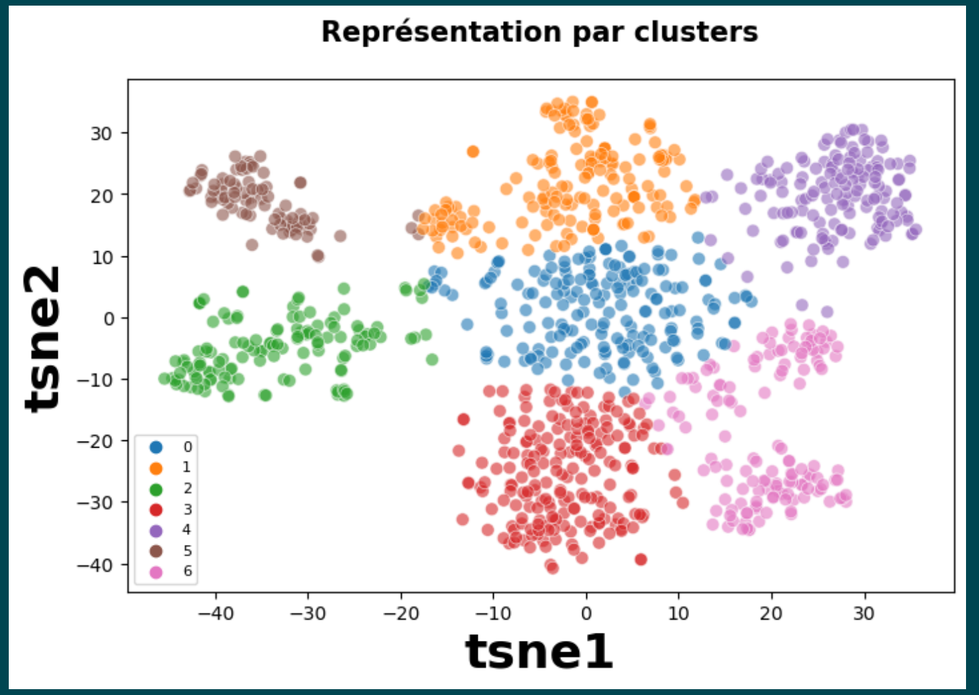
VGG-16

```
Dimensions dataset avant réduction PCA :  (1050, 4096)
Dimensions dataset après réduction PCA :  (1050, 803)
```

ARI : 0.4555







Conclusion

la faisabilité du moteur de classification

Sur certains modèles, la valeur, de l’ordre de 0.4 à 0.5, des ARI confirme la faisabilité de classer automatiquement les produits.

Le Tf-idf a donné des meilleurs résultats que les autres approches pour la classification par la description.

Le VGG-16 a donné de meilleurs résultats que le SIFT pour la classification par l'image.

Recommandations

l'utilisation d’un jeu de données de plus grande taille

La combinaison des deux types d'approches: descriptions et images

MERCI