Projet 6 : Classifiez automatiquement des biens de consommation

Abdessalem BOUZAYANI

03/03/2023

Sommaire

- Problématique
- Jeu de données
- Traitement données textuelles
- Traitement données images
- Conclusions et perspectives

Problématique

Mission :

- Étudier la faisabilité d'un moteur de classification des articles en différentes catégories, avec un niveau de précision suffisant.

Objectifs :

- Améliorer l’expérience des utilisateurs
- Fiabiliser la catégorie des articles avec pertinence et précision

Jeu de données

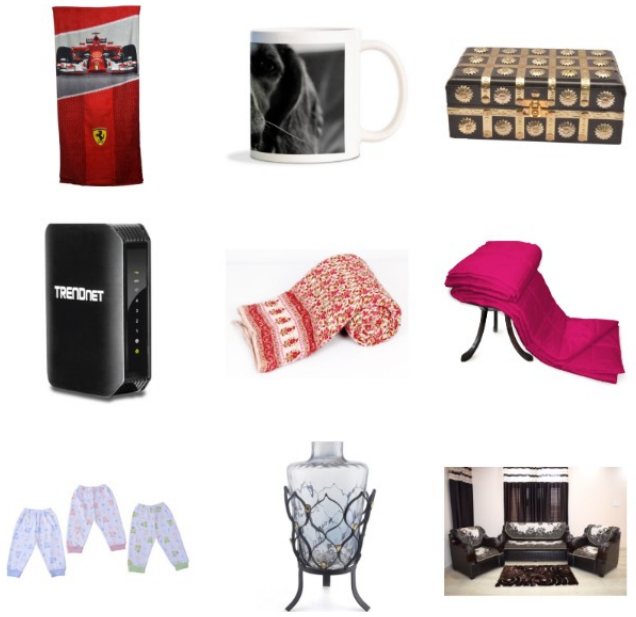
15 colonnes (caractéristiques)

1050 lignes (produits)

2% de valeurs manquantes
12 strings, 2 floattants, 1 booléen

7 catégories principales

categ_level_1	image
Baby Care	150
Beauty and Personal Care	150
Computers	150
Home Decor & Festive Needs	150
Home Furnishing	150
Kitchen & Dining	150
Watches	150



Traitement données textuelles

PRÉ TRAITEMENT

Tokenisation

Normalisation

Stopwords

Lemmatisation

FEATURES EXTRACTION

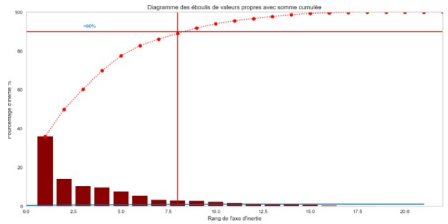
BagsOf Words :
CountVectorizer
TfidfVectorizer

Word embeddings :
Word2Vec
Bert

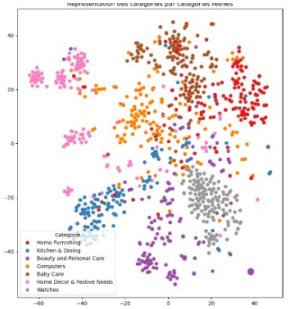
USE

RÉDUCTION DIMENSION

ACP

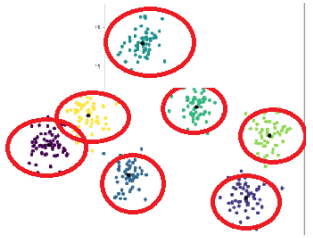


TSNE



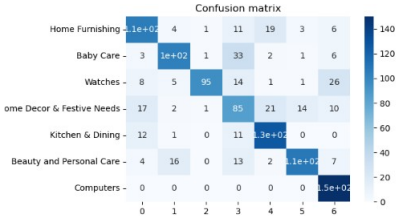
CLUSTERING

Kmeans
7 clusters



ÉVALUATION

ARI
Homogeneity
Completeness
v_measure



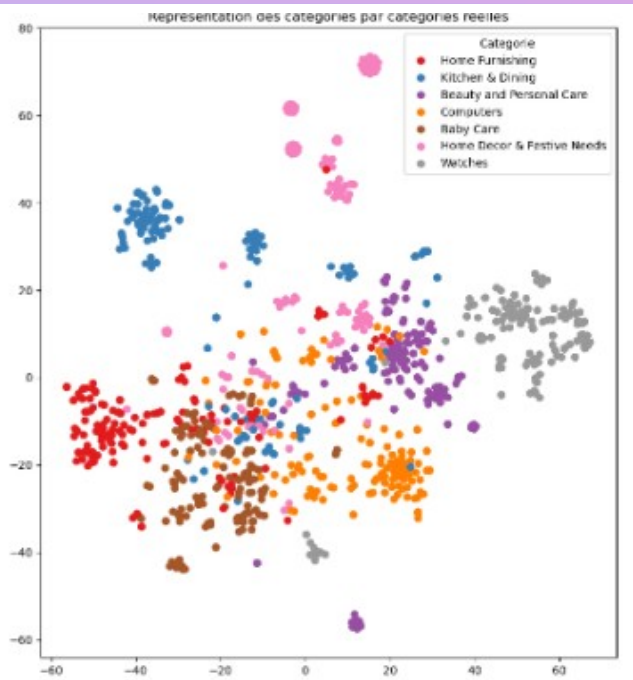
NLP - Prétraitement

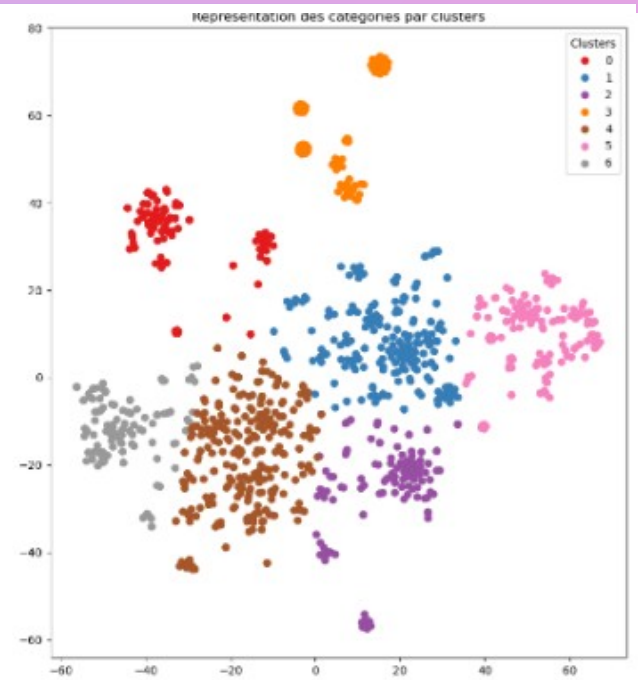
Texte original	Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 This curtain enhances the look of the interiors.This curtain is made from 100% high quality polyester fabric.It features an eyelet style stitch with Metal Ring.It makes the room environment romantic and loving.This curtain is ant- wrinkle and anti shrinkage and have elegant apparance.Give your home a bright and modernistic appeal with these designs. The surreal attention is sure to steal hearts. These contemporary eyelet and valance curtains slide smoothly so when you draw them apart first thing in the morning to welcome the bright sun rays you want to wish good morning to the whole world and when you draw them close in the evening, you create the most special moments of joyous beauty given by the soothing prints. Bring home the elegant curtain that softly filters light in your room so that you get the right amount of sunlight.,Specifications of Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) General Brand Elegance Designed For Door Type Eyelet Model Name Abstract Polyester Door Curtain Set Of 2 Model ID Duster25 Color Multicolor Dimensions Length 213 cm In the Box Number of Contents in Sales Package Pack of 2 Sales Package 2 Curtains Body & Design Material Polyester
Chiffres, ponctuation	Key Features Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain Elegance Polyester Multicolor Abstract Eyelet Door Curtain Height Pack Price This curtain enhances the look the interiors.This curtain made from high quality polyester fabric features eyelet style stitch with Metal Ring makes the room environment romantic and loving This curtain ant wrinkle and anti shrinkage and have elegant apparance.Give your home bright and modernistic appeal with these designs The surreal attention sure steal hearts These contemporary eyelet and valance curtains slide smoothly when you draw them apart first thing the morning welcome the bright sun rays you want wish good morning the whole world and when you draw them close the evening you create the most special moments joyous beauty given the soothing prints Bring home the elegant curtain that softly filters light your room that you get the right amount sunlight Specifications Elegance Polyester Multicolor Abstract Eyelet Door Curtain Height Pack General Brand Elegance Designed For Door Type Eyelet Model Name Abstract Polyester Door Curtain Set Model Duster Color Multicolor Dimensions Length the Box Number Contents Sales Package Pack Sales Package Curtains Body Design Material Polyester
Extraction des tokens	['Key', 'Features', 'Elegance', 'Polyester', 'Multicolor', 'Abstract', 'Eyelet', 'Door', 'Curtain', 'Floral', 'Curtain', 'Elegance', 'Polyester', 'Multicolor', 'Abstract', 'Eyelet', 'Door', 'Curtain', 'Height', 'Pack', 'Price', 'This', 'curtain', 'enhances', 'the', 'look', 'the', 'interiors', 'This', 'curtain', 'made', 'from', 'high', 'quality', 'polyester', 'fabric', 'features', 'eyelet', 'style', 'stitch', 'with', 'Metal', 'Ring', 'makes', 'the', 'room', 'environment', 'romantic', 'and', 'loving', 'This', 'curtain', 'ant', 'wrinkle', 'and', 'anti', 'shrinkage', 'and', 'have', 'elegant', 'apparance', 'Give', 'your', 'home', 'bright', 'and', 'modernistic', 'appeal', 'with', 'these', 'designs', 'The', 'surreal', 'attention', 'sure', 'steal', 'hearts', 'These', 'contemporary', 'eyelet', 'and', 'valance', 'curtains', 'slide', 'smoothly', 'when', 'you', 'draw', 'them', 'apart', 'first', 'thing', 'the', 'morning', 'welcome', 'the', 'bright', 'sun', 'rays', 'you', 'want', 'wish', 'good', 'morning', 'the', 'whole', 'world', 'and', 'when', 'you', 'draw', 'them', 'close', 'the', 'evening', 'you', 'create', 'the', 'most', 'special', 'moments', 'joyous', 'beauty', 'given', 'the', 'soothing', 'prints', 'Bring', 'home', 'the', 'elegant', 'curtain', 'that', 'softly', 'filters', 'light', 'your', 'room', 'that', 'you', 'get', 'the', 'right', 'amount', 'sunlight', 'Specifications', 'Elegance', 'Polyester', 'Multicolor', 'Abstract', 'Eyelet', 'Door', 'Curtain', 'Height', 'Pack', 'General', 'Brand', 'Elegance', 'Designed', 'For', 'Door', 'Type', 'Eyelet', 'Model', 'Name', 'Abstract', 'Polyester', 'Door', 'Curtain', 'Set', 'Model', 'Duster', 'Color', 'Multicolor', 'Dimensions', 'Length', 'the', 'Box', 'Number', 'Contents', 'Sales', 'Package', 'Pack', 'Sales', 'Package', 'Curtains', 'Body', 'Design', 'Material', 'Polyester']
Stop word	['key', 'features', 'elegance', 'polyester', 'multicolor', 'abstract', 'eyelet', 'door', 'curtain', 'floral', 'curtain', 'elegance', 'polyester', 'multicolor', 'abstract', 'eyelet', 'door', 'curtain', 'height', 'pack', 'price', 'curtain', 'enhances', 'look', 'interiors', 'curtain', 'made', 'high', 'quality', 'polyester', 'fabric', 'features', 'eyelet', 'style', 'stitch', 'metal', 'ring', 'makes', 'room', 'environment', 'romantic', 'loving', 'curtain', 'ant', 'wrinkle', 'anti', 'shrinkage', 'elegant', 'apparance', 'give', 'home', 'bright', 'modernistic', 'appeal', 'designs', 'surreal', 'attention', 'sure', 'steal', 'hearts', 'contemporary', 'eyelet', 'valance', 'curtains', 'slide', 'smoothly', 'draw', 'apart', 'first', 'thing', 'morning', 'welcome', 'bright', 'sun', 'rays', 'want', 'wish', 'good', 'morning', 'whole', 'world', 'draw', 'close', 'evening', 'create', 'special', 'moments', 'joyous', 'beauty', 'given', 'soothing', 'prints', 'bring', 'home', 'elegant', 'curtain', 'softly', 'filters', 'light', 'room', 'get', 'right', 'amount', 'sunlight', 'specifications', 'elegance', 'polyester', 'multicolor', 'abstract', 'eyelet', 'door', 'curtain', 'height', 'pack', 'general', 'brand', 'elegance', 'designed', 'door', 'type', 'eyelet', 'model', 'name', 'abstract', 'polyester', 'door', 'curtain', 'set', 'model', 'duster', 'color', 'multicolor', 'dimensions', 'length', 'box', 'number', 'contents', 'sales', 'package', 'pack', 'sales', 'package', 'curtains', 'body', 'design', 'material', 'polyester']
lemmatisation	['key', 'feature', 'elegance', 'polyester', 'multicolor', 'abstract', 'eyelet', 'door', 'curtain', 'floral', 'curtain', 'elegance', 'polyester', 'multicolor', 'abstract', 'eyelet', 'door', 'curtain', 'height', 'pack', 'price', 'curtain', 'enhances', 'look', 'interior', 'curtain', 'made', 'high', 'quality', 'polyester', 'fabric', 'feature', 'eyelet', 'style', 'stitch', 'metal', 'ring', 'make', 'room', 'environment', 'romantic', 'loving', 'curtain', 'ant', 'wrinkle', 'anti', 'shrinkage', 'elegant', 'apparance', 'give', 'home', 'bright', 'modernistic', 'appeal', 'design', 'surreal', 'attention', 'sure', 'steal', 'heart', 'contemporary', 'eyelet', 'valance', 'curtain', 'slide', 'smoothly', 'draw', 'apart', 'first', 'thing', 'morning', 'welcome', 'bright', 'sun', 'ray', 'want', 'wish', 'good', 'morning', 'whole', 'world', 'draw', 'close', 'evening', 'create', 'special', 'moment', 'joyous', 'beauty', 'given', 'soothing', 'print', 'bring', 'home', 'elegant', 'curtain', 'softly', 'filter', 'light', 'room', 'get', 'right', 'amount', 'sunlight', 'specification', 'elegance', 'polyester', 'multicolor', 'abstract', 'eyelet', 'door', 'curtain', 'height', 'pack', 'general', 'brand', 'elegance', 'designed', 'door', 'type', 'eyelet', 'model', 'name', 'abstract', 'polyester', 'door', 'curtain', 'set', 'model', 'duster', 'color', 'multicolor', 'dimension', 'length', 'box', 'number', 'content', 'sale', 'package', 'pack', 'sale', 'package', 'curtain', 'body', 'design', 'material', 'polyester']

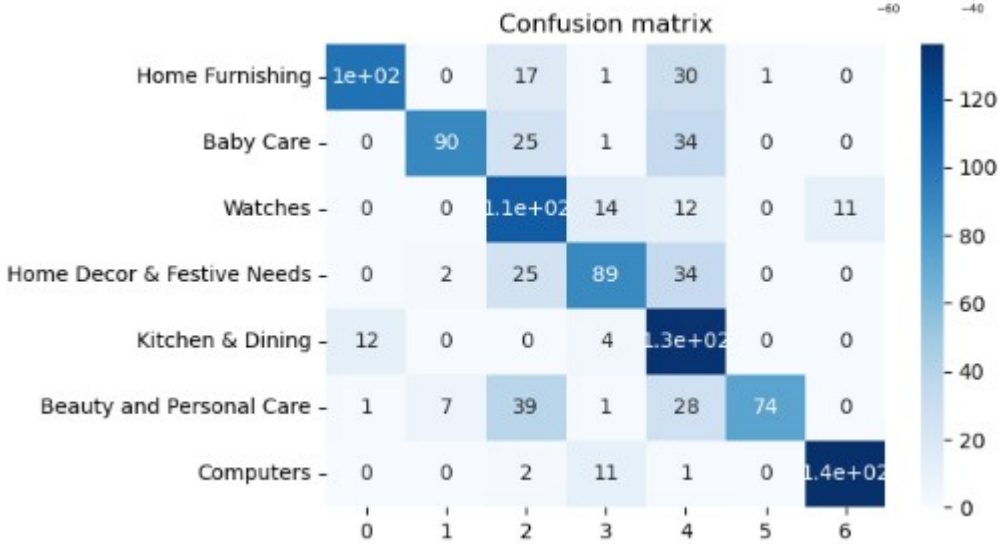
NLP - Bags of words - CountVectorizer

Transformer un texte donné en un vecteur sur la base de la fréquence de chaque mot qui apparaît dans l'ensemble du texte

(1050, 5324)





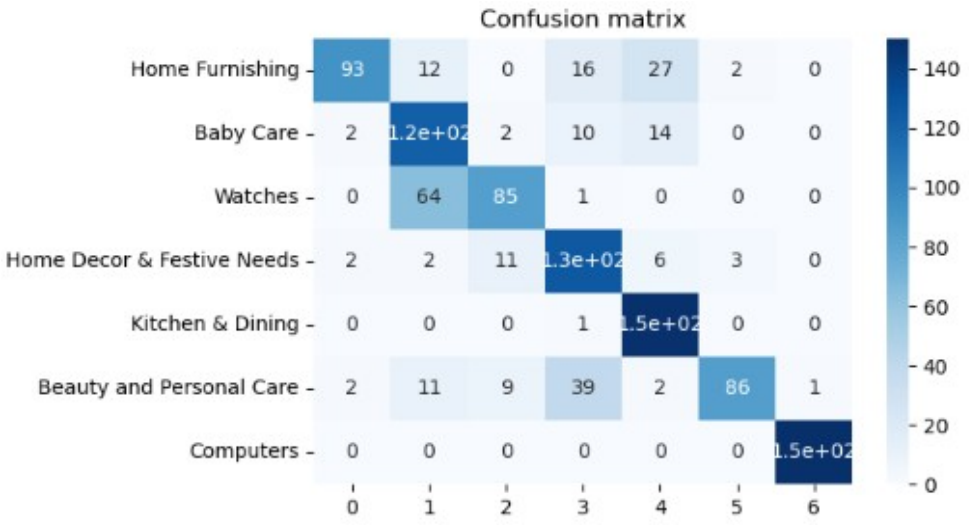


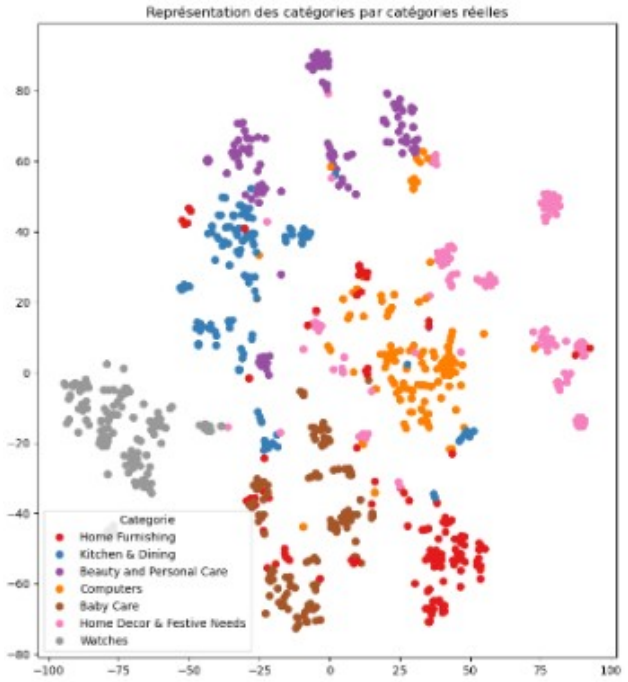
Methode	ARI	homogeneity	completeness	v_measure
countvect	0.4214	0.5347	0.5809	0.5475

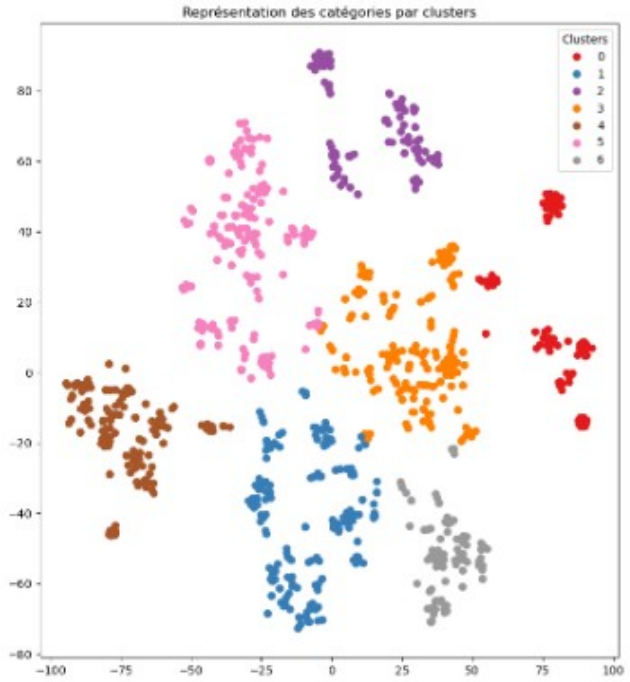
NLP - Bags of words - Tf-idf

L'importance d'un mot est inversement proportionnelle à sa fréquence dans les documents.
TF : fréquence d'un mot dans un document
IDF : rareté relative d'un mot dans le corpus

(1050, 5324)







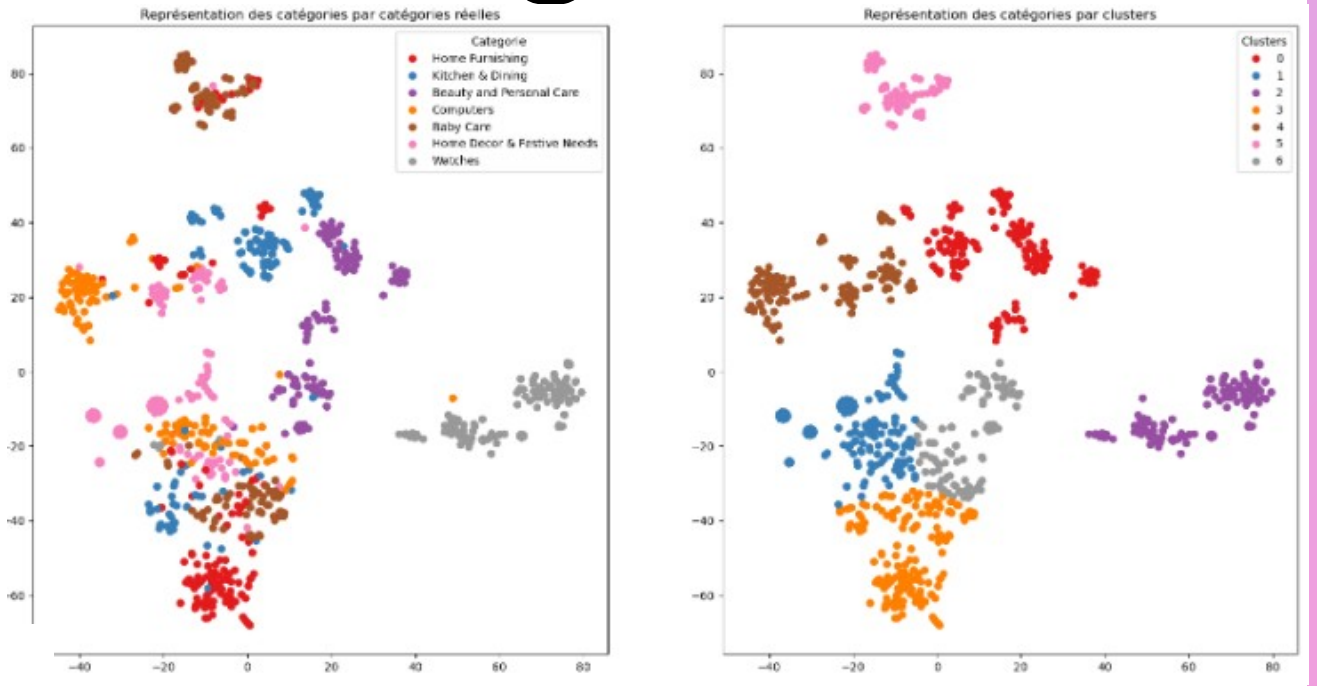
Methode	ARI	homogeneity	completeness	v_measure
tfidf	0.5788	0.8517	0.8693	0.8604

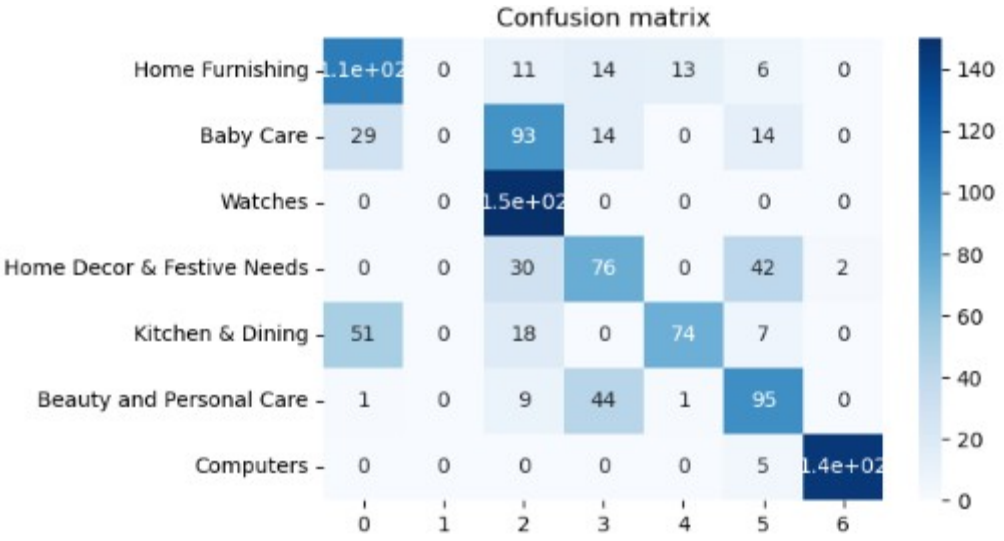
NLP - Word Embedding - w2vec

Générer une représentation vectorielle distribuée de longueur fixe

Utilisation d'un modèle pré entraîné:
word2vec-google-news-300

(1050, 300)



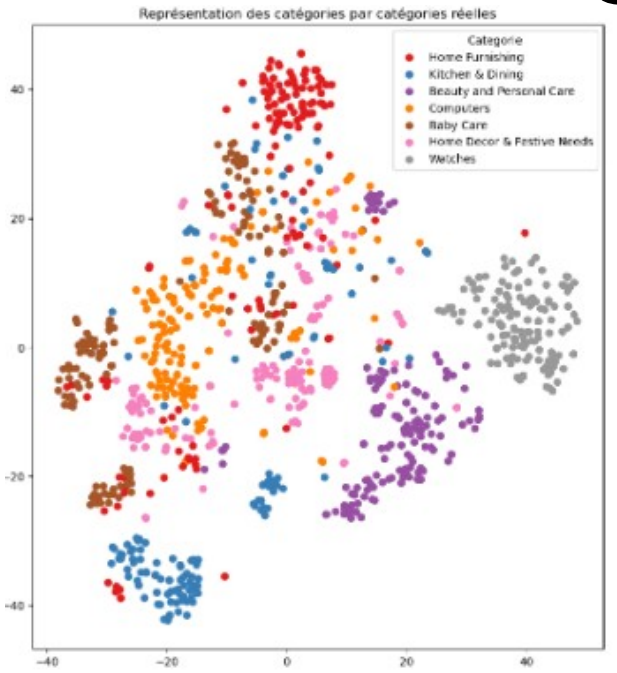


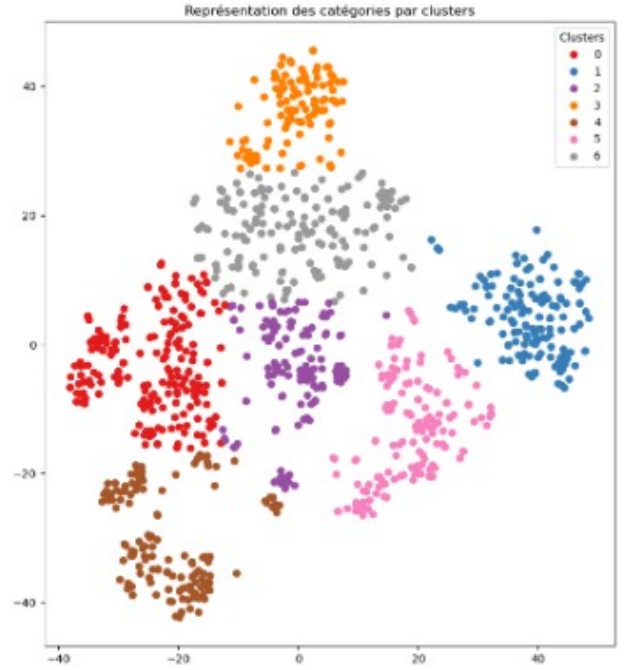
Methode	ARI	homogeneity	completeness	v_measure
word2vec	0.4165	0.5333	0.5409	0.5371

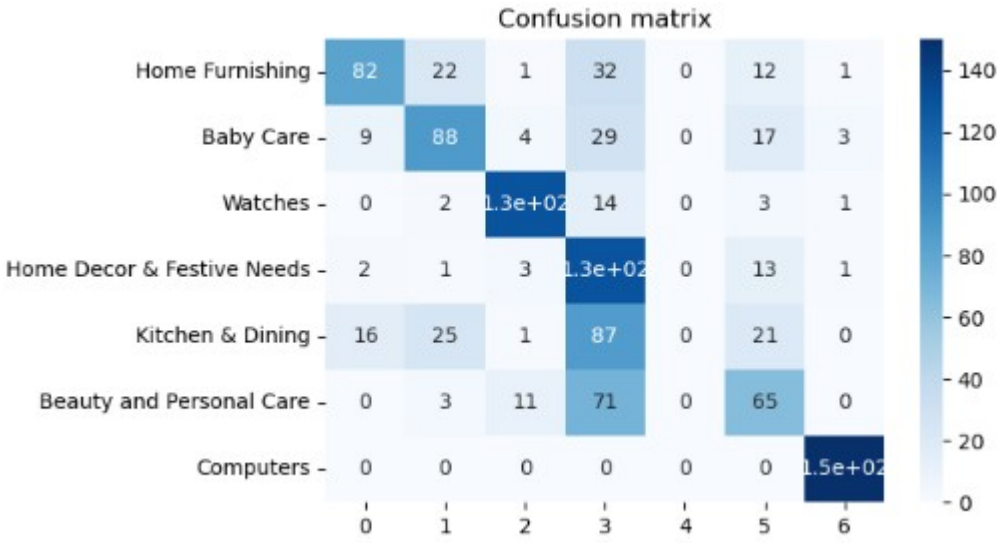
NLP - Word Embedding - Bert

Appliquer l'entraînement bidirectionnel d'un Transformer
Utilisation d'un modèle pré entraîné:
bert-base-uncased

(1050, 768)





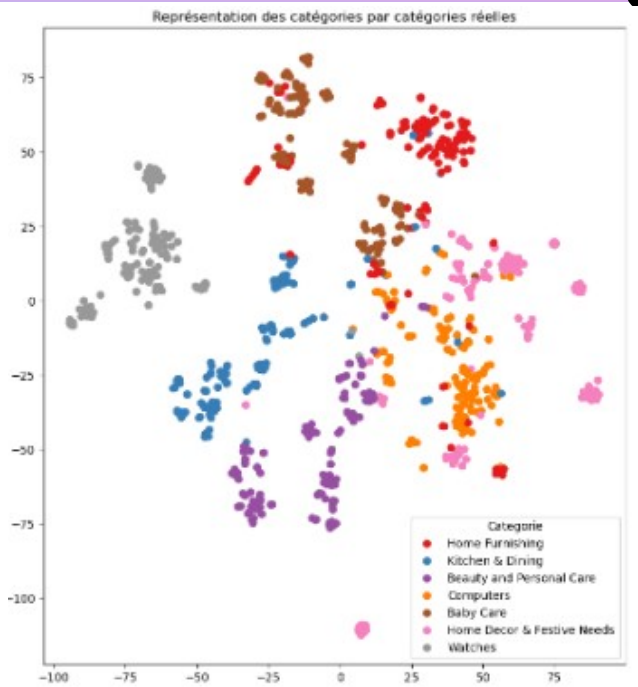


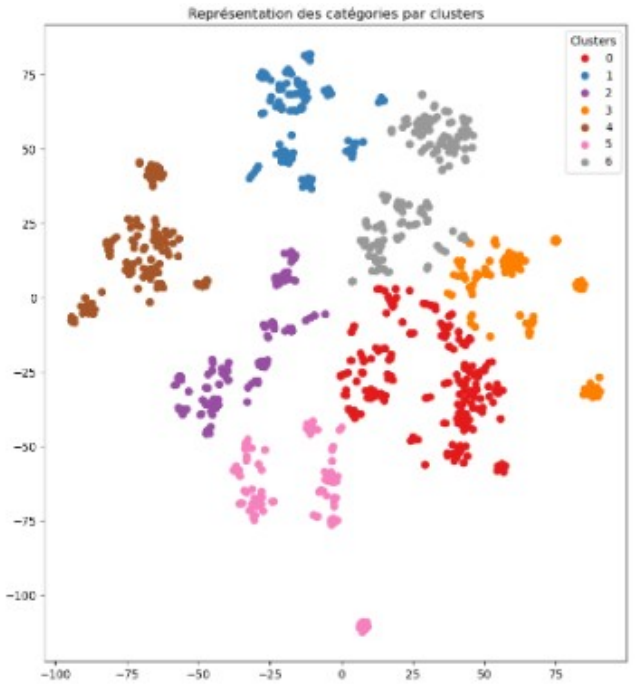
Methode	ARI	homogeneity	completeness	v_measure
bert	0.3888	0.4692	0.4727	0.4709

NLP - Word Embedding - USE

Le modèle calcule d'abord les word embeddings, puis détermine le vecteur de la phrase en calculant la somme par élément des vecteurs de mots.

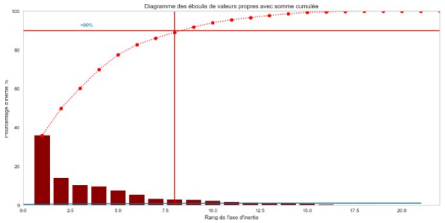
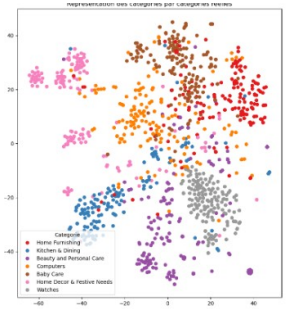
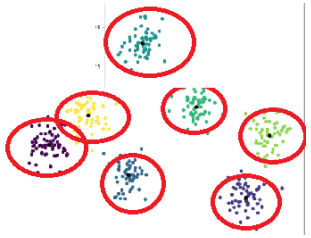
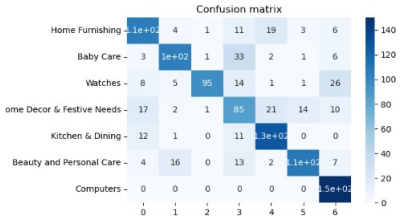
(1050, 512)





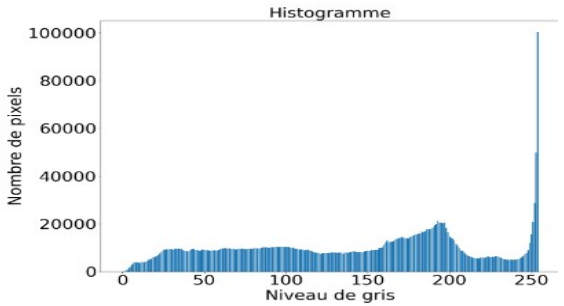
Methode	ARI	homogeneity	completeness	v_measure
use	0.5802	0.6729	0.6840	0.6784

Traitement données visuelles

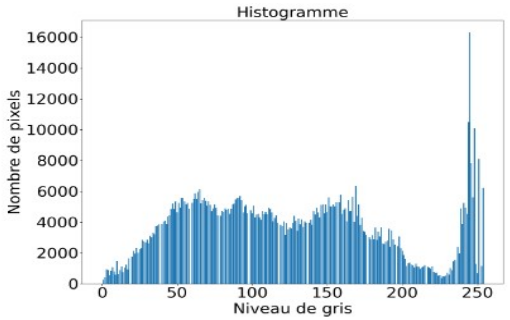
PRÉ TRAITEMENT	FEATURES EXTRACTION	RÉDUCTION DIMENSION	CLUSTERING	ÉVALUATION
Exposition Contraste niveaux de gris Bruit Redimensionnement	BOVW : SIFT ORB Transfert Learning : VGG16 DenseNet201 InceptionResNetV2 InceptionV3 MobileNetV2	ACP  TSNE 	Kmeans 7 clusters 	ARI Homogeneity Completeness v_measure 

CV - Pré-traitement

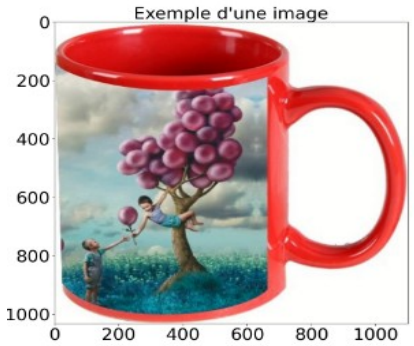
Image originale

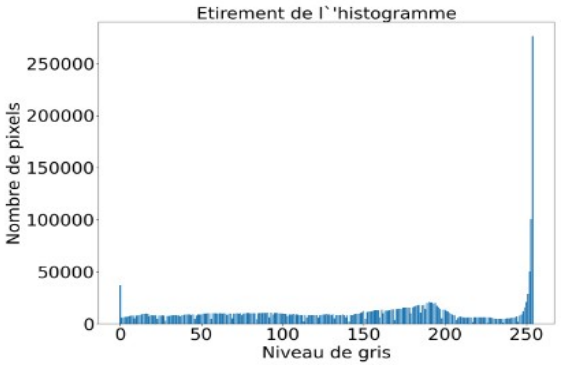


Niveaux de gris

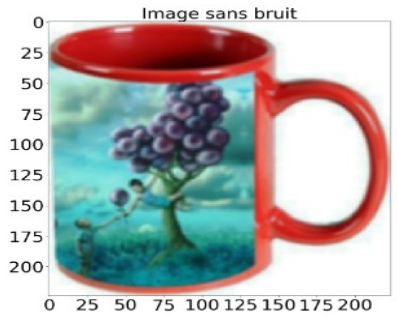


Étirement de l'histogramme

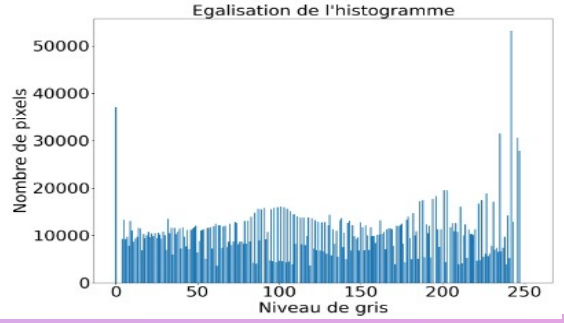




Réduction de bruit

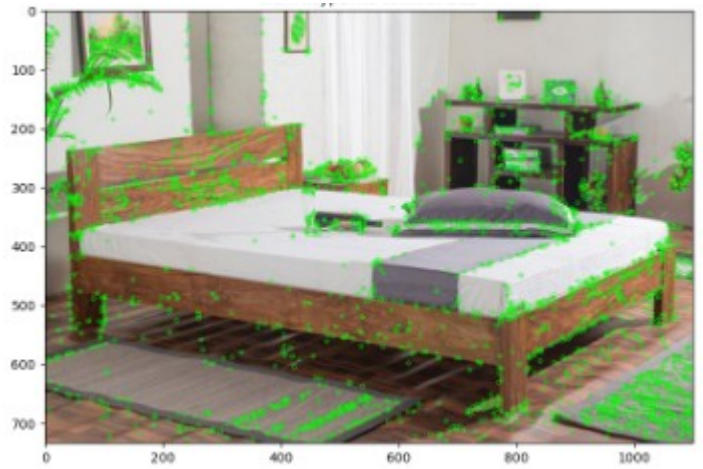


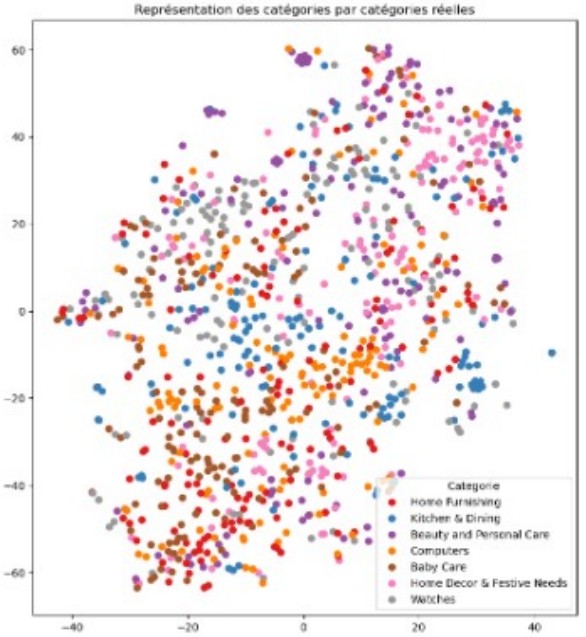
Égalisation de l'histogramme

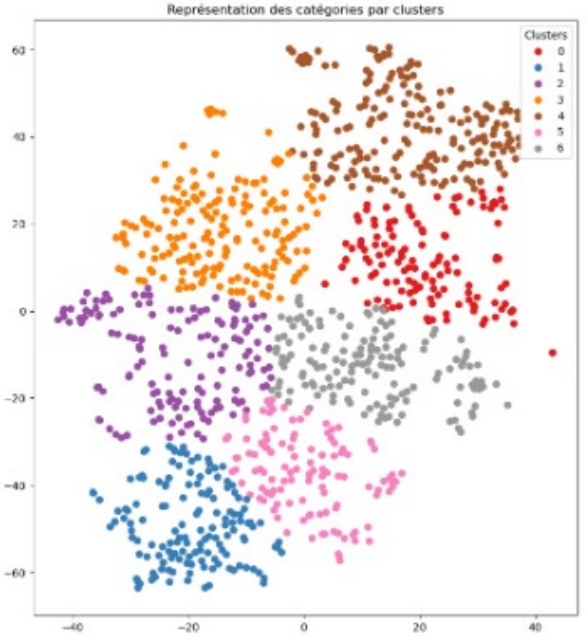


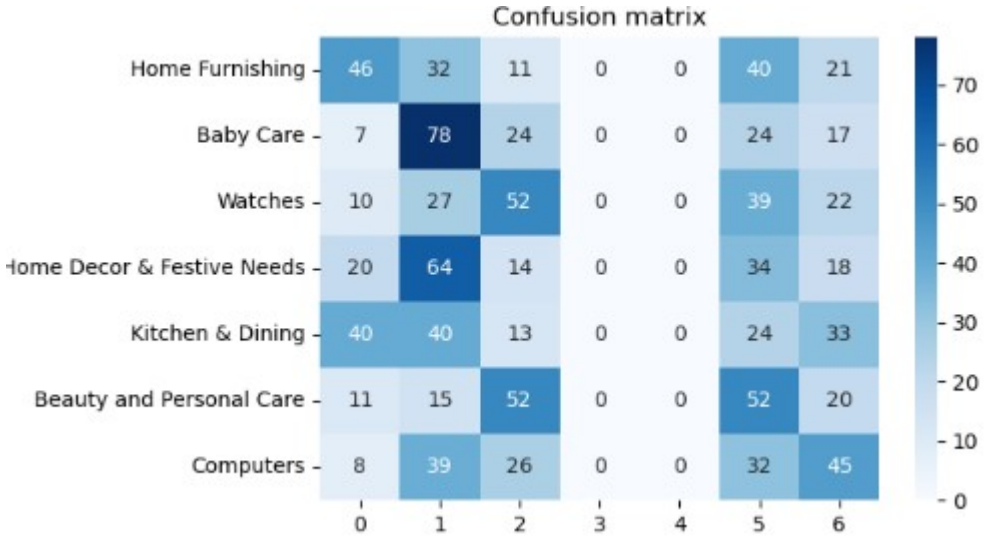
Redimentionnement

CV - SIFT



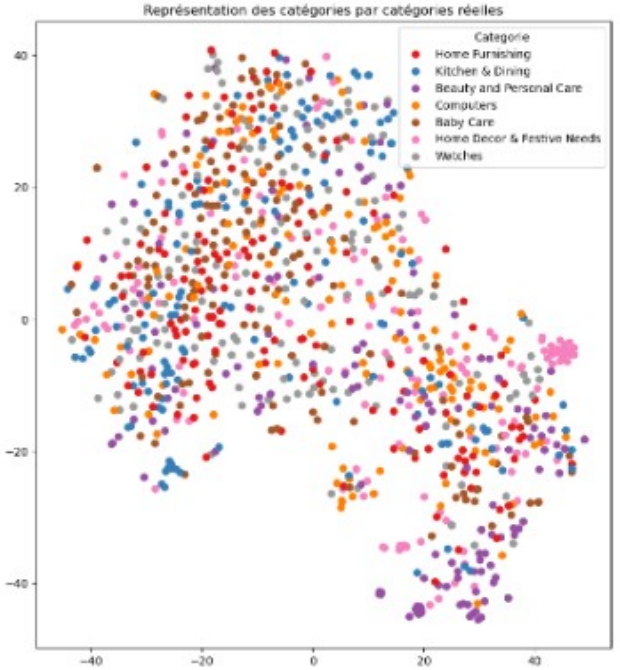


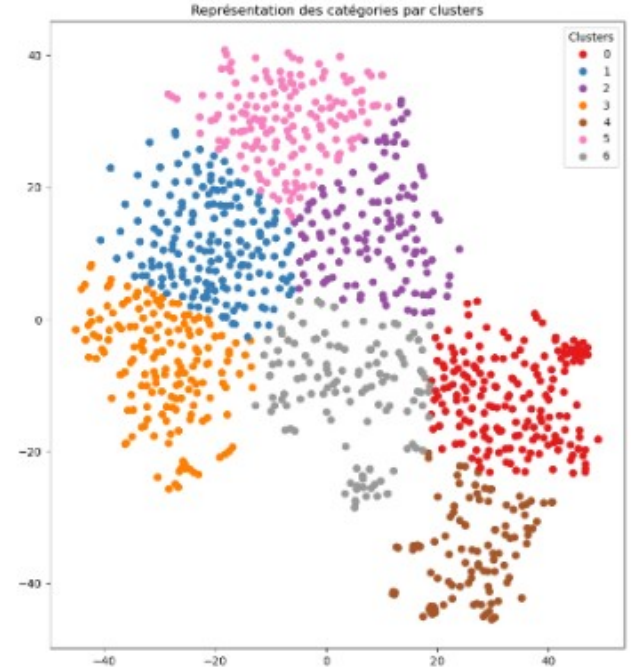


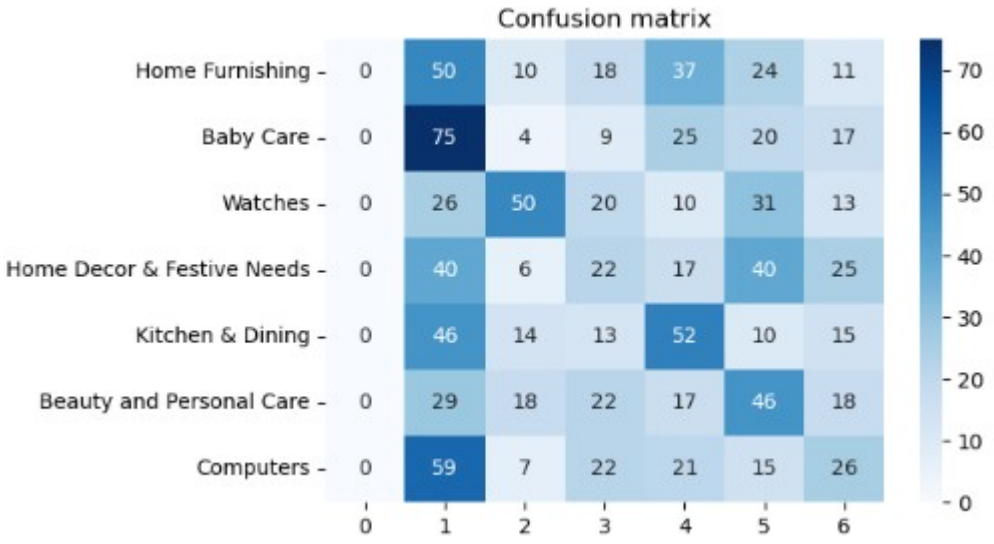


Methode	ARI	homogeneity	completeness	v_measure
sift	0.0427	0.0714	0.0718	0.0716

CV - ORB

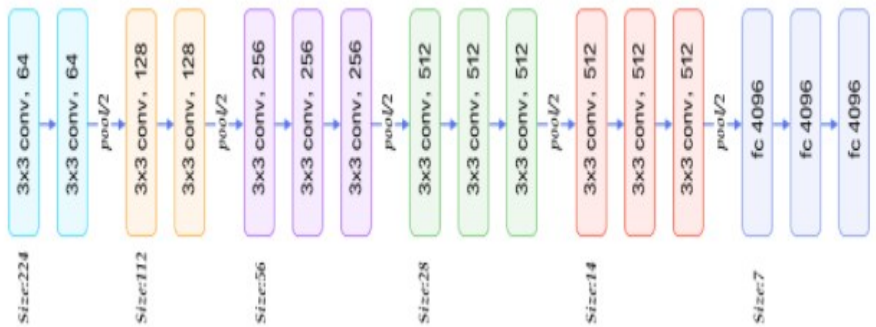




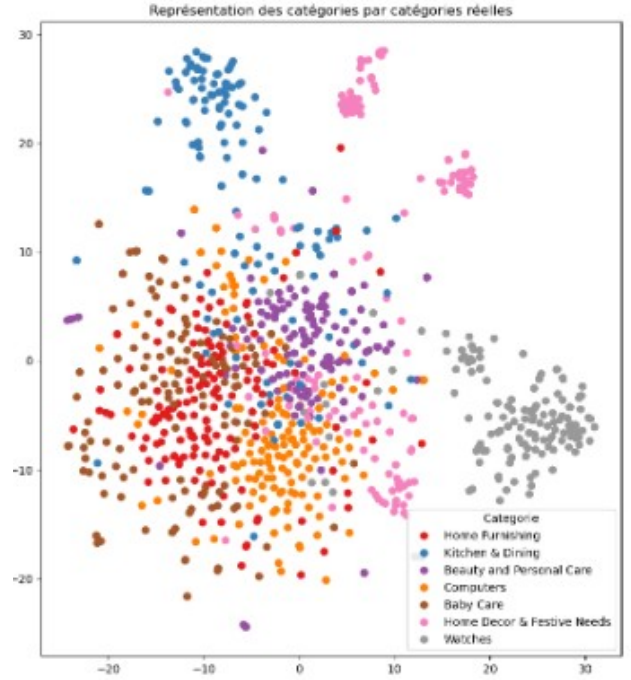


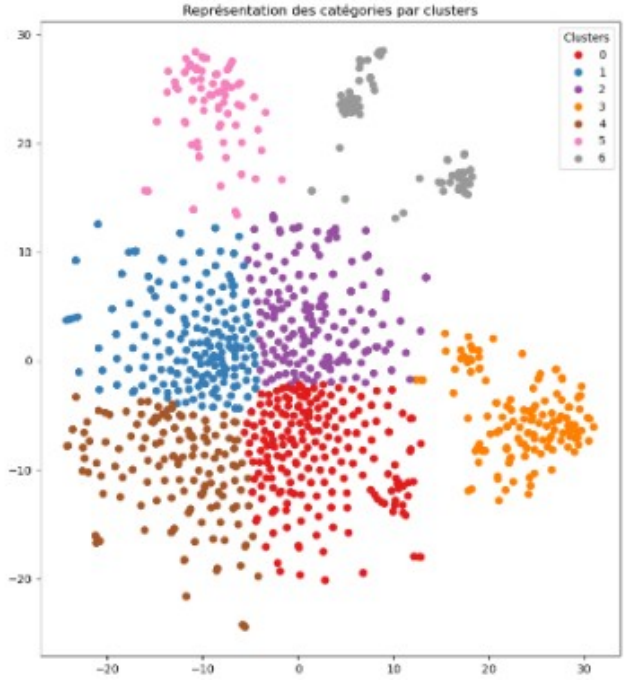
Methode	ARI	homogeneity	completeness	v_measure
orb	0.0339	0.0580	0.0585	0.0582

CV - VGG16



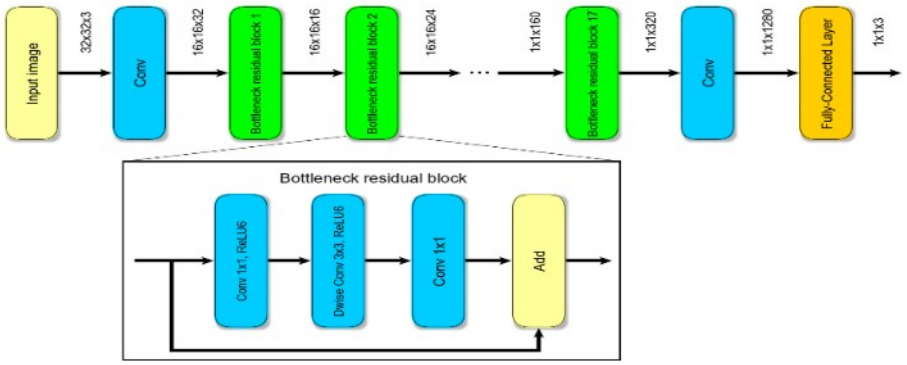
(1050, 25088)



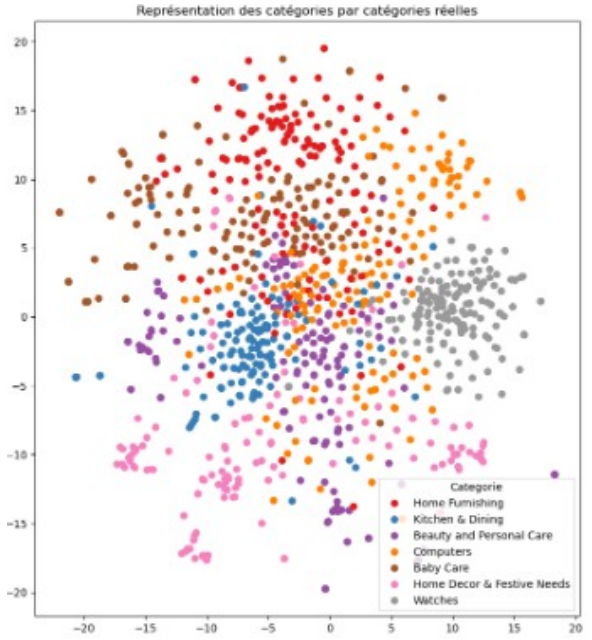


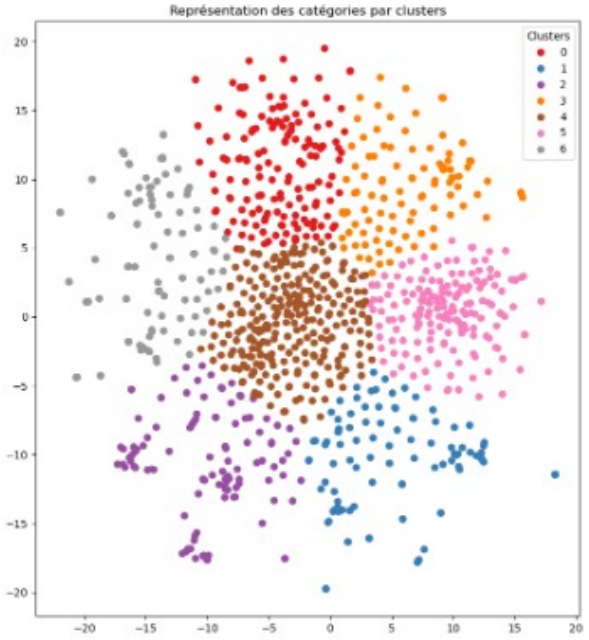
Methode	ARI	homogeneity	completeness	v_measure
vgg16	0.3521	0.4601	0.4722	0.4661

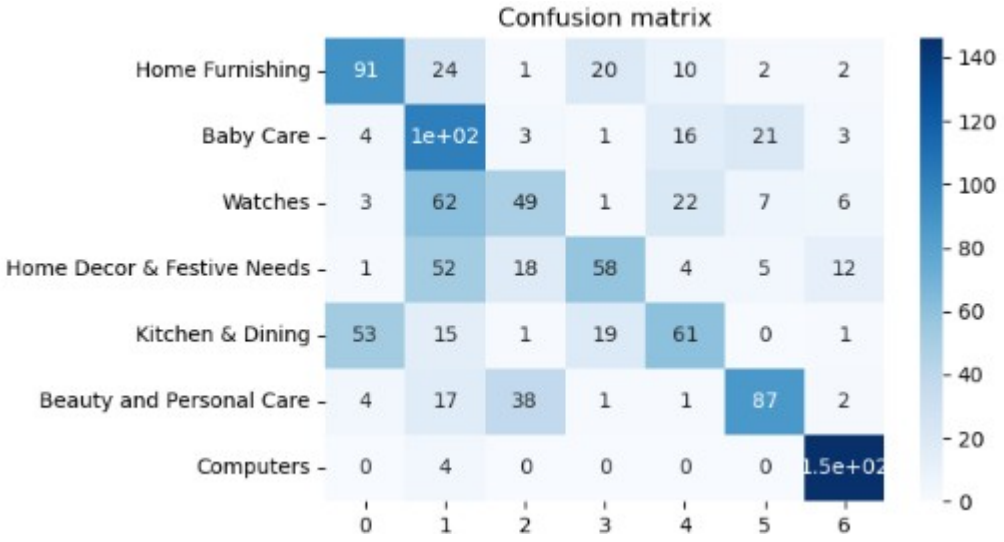
CV - MobileNet



(1050, 62720)

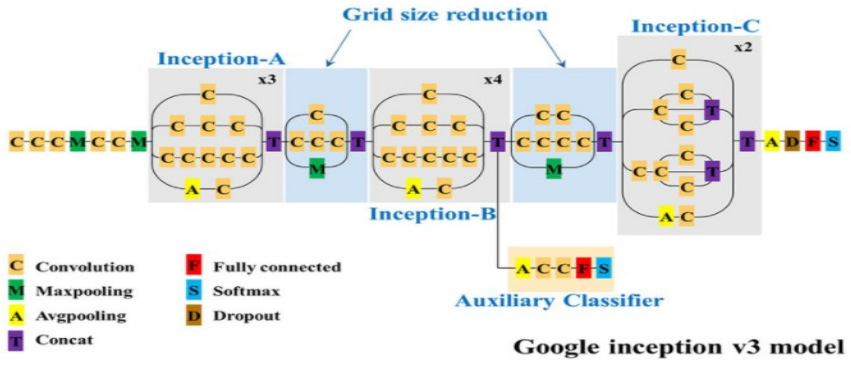






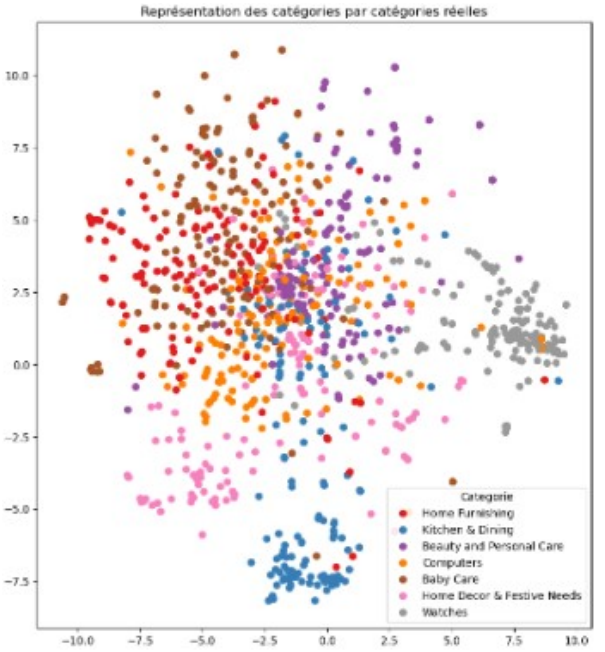
Methode	ARI	homogeneity	completeness	v_measure
mobnet	0.3178	0.4089	0.4208	0.4138

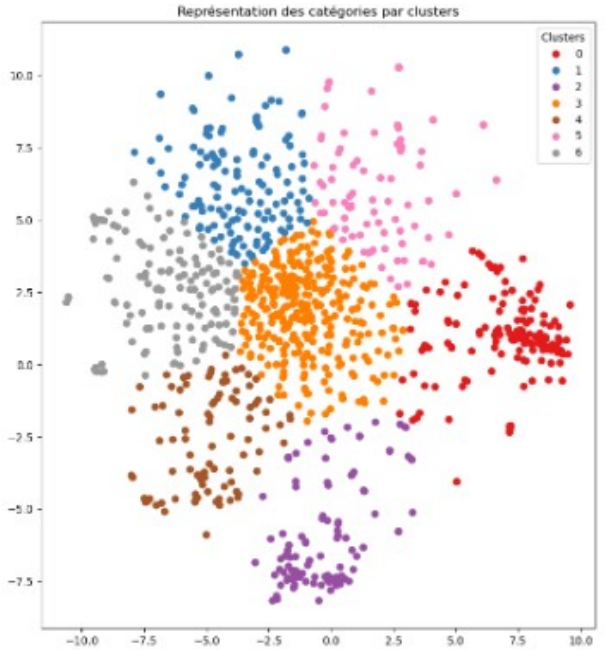
CV - InceptionV3

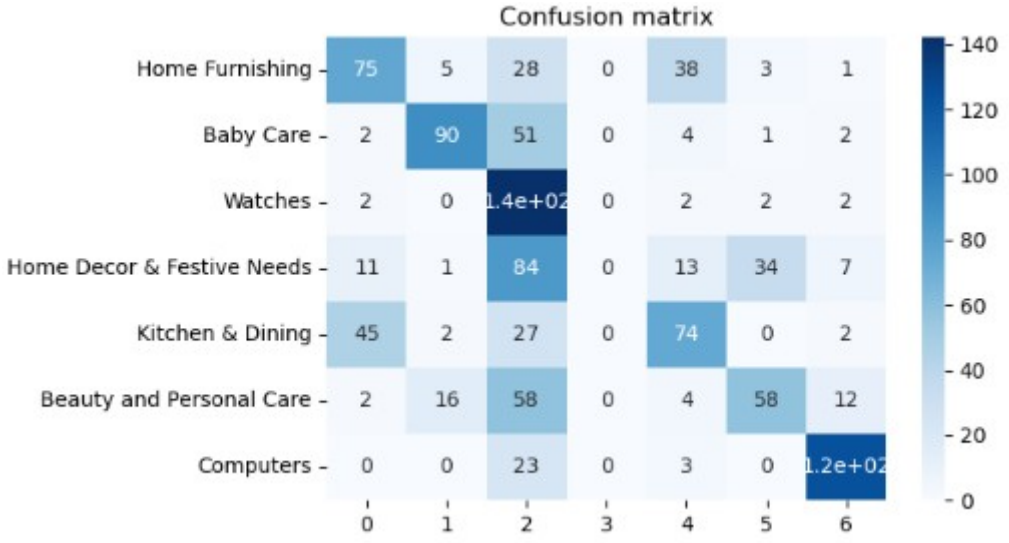


Google inception v3 model

(1050, 51200)

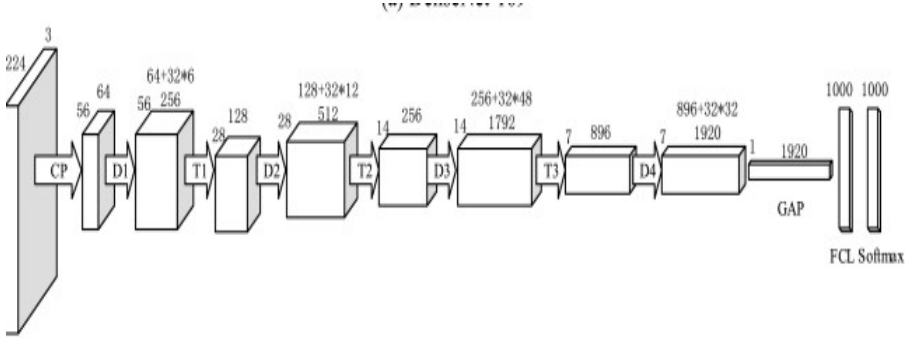




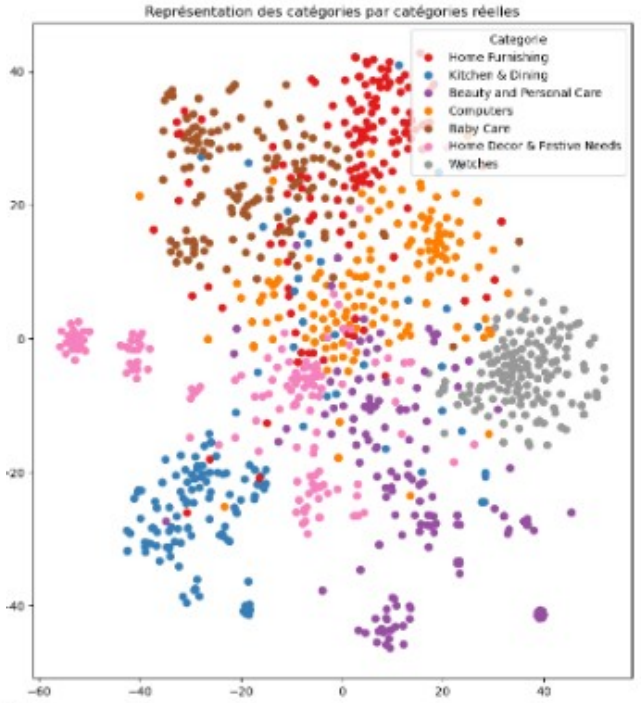


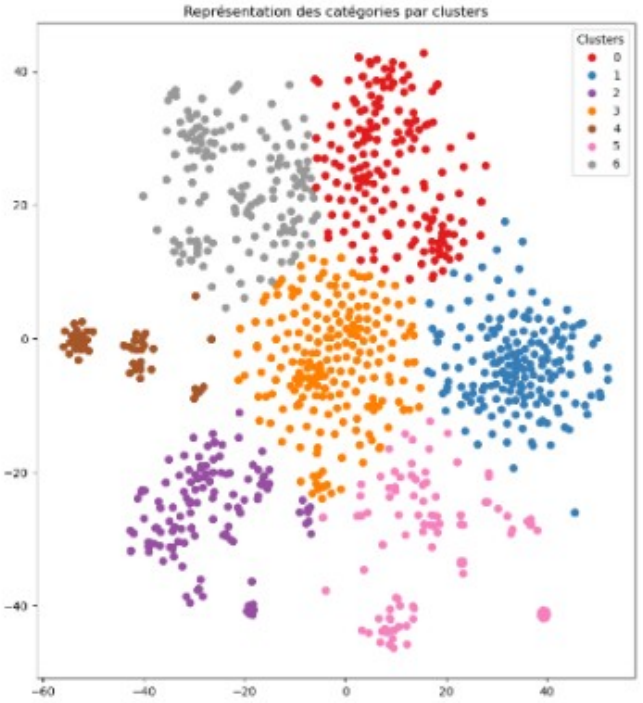
Methode	ARI	homogeneity	completeness	v_measure
inception	0.2562	0.3675	0.3847	0.3759

CV - DenseNet



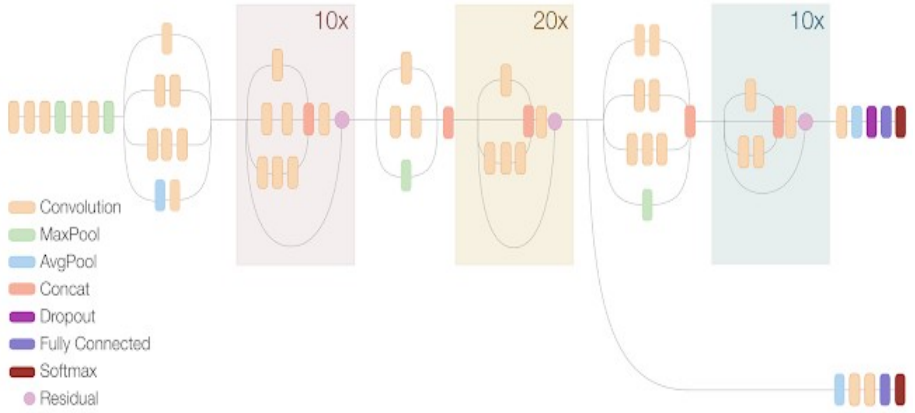
(1050, 94080)



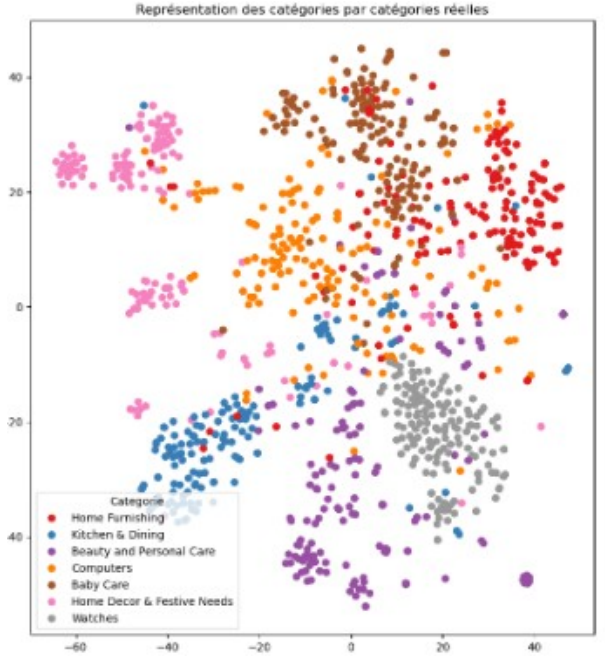


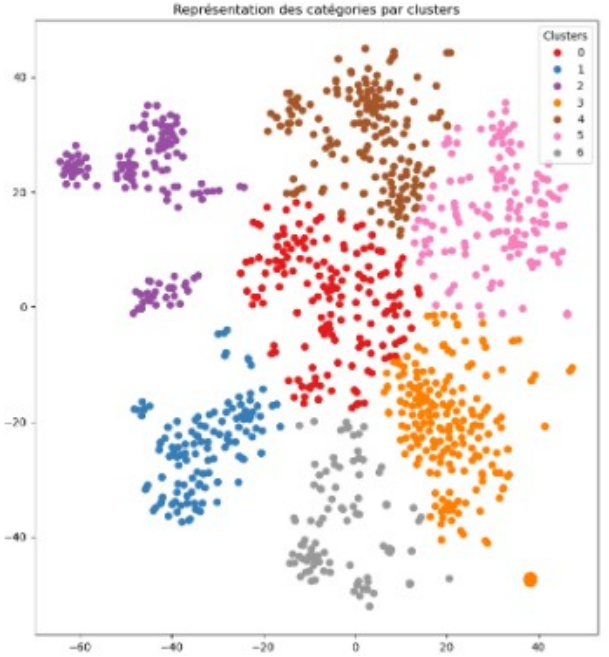
Methode	ARI	homogeneity	completeness	v_measure
dense	0.4554	0.5372	0.5584	0.5486

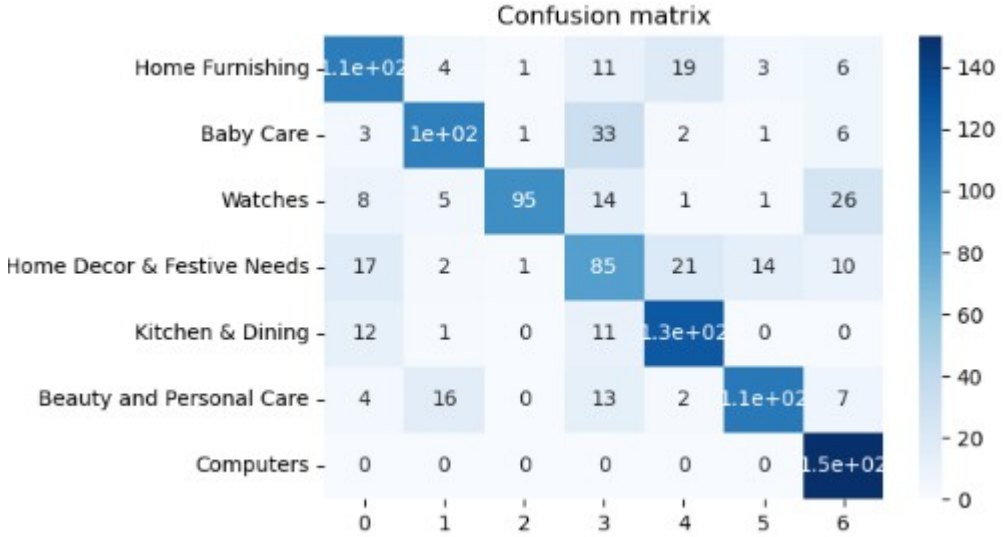
CV - InceptionResNetV2



(1050, 38400)



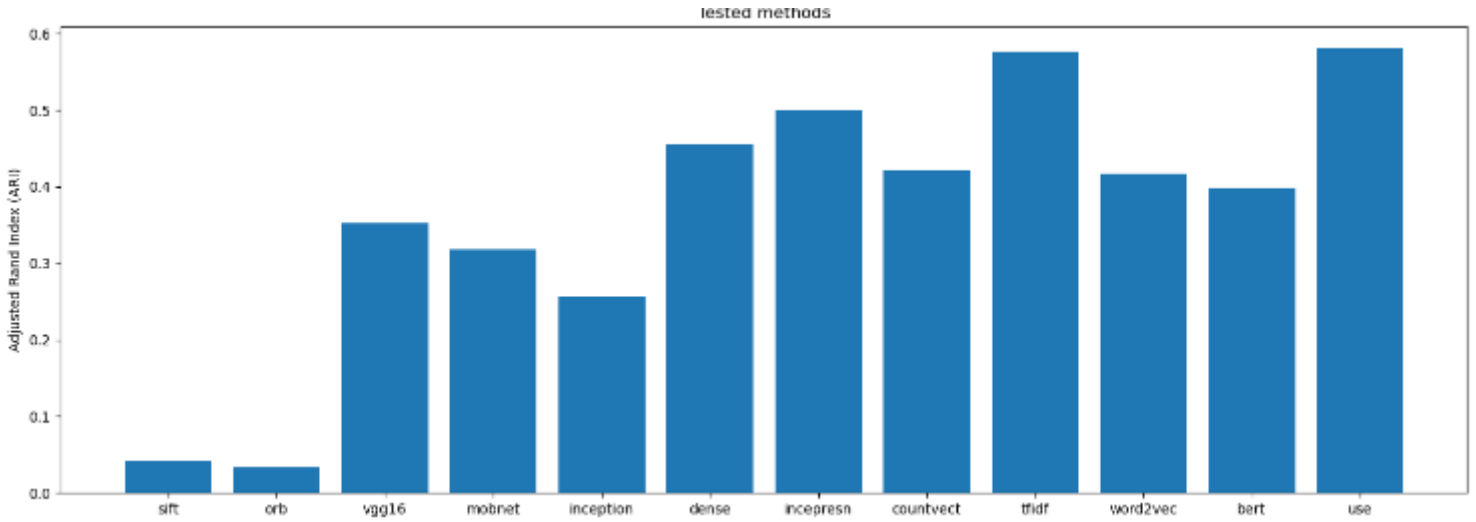




Methode	ARI	homogeneity	completeness	v_measure
incepresn	0.5003	0.5475	0.5542	0.5508

Comparaison

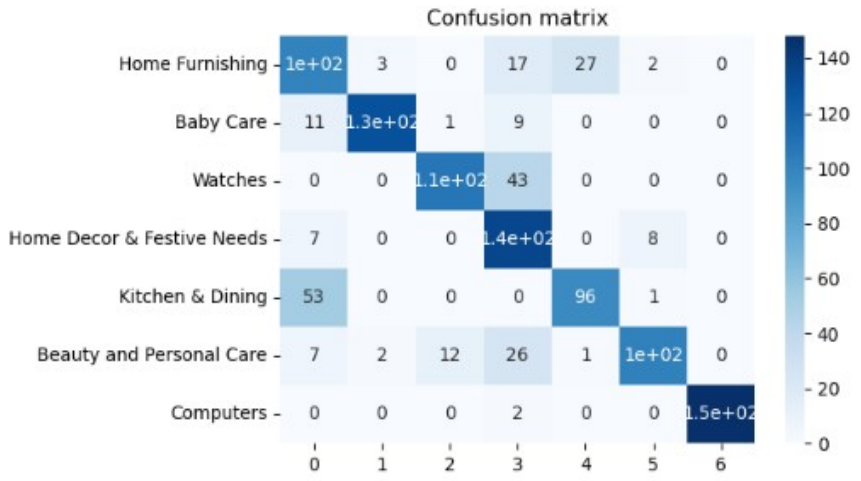
	Methode	ARI	homogeneity	completeness	v_measure
0	sift	0.0427	0.0714	0.0718	0.0716
1	orb	0.0339	0.0580	0.0585	0.0582
2	vgg16	0.3521	0.4601	0.4722	0.4661
3	mobnet	0.3176	0.4069	0.4206	0.4136
4	inception	0.2562	0.3675	0.3847	0.3759
5	dense	0.4554	0.5372	0.5564	0.5466
6	incepresn	0.5003	0.5475	0.5542	0.5508
7	countvect	0.4214	0.5347	0.5609	0.5475
8	tfidf	0.5766	0.6517	0.6693	0.6604
9	word2vec	0.4165	0.5333	0.5409	0.5371
10	bert	0.3988	0.4692	0.4727	0.4709
11	use	0.5802	0.6729	0.6840	0.6784

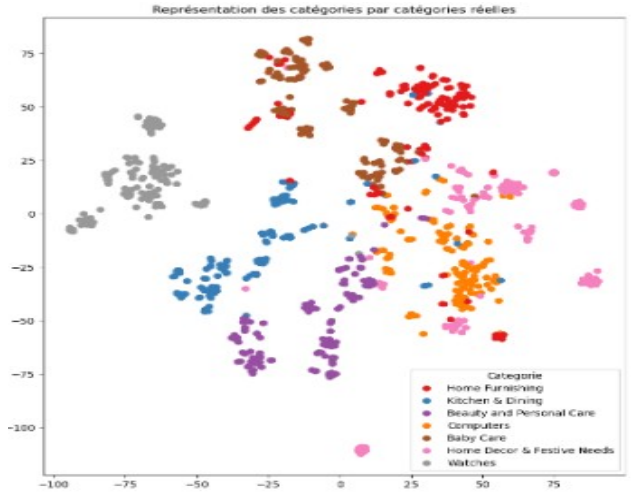


Conclusion

- **Étude de faisabilité d'un moteur de classification validée**

Meilleur	Données visuelles	Données textuelles
Modèle	InceptionResNetV2	**USE**
ARI	0,51	**0,58**





Perspectives

- Enrichir le jeu de données
- Combinaison textes et images
- Utiliser des méthodes d’apprentissage supervisées

Merci de votre attention

Contact : bouzaieni@gmail.com