

Projet 6 : Classifiez automatiquement des biens de consommation

Présenté par :

Bourama FANE
Etudiant Data Scientist

Dirigé par :

Babou M'BAYE
Mentor chez OpenClassrooms

25 Aout 2023

- 1 Problématique
- 2 Nettoyage des données
- 3 Méthodologie
- 4 Données Textuelles
- 5 Données images
- 6 Classification supervisée



Plan de la présentation

1 Problématique

2 Nettoyage des données

3 Méthodologie

4 Données Textuelles

5 Données images

6 Classification supervisée

Problématique

Sur la **place de marché**, des vendeurs proposent des articles à des acheteurs en postant une photo et une description. L'entreprise **Place de marché** souhaite lancer une marketplace **e-commerce**.



Pour l'instant, l'**attribution de la catégorie** d'un article est **manuelle**, et est donc **peu fiable**. De plus, le volume des articles est pour l'instant **très petit**.

Afin de passer à une plus large échelle et faciliter le processus, il devient **nécessaire d'automatiser cette tâche**.

Votre objectif est de réaliser une première étude de faisabilité d'un moteur de **classification** d'articles.



Données

Sources de données

Une base de données de **1050 images** avec les catégories, les images et la description du produit est mise à notre disposition.

Baby Care



Watches



Travail à faire

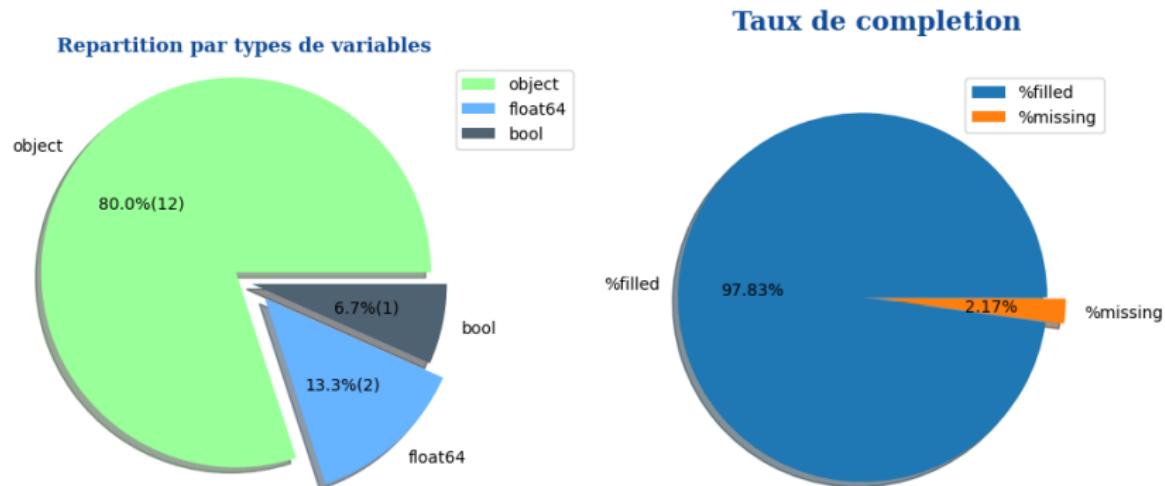
- ☛ Analyser le jeu de données,
- ☛ Prétraitement des descriptions des produits,
- ☛ Prétraitement des images,
- ☛ Extraction de features,
- ☛ Réduction de dimension,
- ☛ Clustering,
- ☛ En fin, classification supervisée.



Plan de la présentation

- 1 Problématique
- 2 Nettoyage des données
- 3 Méthodologie
- 4 Données Textuelles
- 5 Données images
- 6 Classification supervisée

Repartition & Taux de compléction



- 80% des variables sont de types object ;
- La base contient de 2% de valeurs manquantes.

Nettoyage

- ☛ Focus sur les colonnes **product_category_tree**

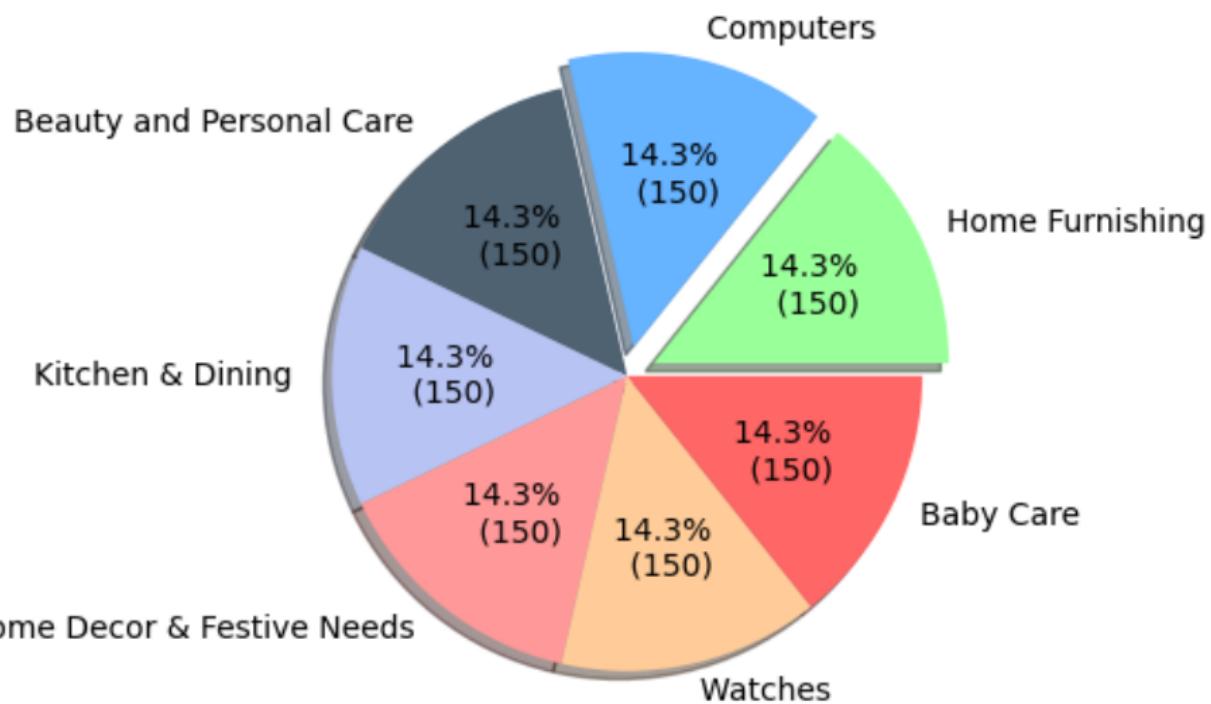
- ☛ La catégorie est un arbre séparé par '> >', de profondeur 7.

Valeurs uniques catégorie niveau 1 = 7
Valeurs uniques catégorie niveau 2 = 62
Valeurs uniques catégorie niveau 3 = 241
Valeurs uniques catégorie niveau 4 = 349
Valeurs uniques catégorie niveau 5 = 297
Valeurs uniques catégorie niveau 6 = 117
Valeurs uniques catégorie niveau 7 = 57

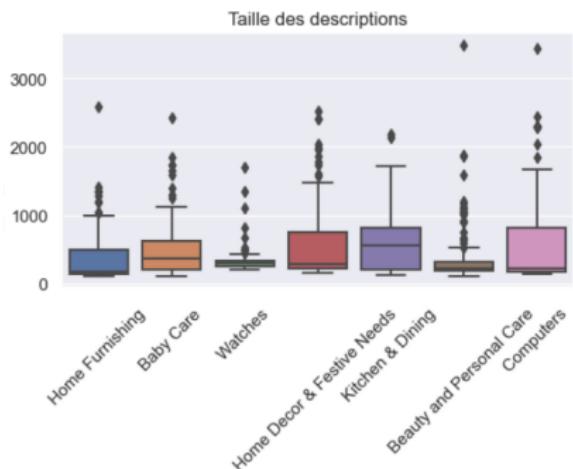
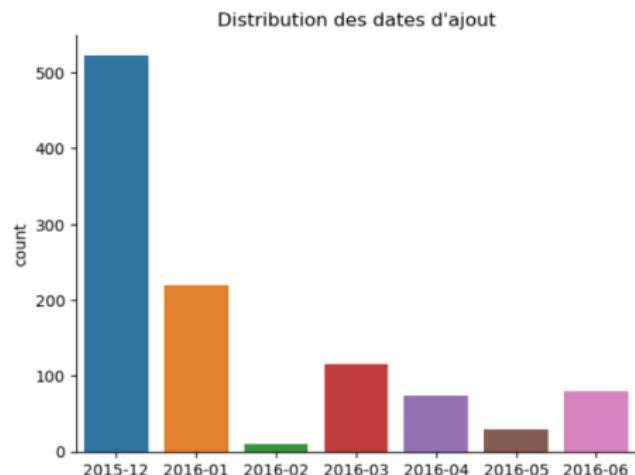
- ☛ Pour les variables prix (**retail_price** et **discounted_price**), nous avons imputé par la **moyenne**.
- ☛ Pour les variables **brand** et **product_specifications**, nous avons remplacé les valeurs manquantes par l'expression <<**Inconnu**>>.

Répartition

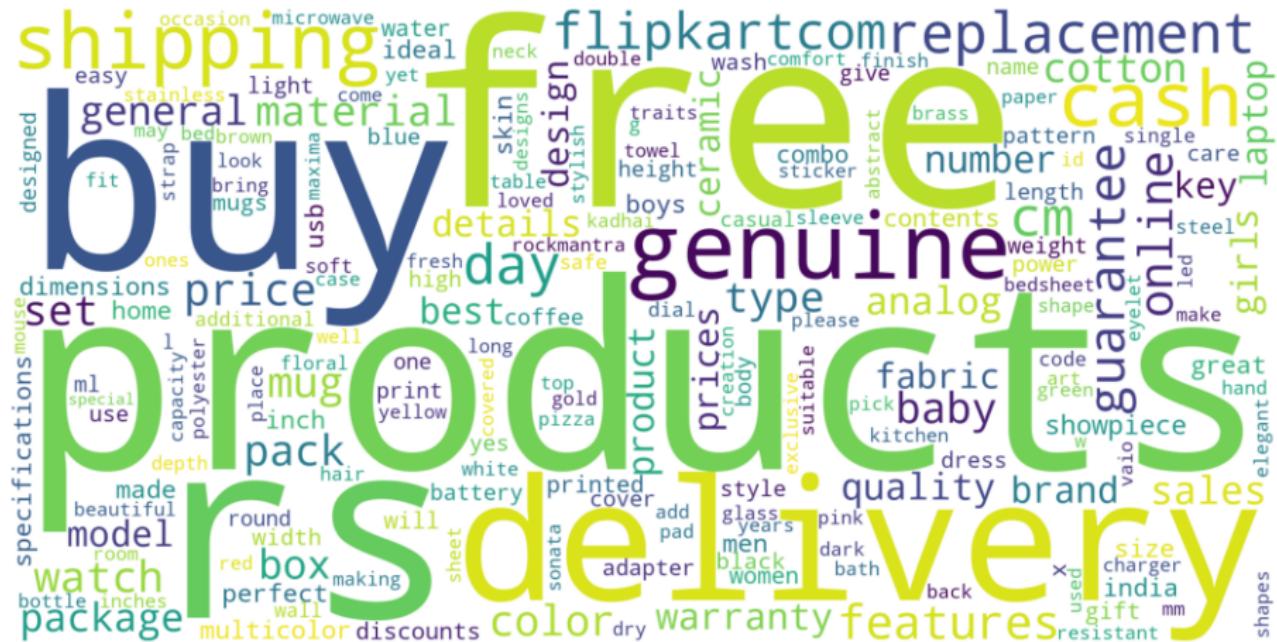
Repartition selon la categorie



Distribution dates ajout, longueur du champ description



Nuage de mots : description



Nuage de mots : brand



Plan de la présentation

- 1 Problématique
- 2 Nettoyage des données
- 3 Méthodologie
- 4 Données Textuelles
- 5 Données images
- 6 Classification supervisée



Comment allons-nous capter la faisabilité ?

	Pré traitement	Features extraction et description <i>Construction d'un vecteur numérique</i>	Réduction de dimension	Clustering	Visualisation	Evaluation
Données textuelles	Récupération des tokens, nettoyage et création d'un vocabulaire	<ul style="list-style-type: none"> ▪ Bag of word : count-vectorizer, TF-IDF ▪ Words embedding word2vec, BERT, USE 				
Données images	Récupération des images et réduction de la taille	<ul style="list-style-type: none"> ▪ Bag of visual word : SIFT, ORB ▪ Embedding : CNN 	ACP	K-means, GMM	ACP, TSNE	ARI, Accuracy

Classification Supervisée et Data Augmentation

Après avoir démontré la faisabilité de regrouper automatiquement des produits de même catégorie, on fera:

1.Une classification supervisée à partir des images des articles.

2.Une mise en place de la data augmentation pour optimiser le modèle.

Plan de la présentation

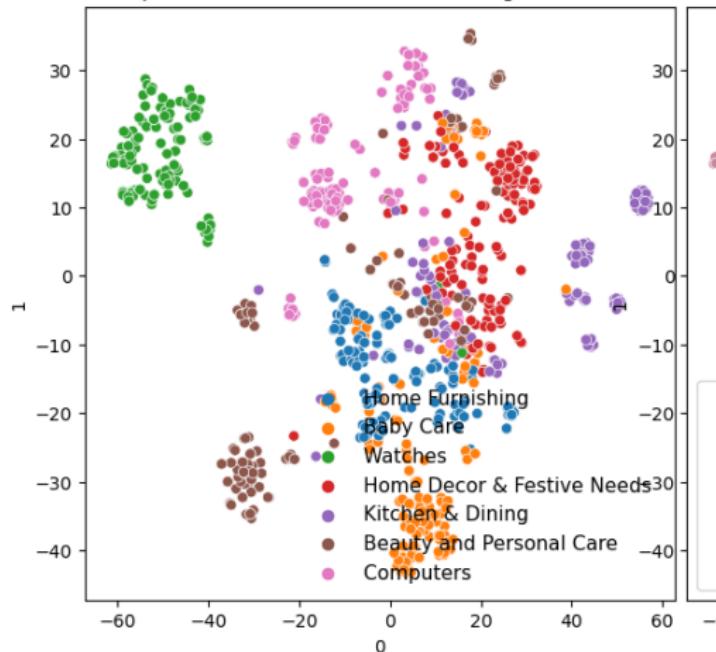
- 1 Problématique
- 2 Nettoyage des données
- 3 Méthodologie
- 4 Données Textuelles
- 5 Données images
- 6 Classification supervisée

Comment avons-nous procédé ?

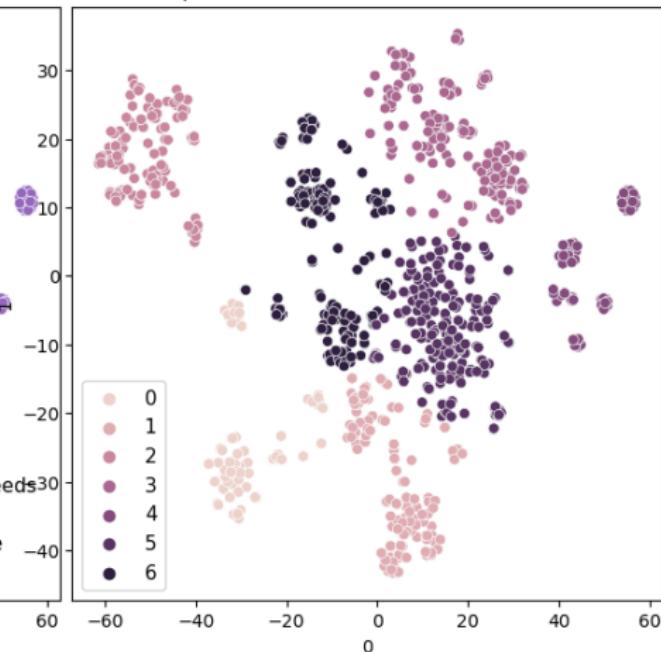


Approche Bag-of-words

Représentation en fonction des catégories réelles



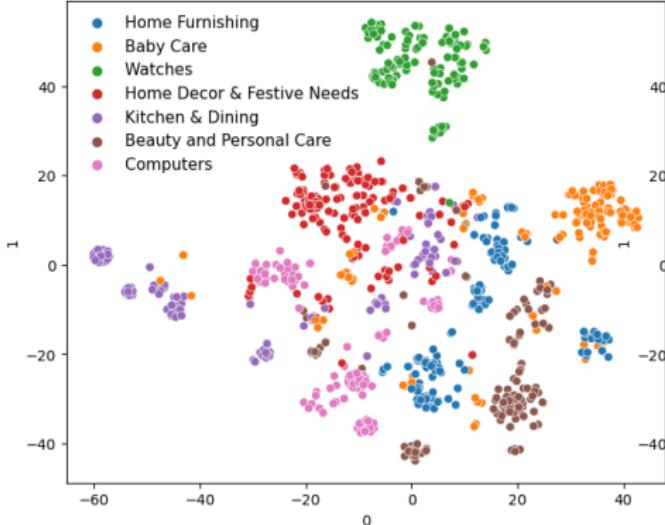
Représentation en fonction des clusters



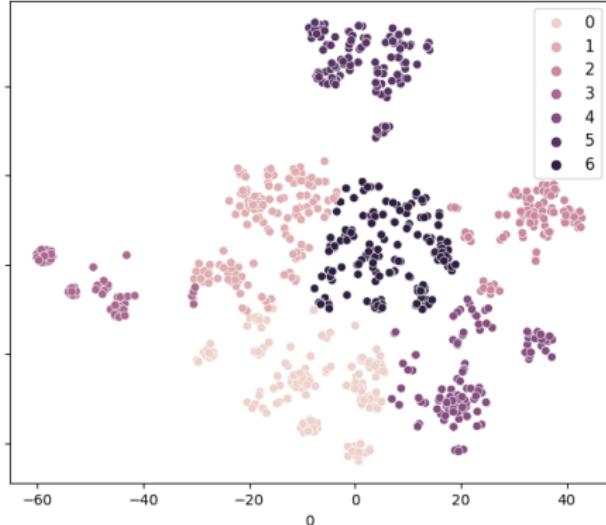
- ☞ Le score est **ARI : 0.3969**,
- ☞ Les catégories ne sont pas bien séparées.

Tf-idf (term frequencyinverse document frequency)

Représentation en fonction des catégories réelles



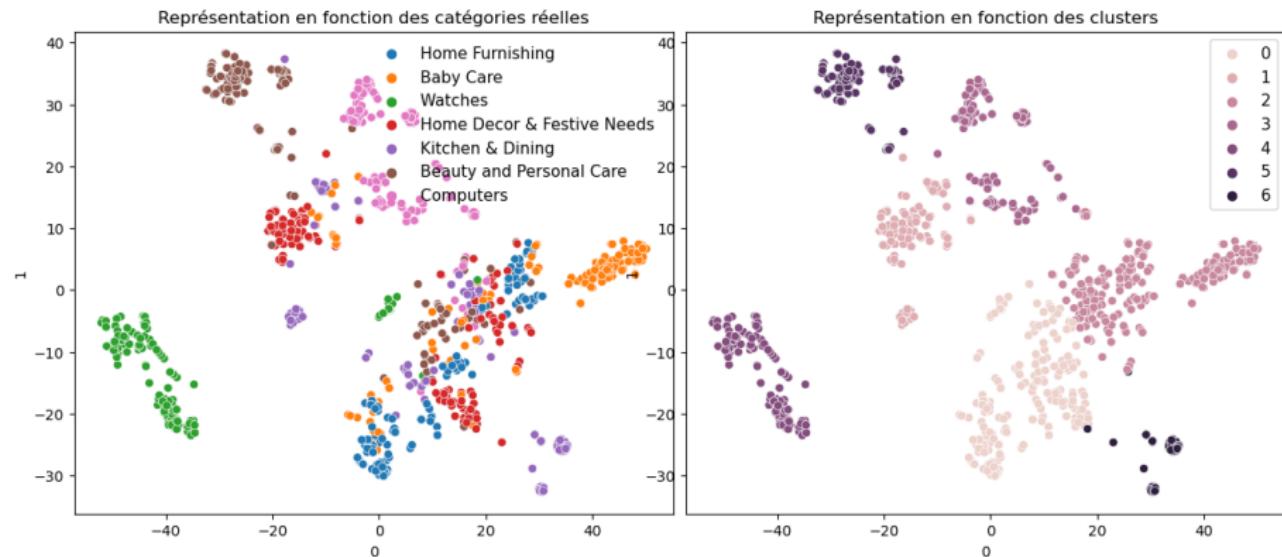
Représentation en fonction des clusters



- ☛ Le score est **ARI : 0.4055**,
- ☛ Les catégories sont un mieux séparées.



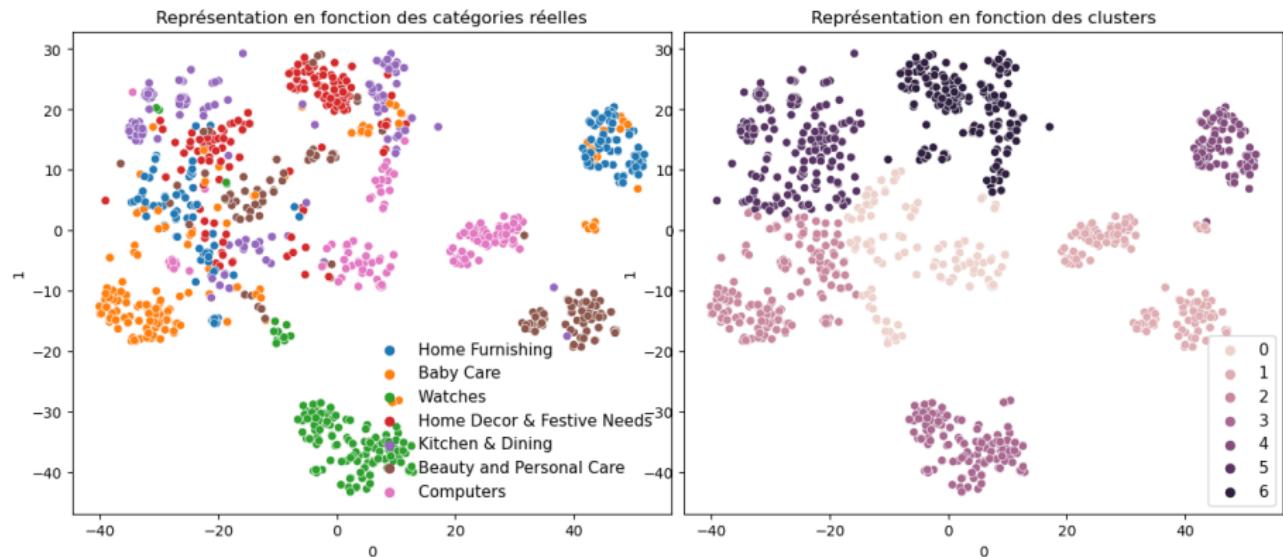
Approche wordsentence embedding classique avec Word2Vec



- Le score est **ARI : 0.3465**,
- Les catégories ne sont pas bien séparées.



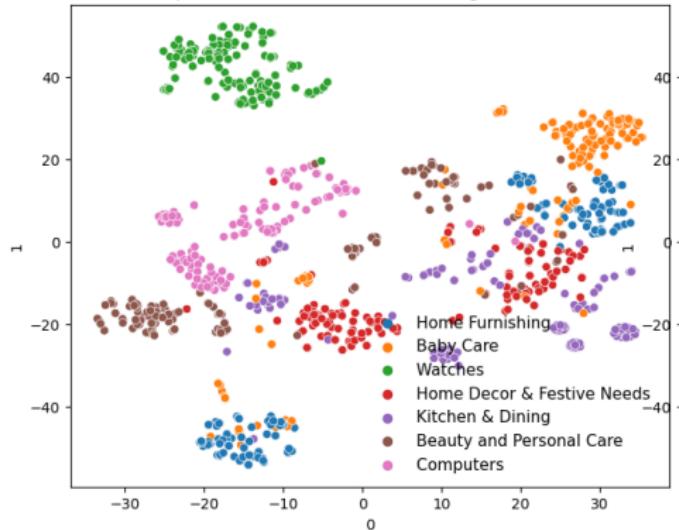
BERT (Bidirectional Encoder Representations from Transformers)



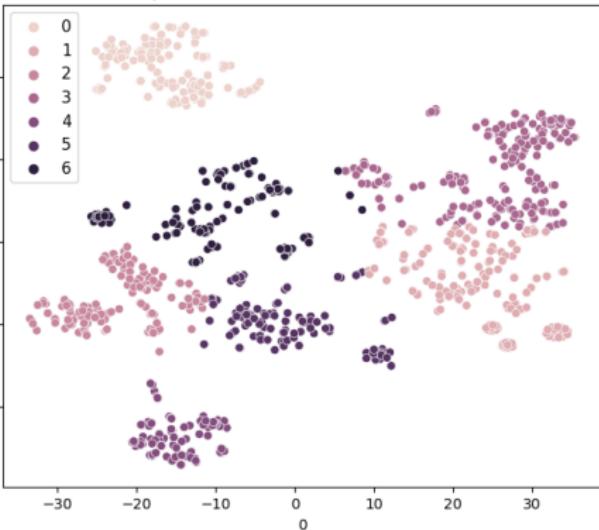
- Le score est **ARI : 0.34**,
- Les catégories ne sont pas bien séparées.

Approche USE

Représentation en fonction des catégories réelles



Représentation en fonction des clusters

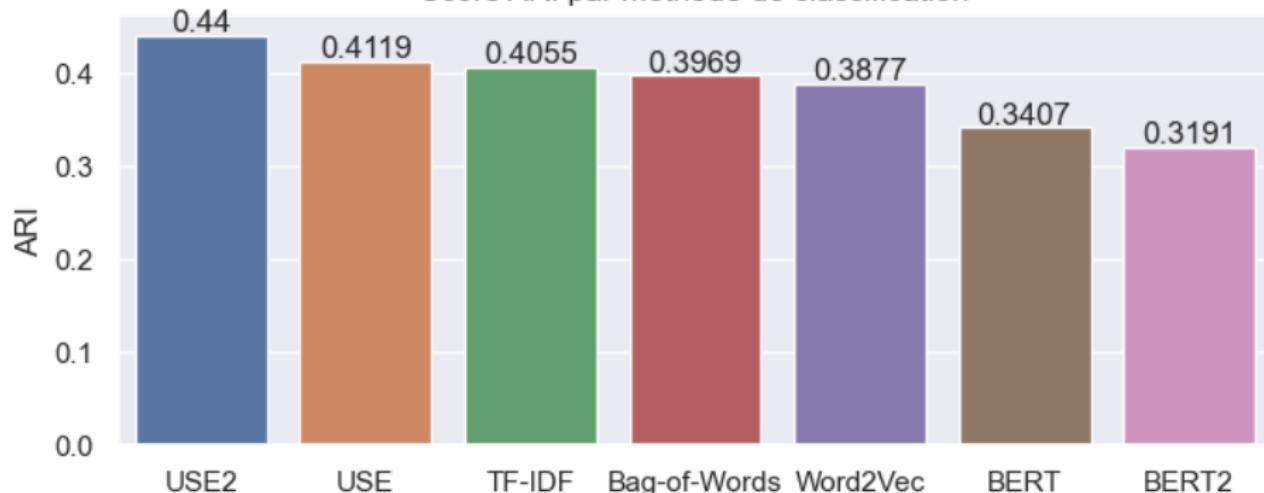


- Le score est **ARI : 0.4119**,
- Les catégories sont mieux séparées.



Bilan données textuelles

Score ARI par méthode de classification



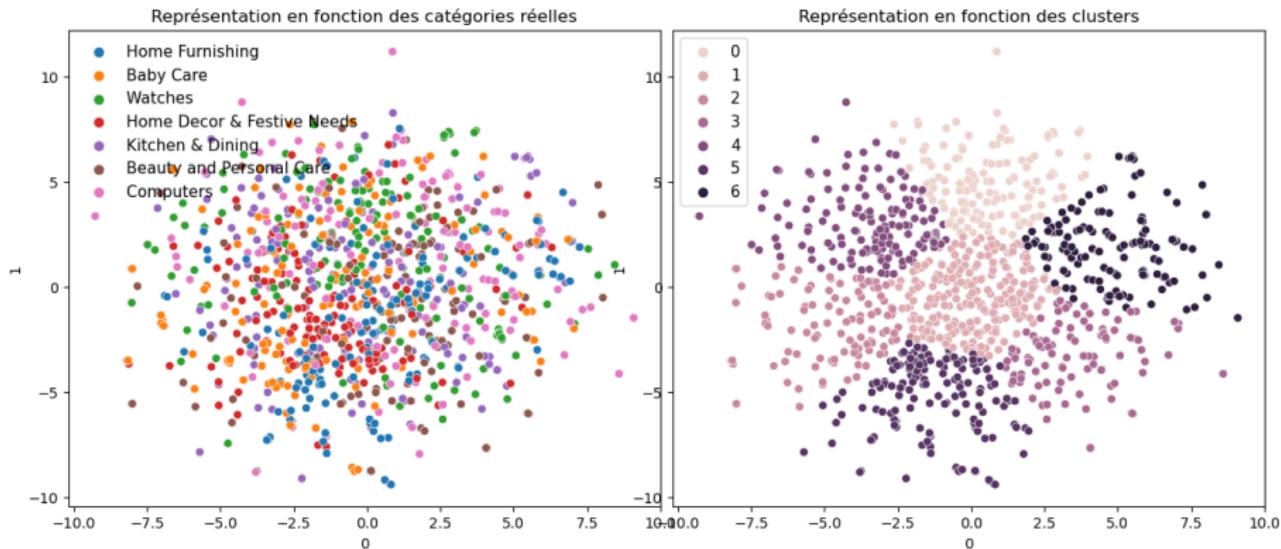
- La méthode USE présente le **meilleur score ARI**.
- Elle est suivie de la méthode **TF-IDF**.



Plan de la présentation

- 1 Problématique
- 2 Nettoyage des données
- 3 Méthodologie
- 4 Données Textuelles
- 5 Données images
- 6 Classification supervisée

SIFT- (scale-invariant feature transform)



- ☛ Le score est **ARI : 0.0277**,
- ☛ Les catégories ne sont pas bien séparées.



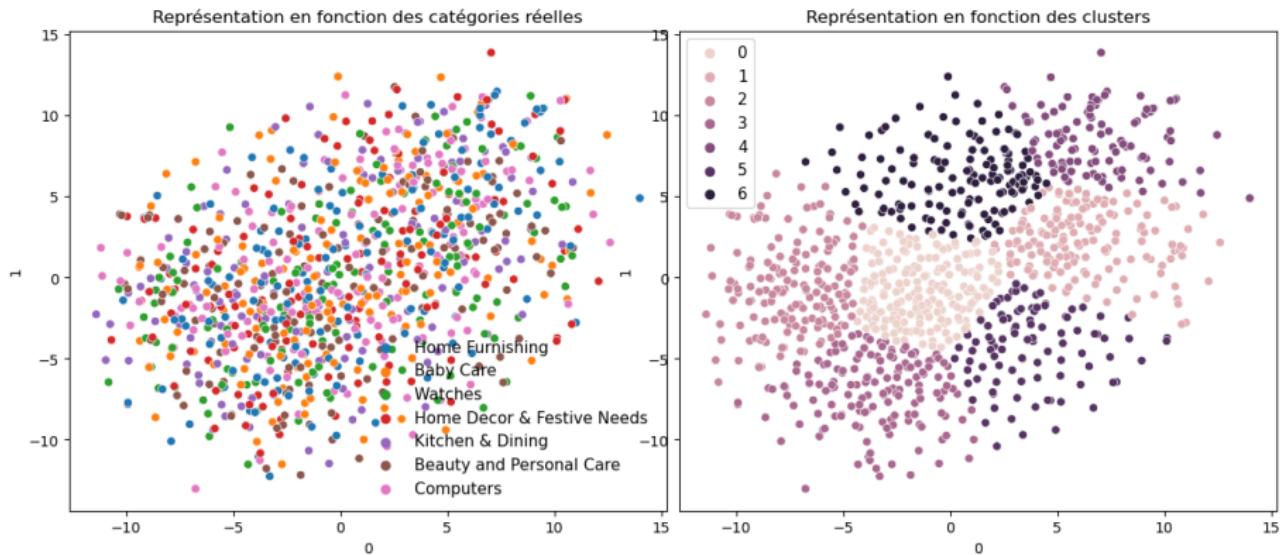
SIFT : Matrice de confusion

Matrice de confusion

		0	1	2	3	4	5	6	
Baby Care -	0	3	33	68	2	44	0		- 80
Beauty and Personal Care -	0	36	27	74	0	13	0		- 60
Computers -	1	13	99	22	6	8	1		- 40
Home Decor & Festive Needs -	0	4	15	84	2	45	0		- 20
Home Furnishing -	0	4	21	68	6	51	0		- 0
Kitchen & Dining -	0	12	39	76	1	22	0		
Watches -	0	21	59	65	0	5	0		
	0	1	2	3	4	5	6		

- 👉 Le score est **ARI : 0.0277**,
- 👉 Les catégories ne sont pas bien séparées.

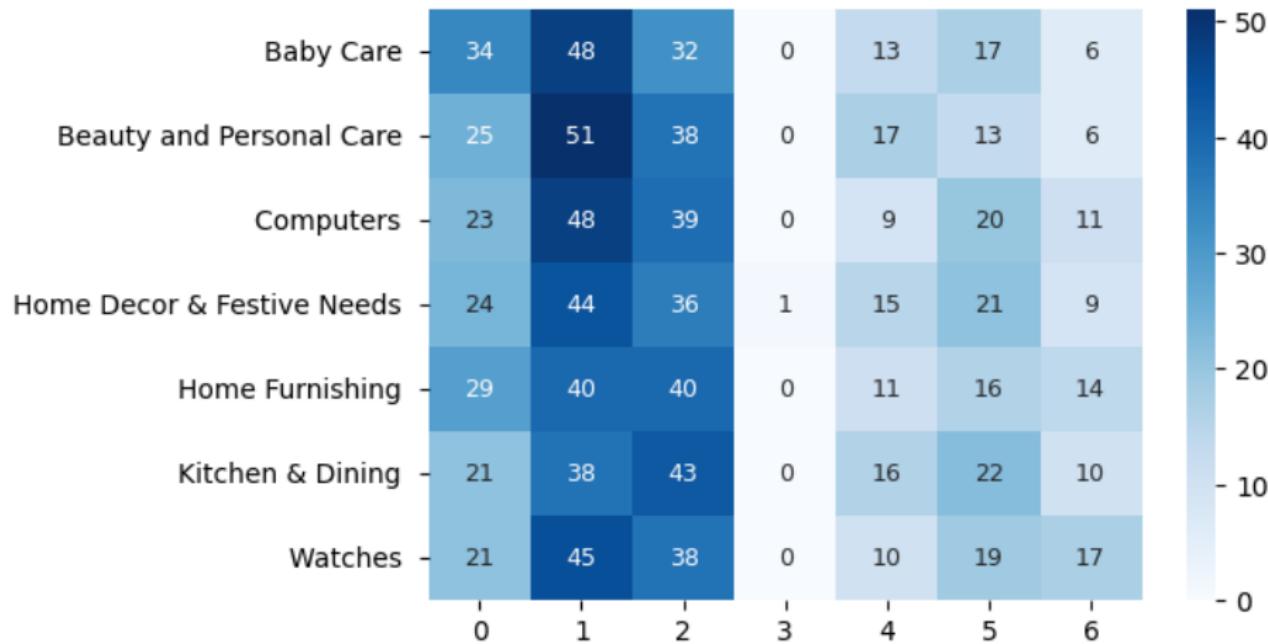
ORB (Oriented FAST and Rotated BRIEF)



- Le score est **ARI : 0.0001**,
- Les catégories ne sont pas bien séparées.

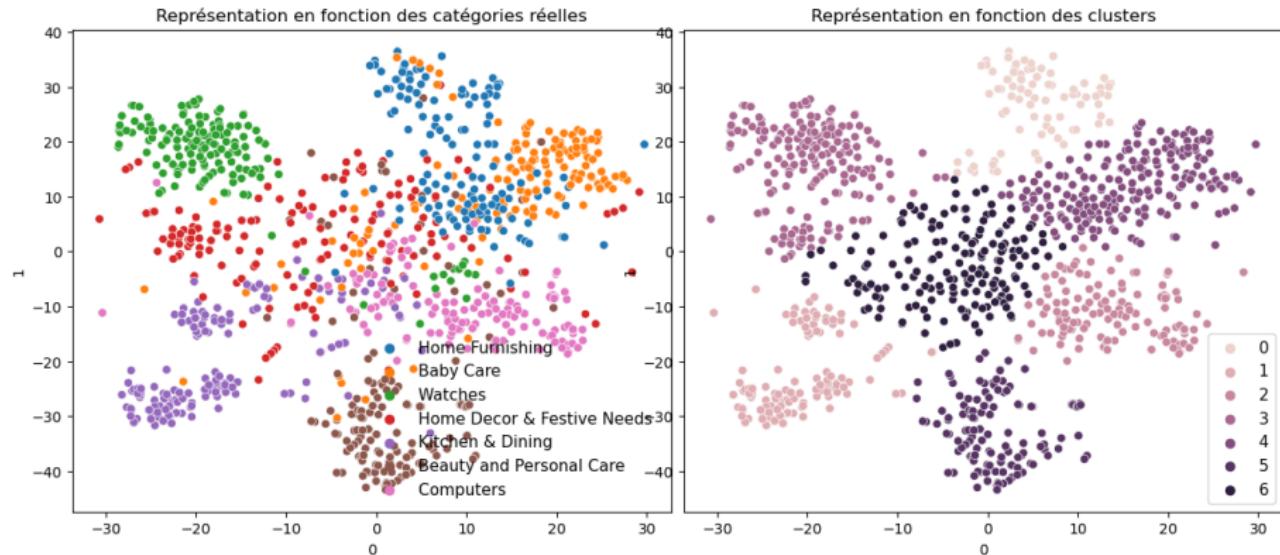
ORB : Matrice de confusion

Matrice de confusion



- Le score est **ARI : 0.0001**,
- Les catégories ne sont pas bien séparées.

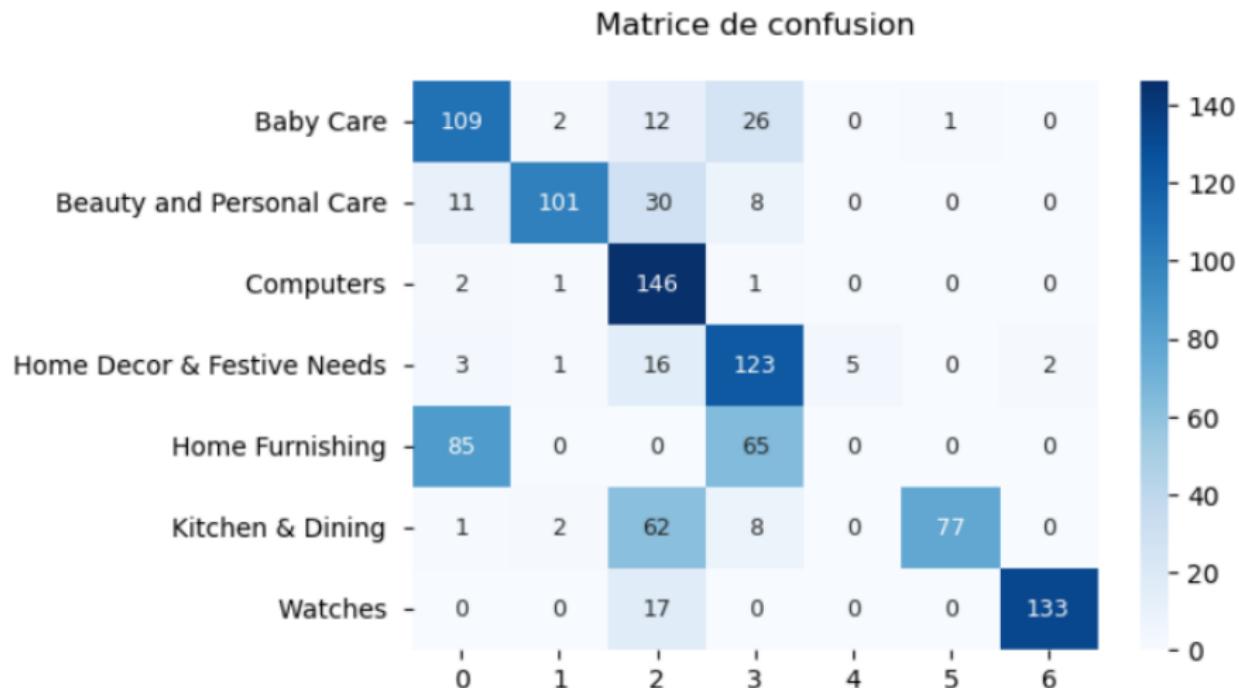
Approche CNN : VGG16



- ☞ Le score est **ARI : 0.4577**,
 - ☞ Les catégories sont assez bien séparées.



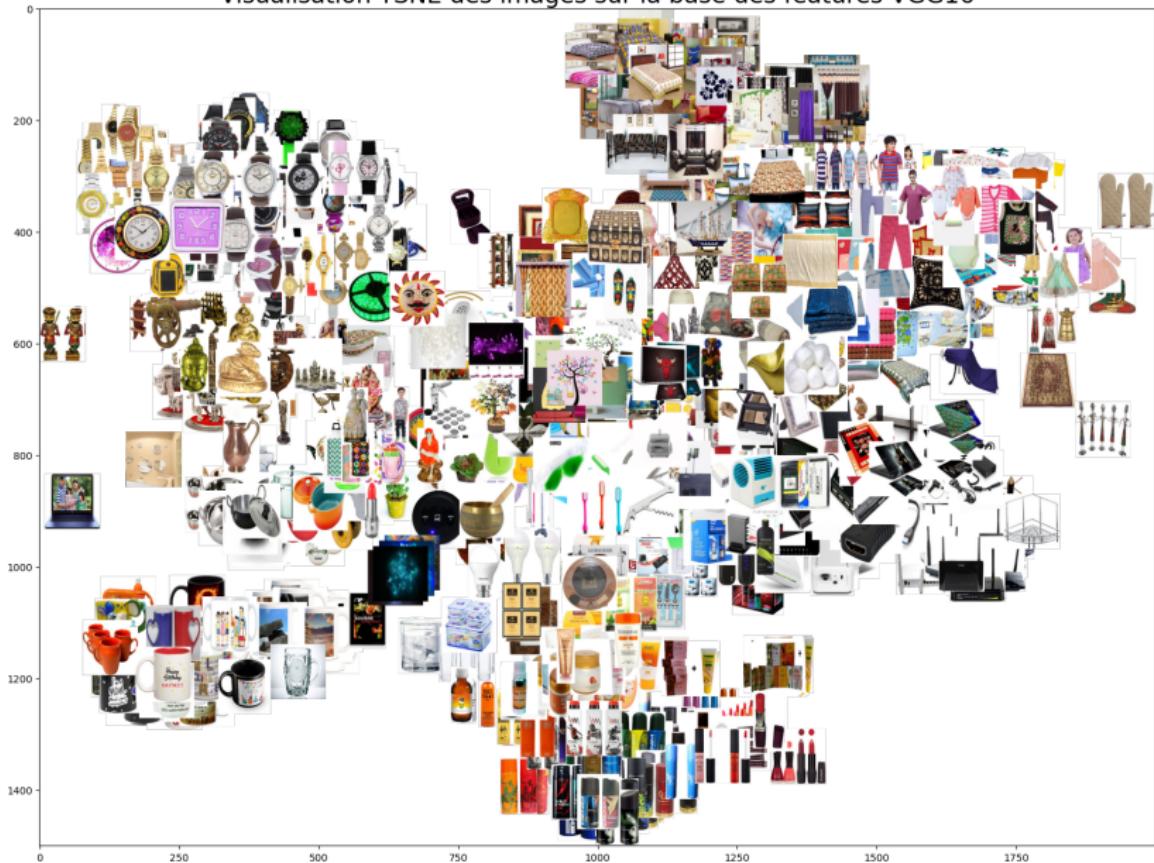
Approche CNN : Matrice de confusion VGG16



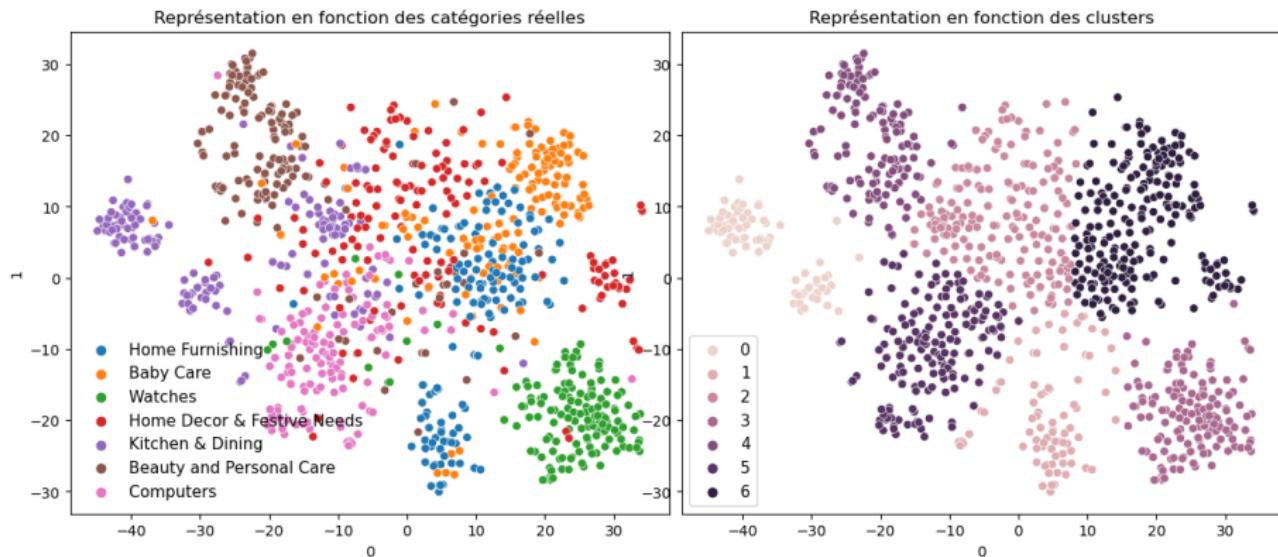
- Le score est **ARI : 0.4577**,
- Les catégories ne sont pas bien séparées.

Approche CNN : Regroupement des images TSNE

Visualisation TSNE des images sur la base des features VGG16



Approche CNN : VGG19



- Le score est **ARI : 0.4442**,
- Les catégories sont assez bien séparées.



Approche CNN : Matrice de confusion VGG19

Matrice de confusion



Le score est ARI : 0.4442

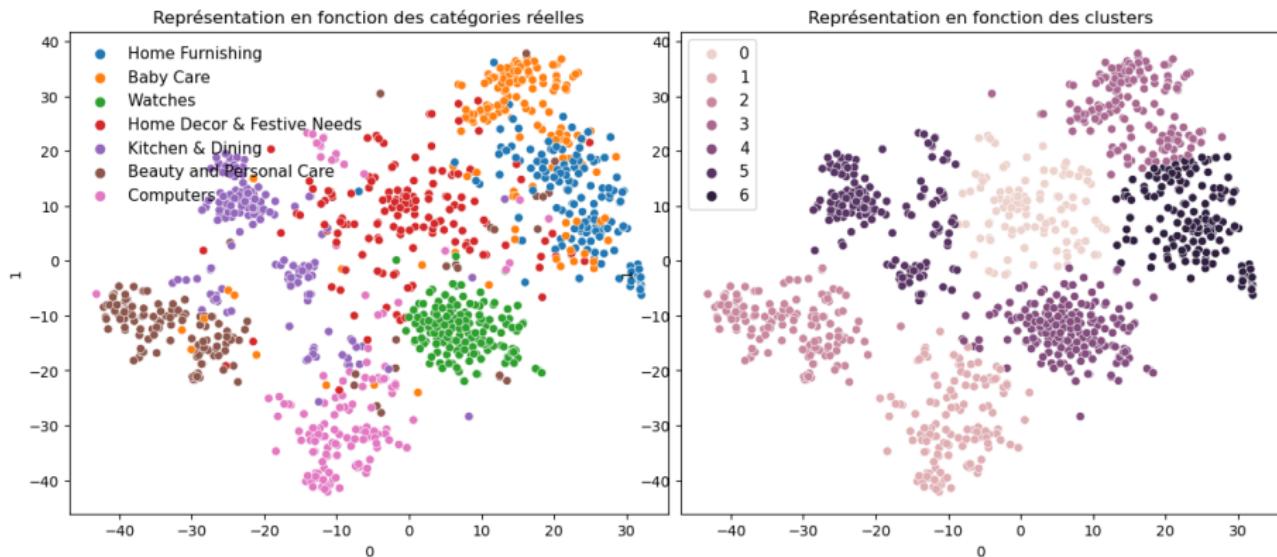
Les catégories ne sont pas bien séparées.

Approche CNN : Regroupement des images TSNE

Visualisation TSNE des images sur la base des features VGG19



Approche CNN : Xception



- Le score est **ARI : 0.5732**,
- Les catégories sont assez bien séparées.



Approche CNN : Matrice de confusion Xception

Matrice de confusion



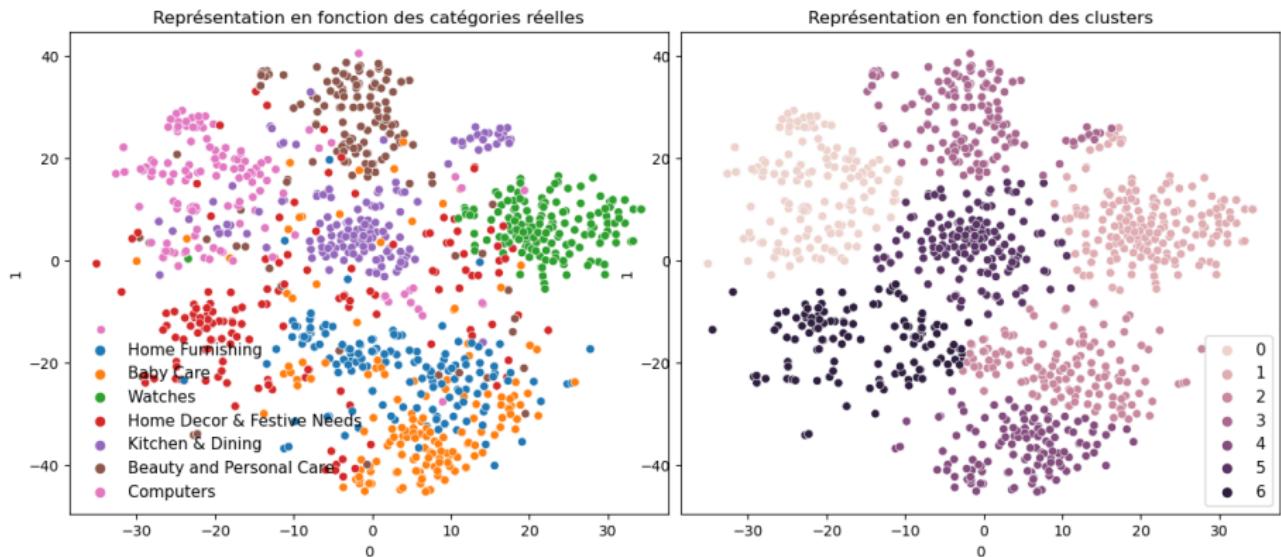
- Le score est ARI : 0.5732,
- Les catégories ne sont pas bien séparées.

Approche CNN : Regroupement des images TSNE

Visualisation TSNE des images sur la base des features Xception



Approche CNN : ResNet50



- Le score est **ARI : 0.4762**,
- Les catégories sont assez bien séparées.



Approche CNN : Matrice de confusion ResNet50

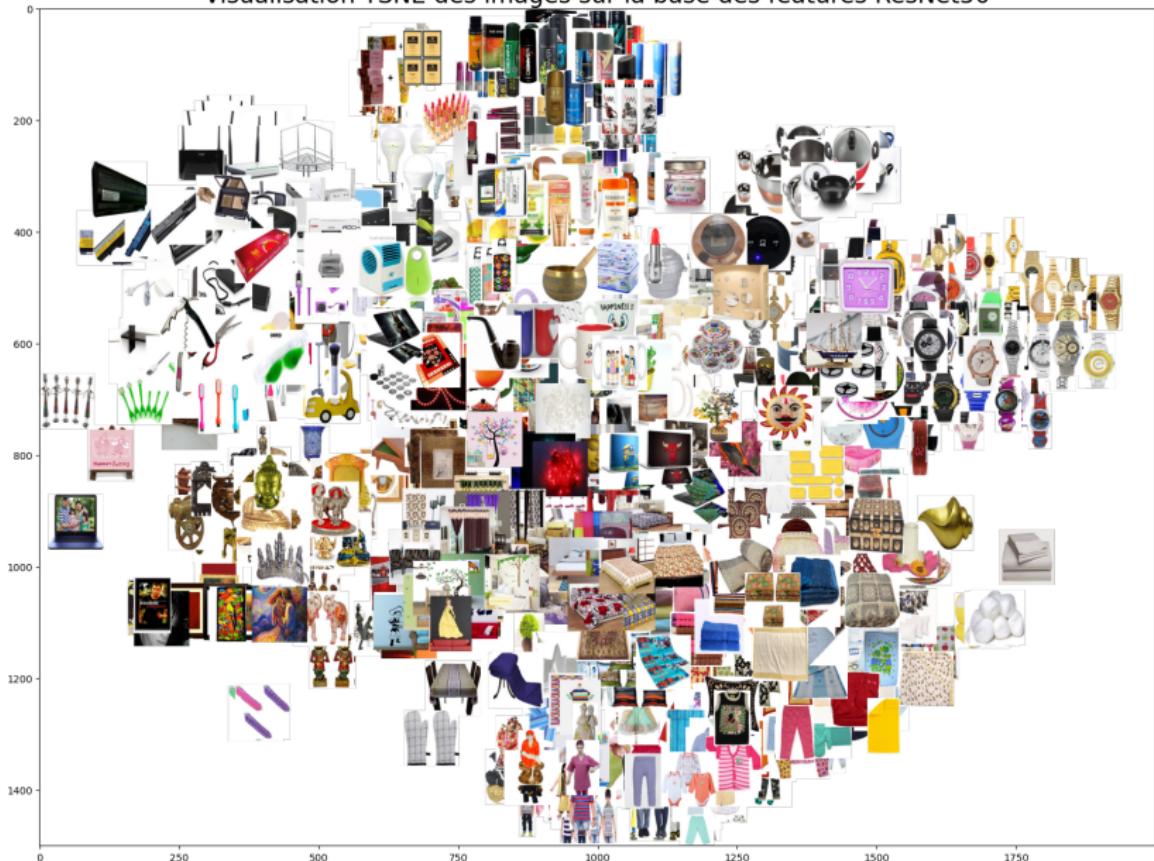
Matrice de confusion



- Le score est **ARI : 0.4762**,
- Les catégories ne sont pas bien séparées.

Approche CNN : Regroupement des images TSNE

Visualisation TSNE des images sur la base des features ResNet50

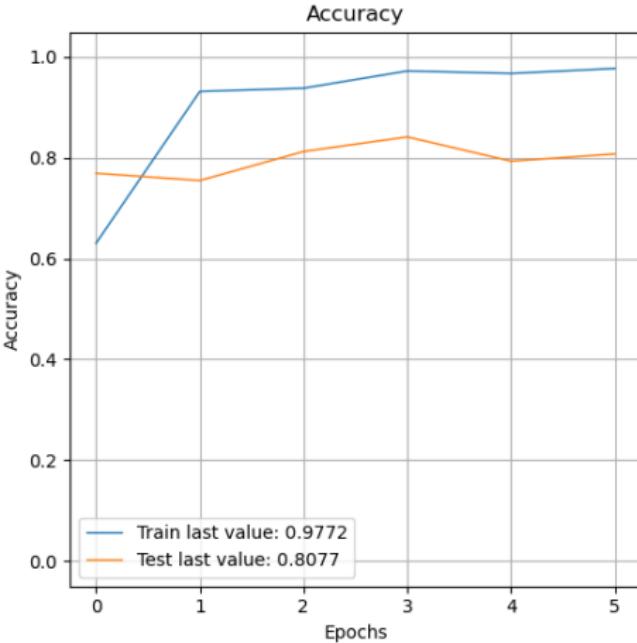
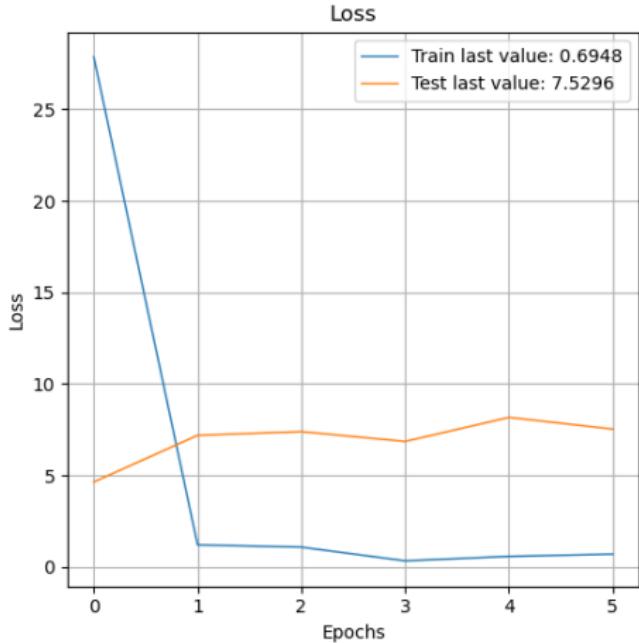


Plan de la présentation

- 1 Problématique
- 2 Nettoyage des données
- 3 Méthodologie
- 4 Données Textuelles
- 5 Données images
- 6 Classification supervisée



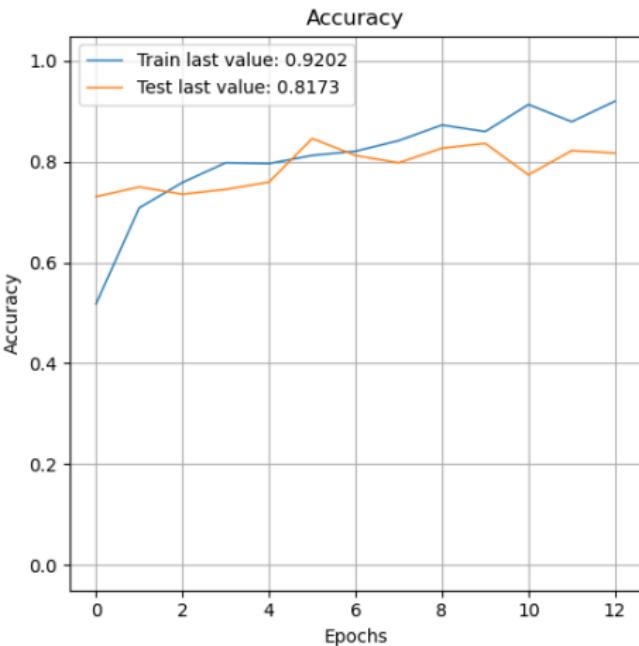
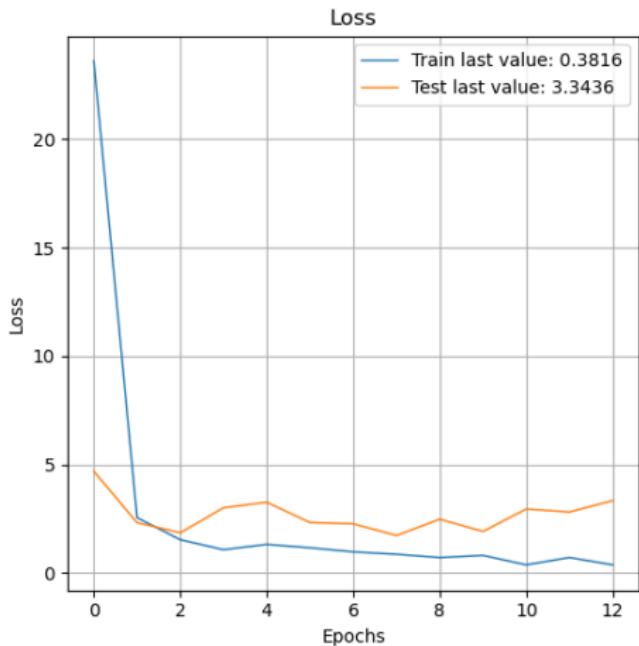
Modèle pré-entraîné VGG16



Réultats du best model : Model1_VGG16.h5

[INFO] Test Loss: 6.7164
[INFO] Test accuracy: 77.62%

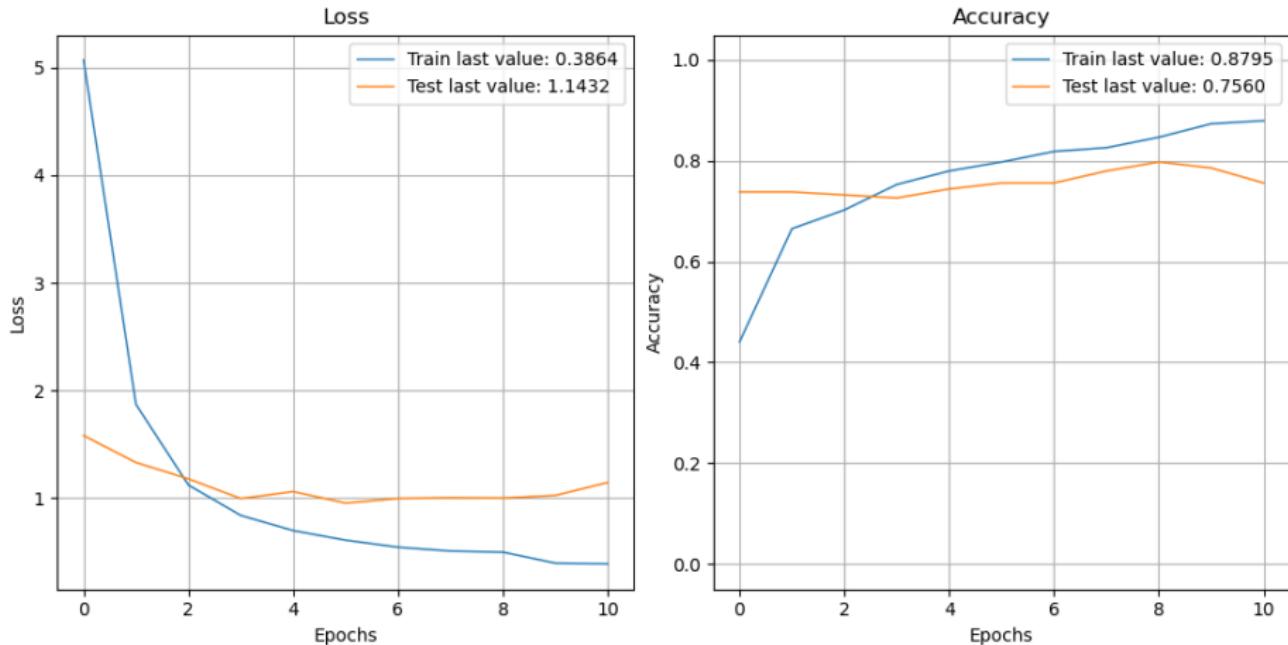
Data augmentation : flow_from_directory



Resultats du best model : Model2_VGG16_augmented.h5

[INFO] Test Loss: 2.3316
[INFO] Test accuracy: 77.14%

Data augmentation : intégrée dans le modèle



Réultats du best model : Model3_VGG16_aug.h5

[INFO] Test Loss: 1.0900
[INFO] Test accuracy: 80.00%

Bilan data augmentation

	Model	Accuracy	Loss	Time_min
0	Model1_VGG16.h5	77.62%	6.716	15.57
0	Model2_VGG16_augmented.h5	77.14%	2.332	31.57
0	Model3_VGG16_aug.h5	80.00%	1.090	25.11

- La data augmentation est permis d'améliorer la performance du modèle. En effet, nous remarquons des performances similaires sur les données **d'entraînement** et de **test**.
- Donc, il est en mesure de **generaliser sur de nouvelles données**.



Conclusion

- ☞ Le moteur de classification est **possible**.
- ☞ Les résultats sont **meilleurs** avec l'**approche CNN**.
- ☞ La data augmentation **améliore** les résultats de la **classification supervisée**.



MERCI POUR VOTRE
AIMABLE ATTENTION

