PROJET 6 :

Classifiez automatiquement des biens de consommation

Soutenance de projet, octobre 2023

Mauyves NKONDO

Data Scientist

Sommaire

Problématique

- Rappel de la problématique
- Présentation du jeu de données

OPENCLASSROOMS

Rappel de la problématique

- **Contexte** : “ Place de marché ” fait référence à une plateforme de commerce en ligne.

- **Moyen** : Consiste à automatiser l'assignation des catégories aux articles.

- **Objectif** : est d'augmenter la convivialité pour les utilisateurs et d'accroître la fiabilité de la catégorisation.

- **But du projet** : **étudier la faisabilité** de cette catégorisation :
  - Extraction de données
  - Analyse et prétraitement du jeu de données : visuelles / textuelles
  - Clustering

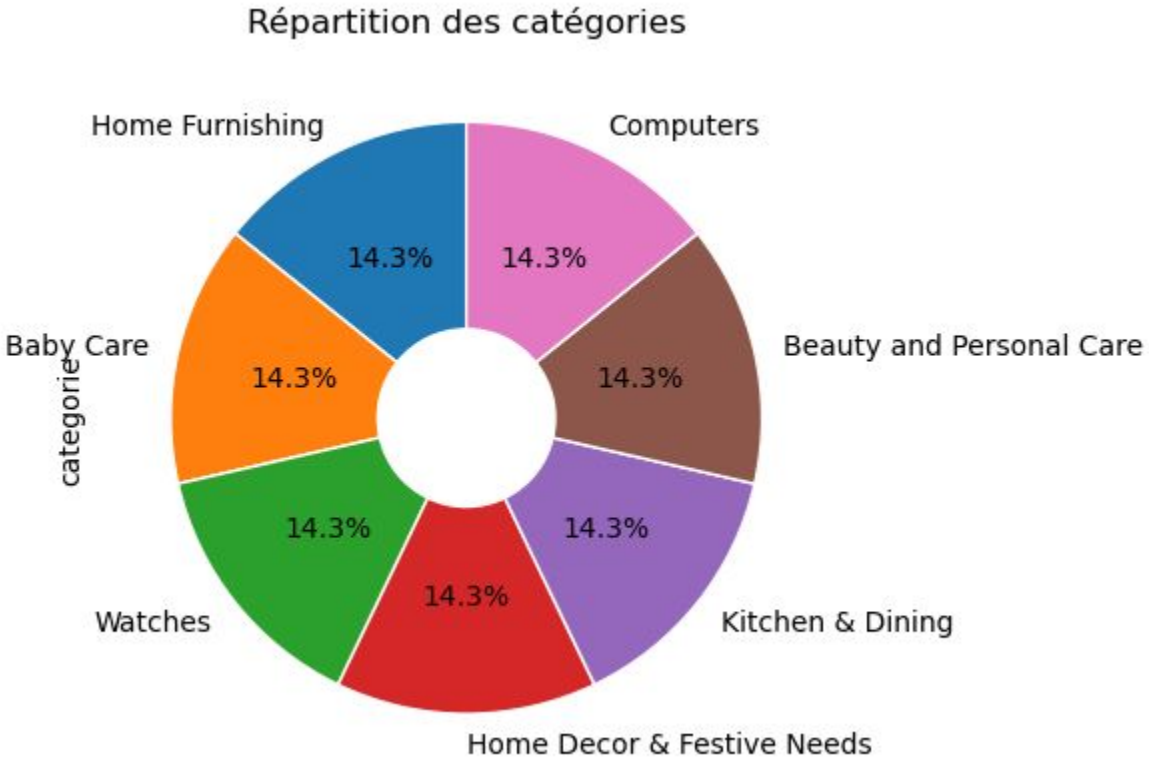
Etude de faisabilité : Processus

Prétraitement des données
- Données textuelles
- Données visuelles

Essais de classification non supervisée
- Données textuelles
- Données visuelles

Essais de classification supervisée
- Données visuelles

Jeu de données



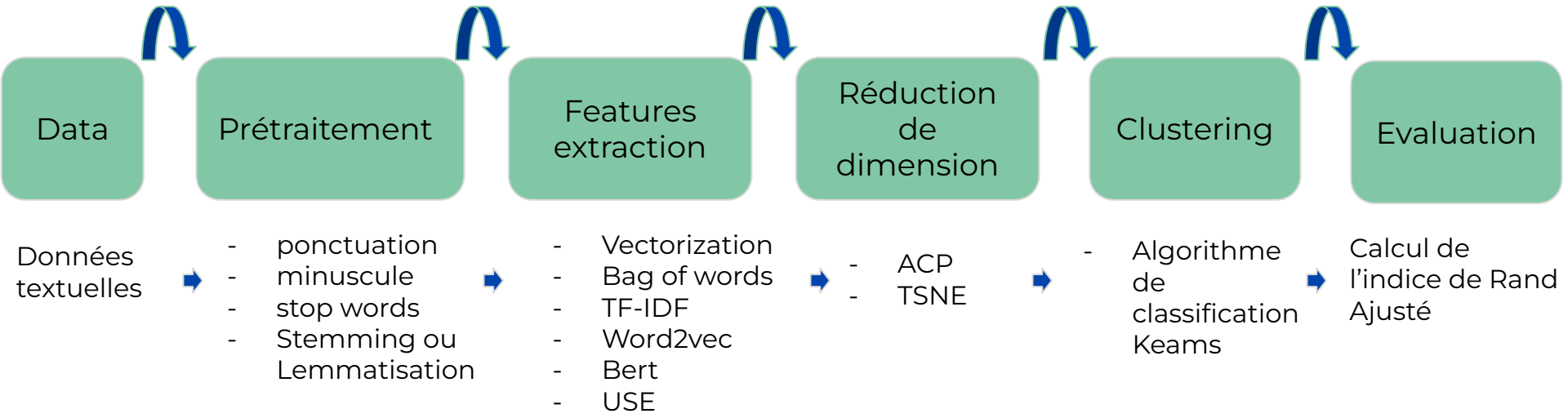
On dispose d'un ensemble de données comprenant 1050 produits répartis en 7 catégories, avec chacune contenant 150 produits et nous créons un diagramme en camembert qui illustre la distribution des diverses catégories, affichant les pourcentages sur chaque portion pour indiquer la part de chaque catégorie.

Traitement et Clusters

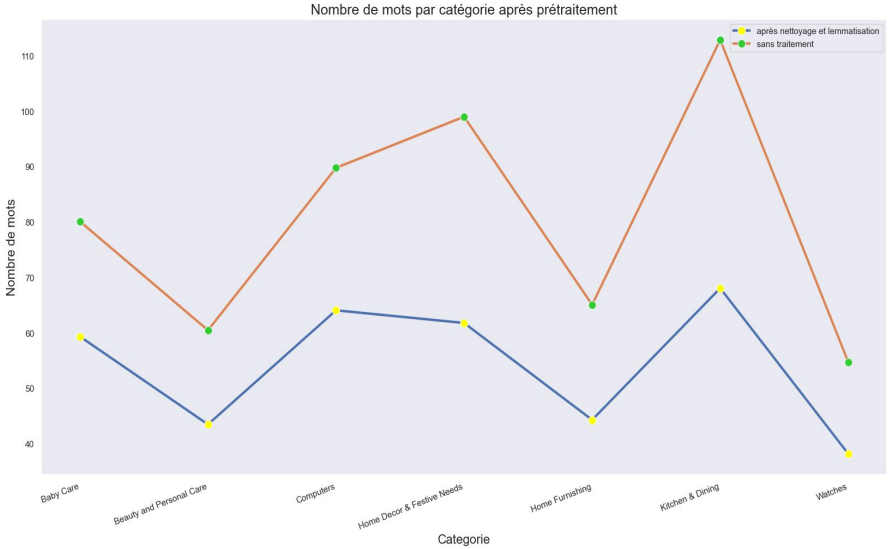
- Données textuelles
  - Les 3 embeddings (Word2Vec, Bert, Use)
- Données visuelles
- VGG 16

Données textuelles : Traitement

- Comment fonctionne le processus de classification des données textuelles ?



Données textuelles : Traitement - Suite



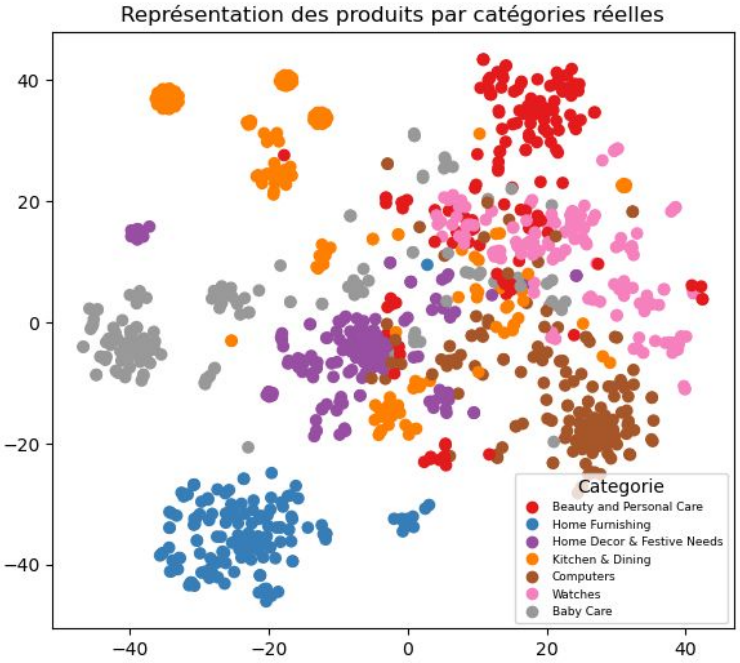
En moyenne, le processus de nettoyage a permis de supprimer 33 % des mots superflus de nos descriptions de produits.

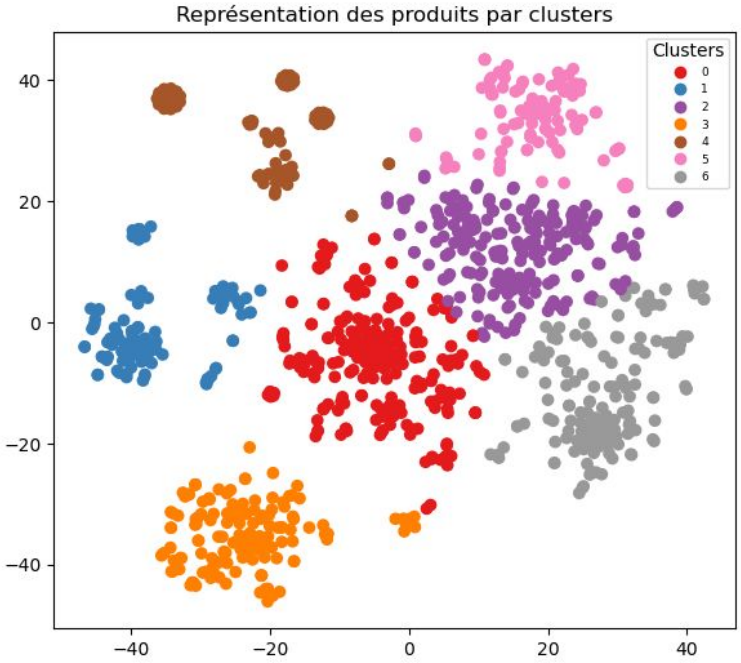
Selon l'analyse de nuage de mots des données Flipkart, les termes les plus fréquemment observés sont : "genuine" (authentique), "product" (produit), "shipping" (expédition), "price" (prix) ...

Données textuelles : Traitement - **Bag of Words**

Les algorithmes d'apprentissage automatique ne peuvent pas traiter directement le texte brut ; il est nécessaire de le convertir en des vecteurs de longueur fixe bien définis.

ARI : 0.4597

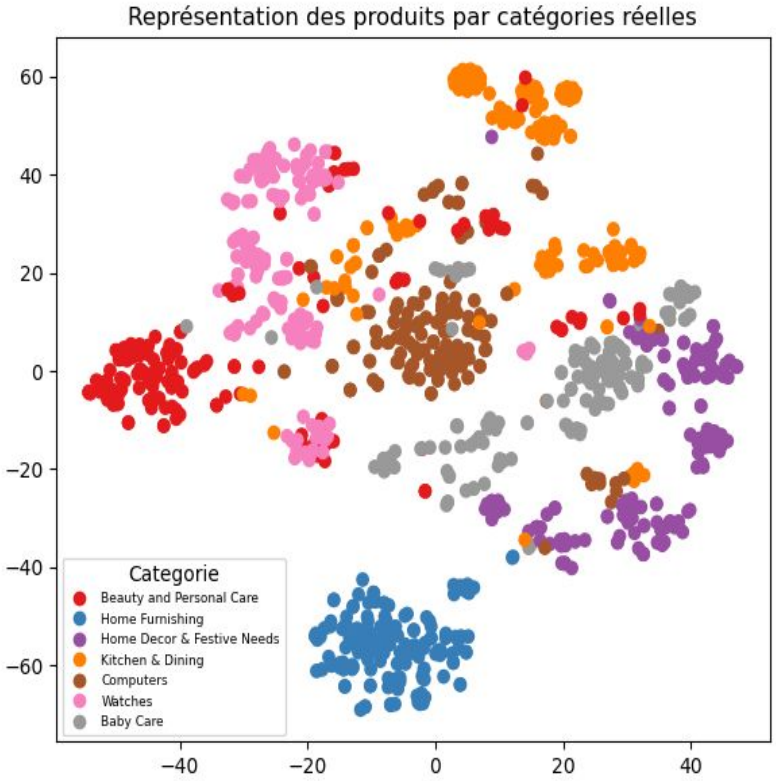


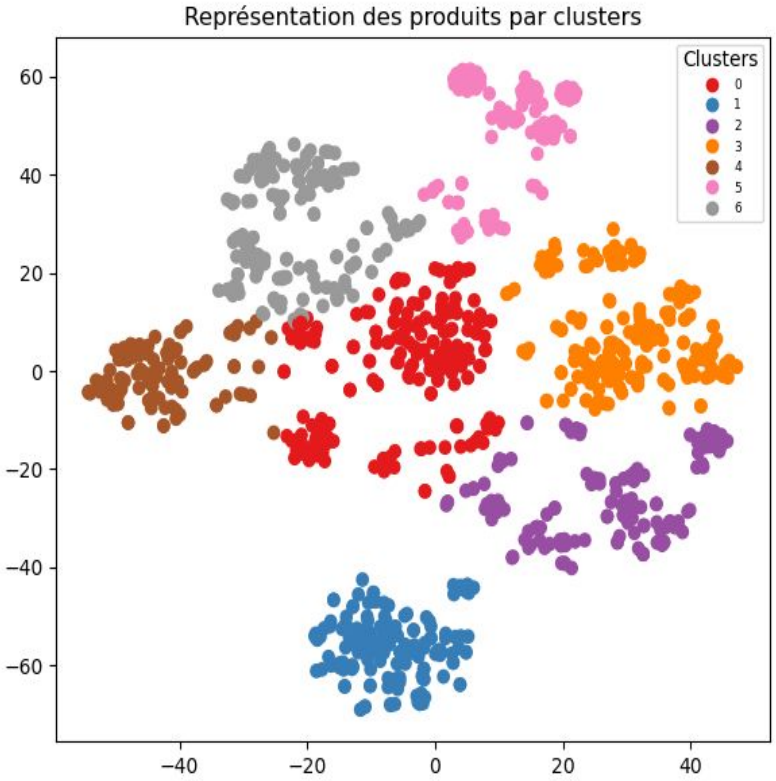


Les catégories sont clairement distinctes, mais notre score ARI est faible.

Données textuelles : Traitement - **TF - IDF**

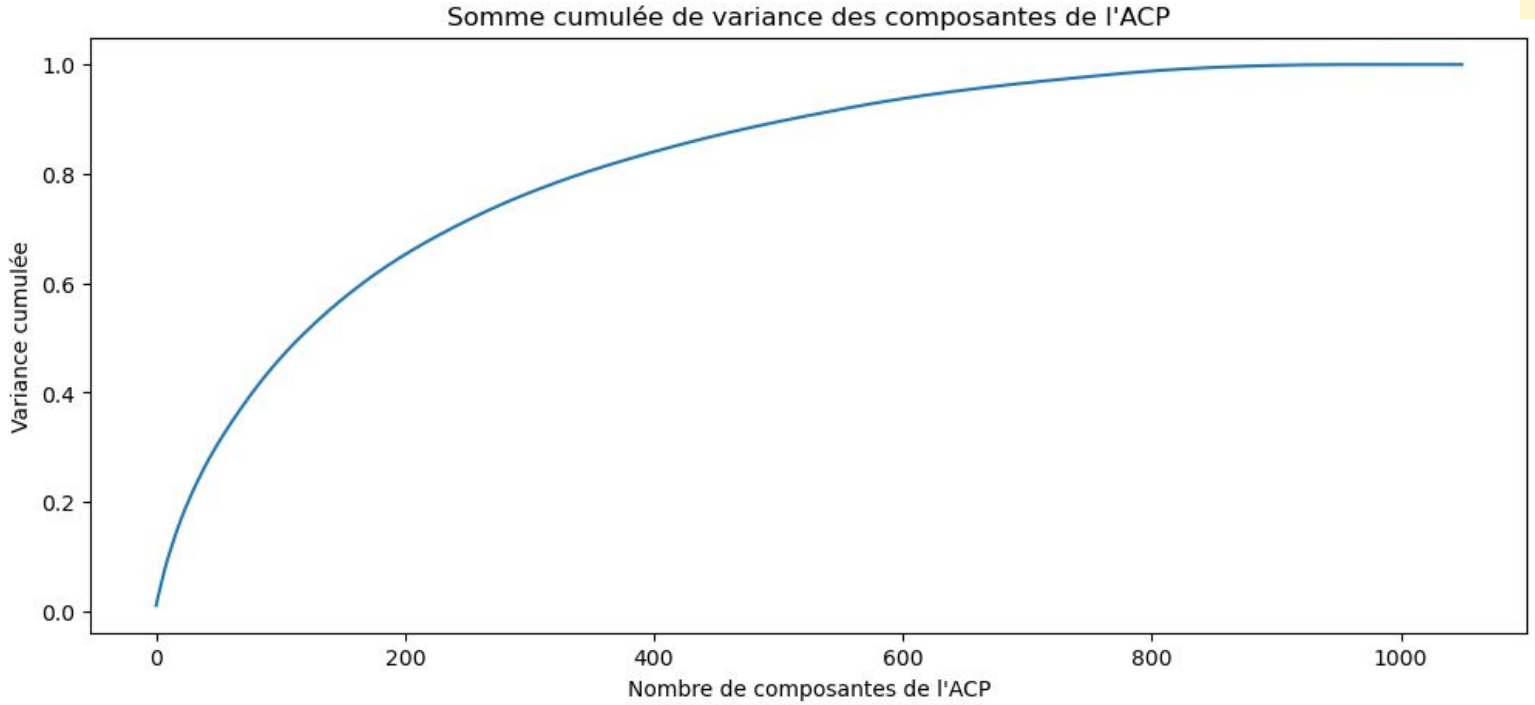
ARI : 0.437





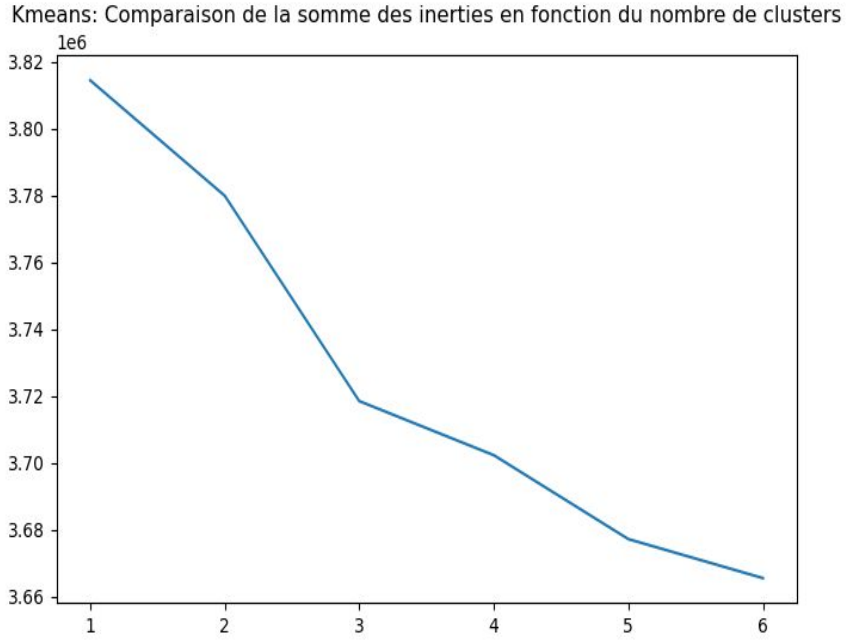
Les catégories sont clairement distinctes, mais notre score ARI est encore faible que celui du Bag of Words.

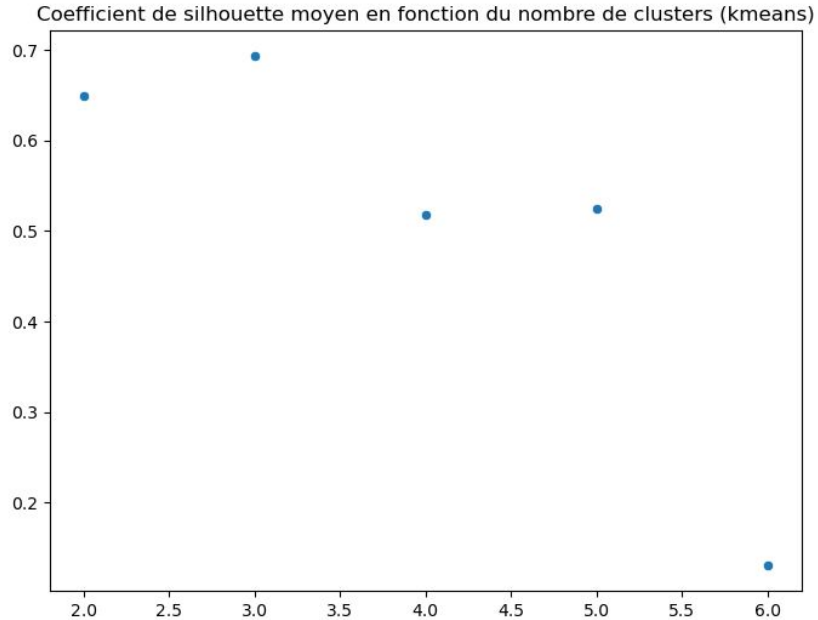
Classifieurs non supervisés sur données textuelles - **ACP**



Ce graphique montre comment la variance cumulée évolue à mesure que l'on augmente le nombre de composantes de l'Analyse en Composantes Principales (ACP). Cela permet de visualiser combien d'informations sont capturées par un nombre croissant de composantes.

Données textuelles - **Catégorisation**



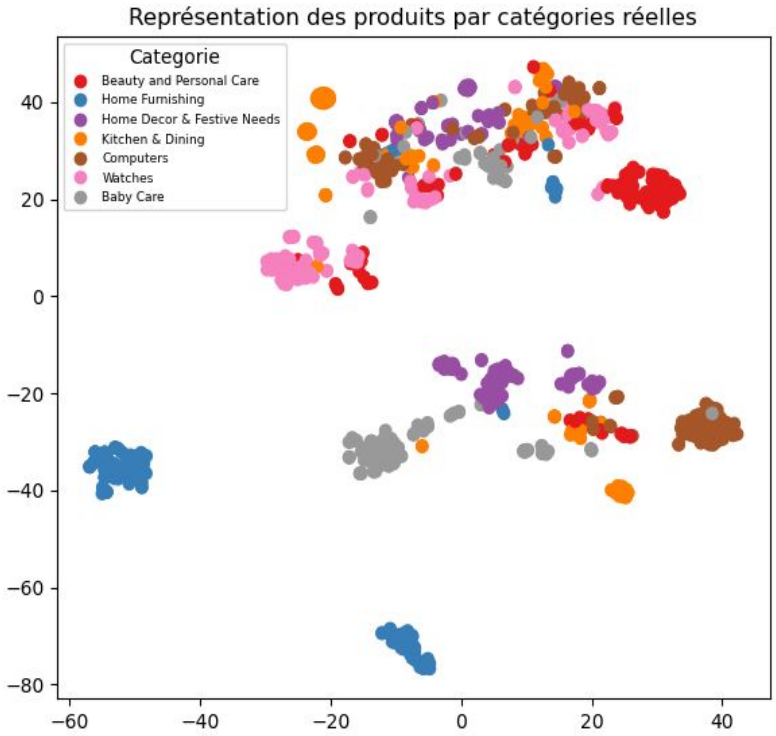


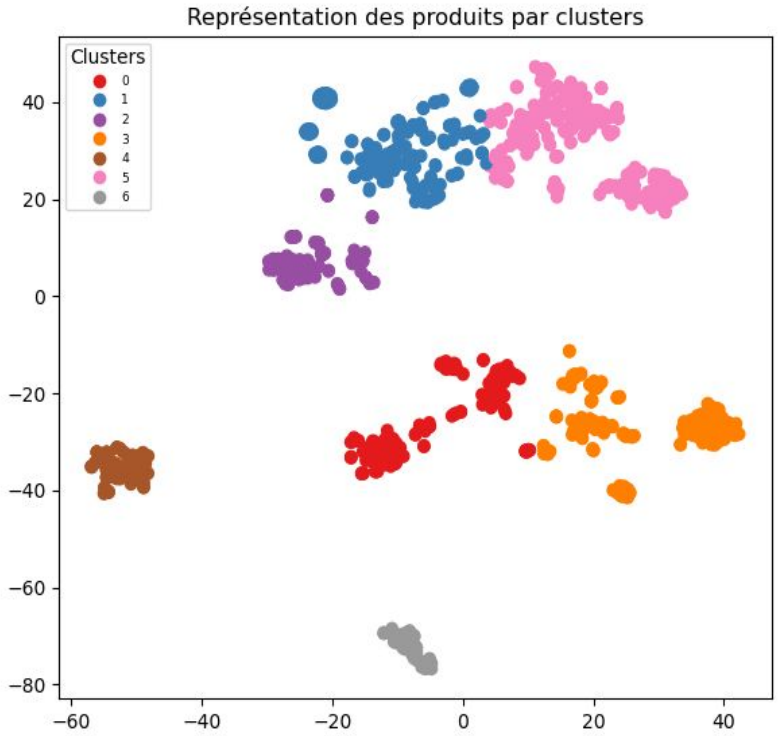
Les résultats exécute l'analyse K-Means pour différents nombres de clusters de 1 à 6. Les valeurs d'inertie ou somme des carrés des distances des points au centre du cluster le plus proche sont calculées pour chaque modèle K-Means, ce qui permettra d'évaluer la qualité de la segmentation. Une barre de progression est utilisée pour suivre l'avancement de l'analyse.

Ce graphique de dispersion représente ces coefficients en fonction du nombre de clusters. Les coefficients de silhouette donnent une indication de la qualité de la segmentation des données en clusters, et ce graphique permet de visualiser comment cette qualité varie en fonction du nombre de clusters.

Données textuelles - **Word2Vec**

ARI : 0.2192

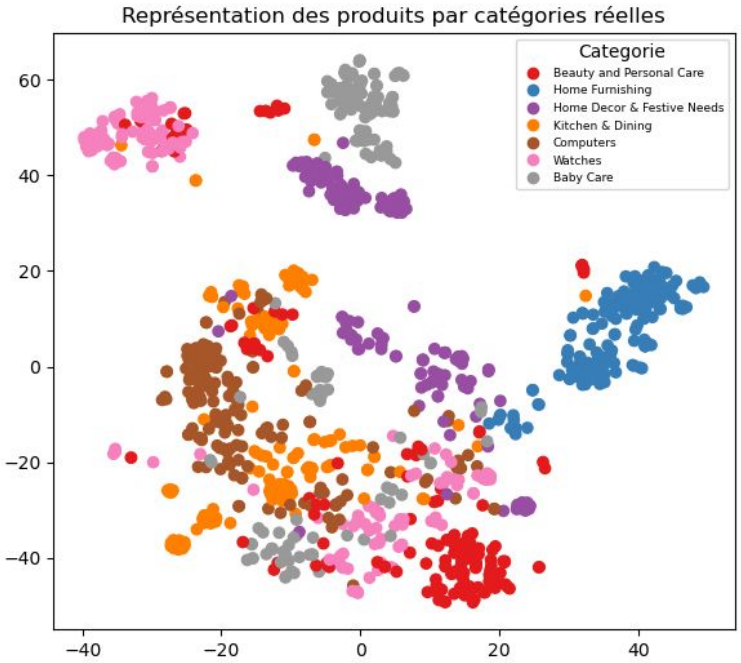


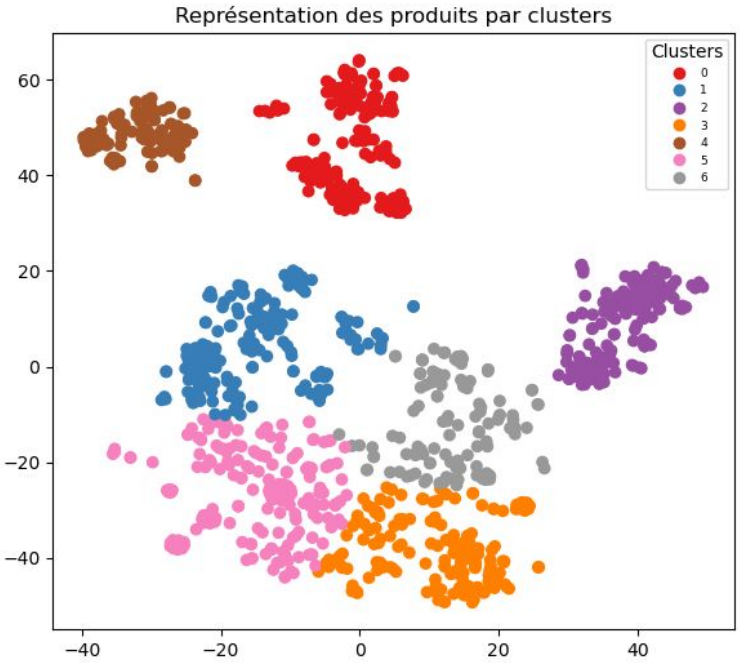


Le clustering a clairement distingué les catégories, mais nous observons un score ARI très bas.

Données textuelles - **Bert (** Bidirectional Representations from Transformers **)**

ARI : 0.327

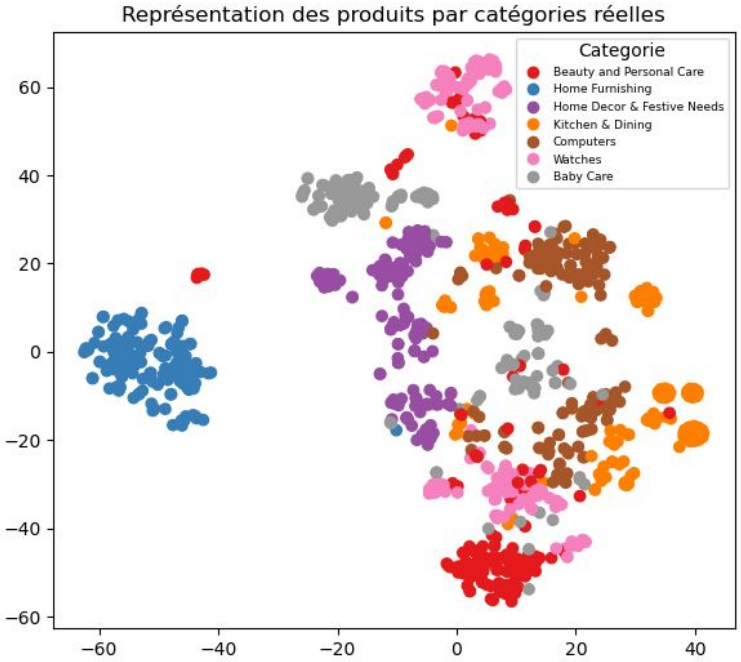


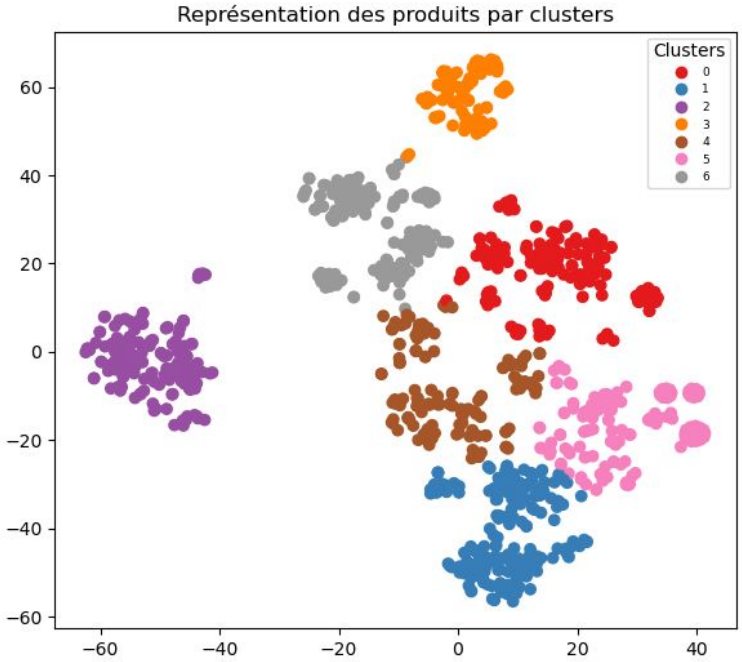


Les catégories sont fortement distinctes, cependant, le Score ARI est peu satisfaisant.

Données textuelles - **USE (** Universal Sentence Encoder **)**

ARI : 0.4219





Les catégories sont clairement distinctes, et notre score ARI n'est pas loin à celui obtenu avec la méthode Bag of Words.

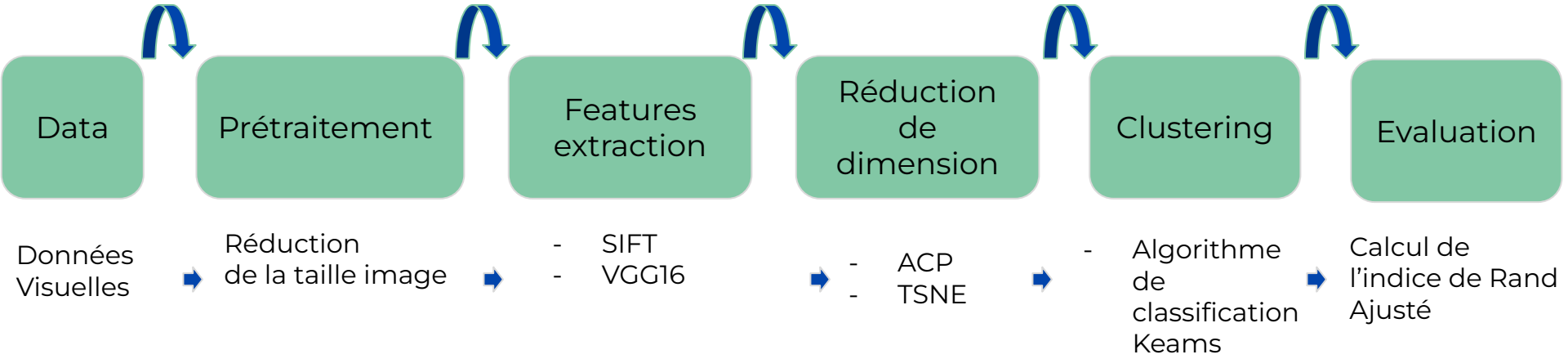
Données textuelles - **Comparaison des Scores ARI**

Modèle	ARI Score
Bag of Word	0.4597
TF - IDF	0.437
Word2Vec	0.2192
Bert	0.327
USE	0.4219

Dans notre cas, le modèle le plus performant est le Bag of Words.

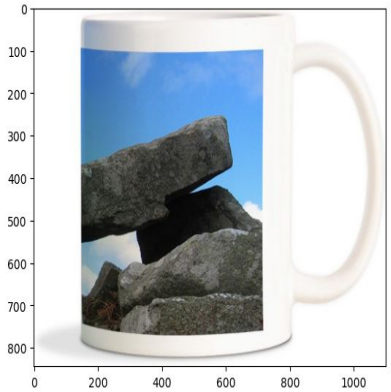
Données Visuelles : Traitement

- Comment fonctionne le processus de classification d'images ?

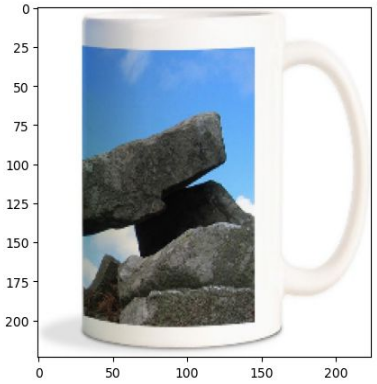


Données Visuelles - **SIFT**

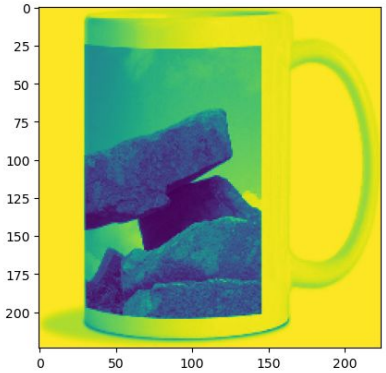
Image Original



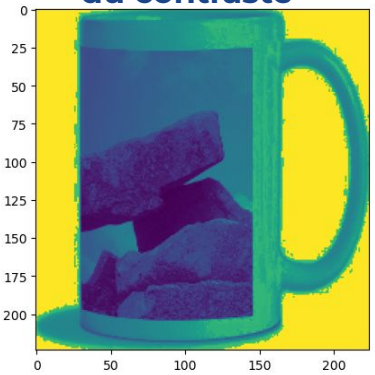
Redimensionnement

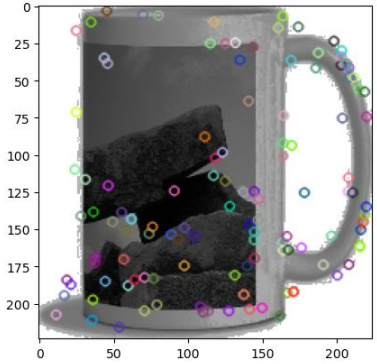


Passage de l'image en gris



Amélioration du contraste

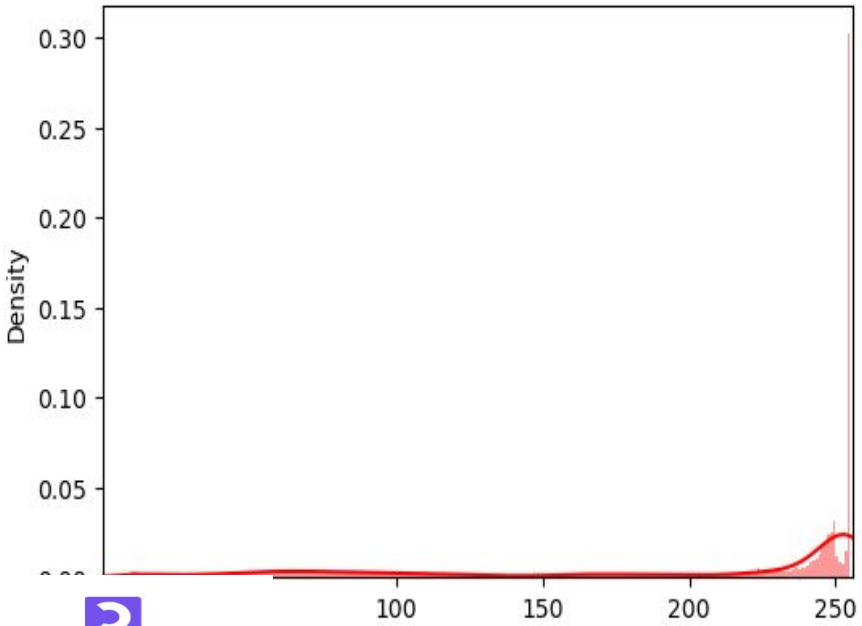


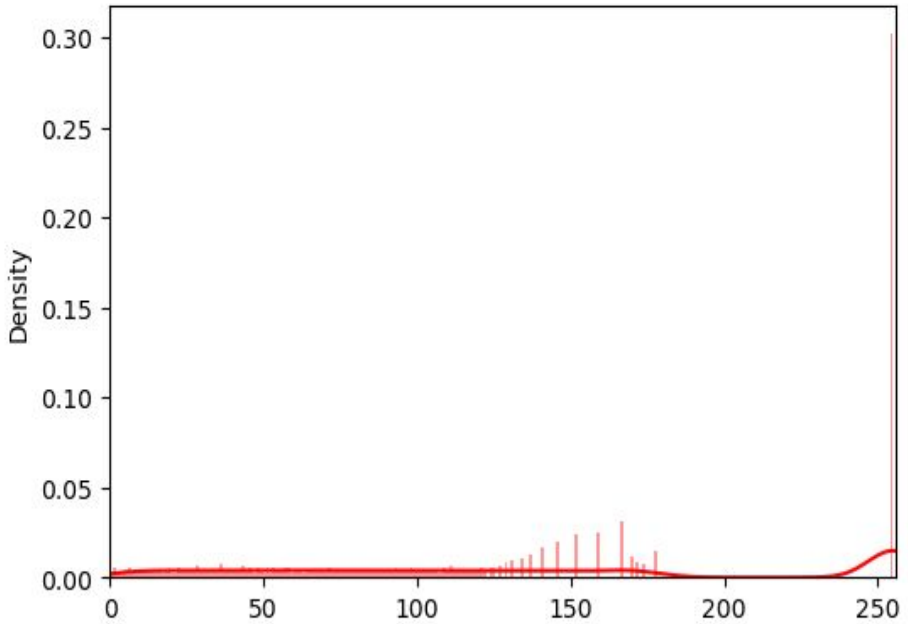


Détection et extraction des features avec **SIFT**

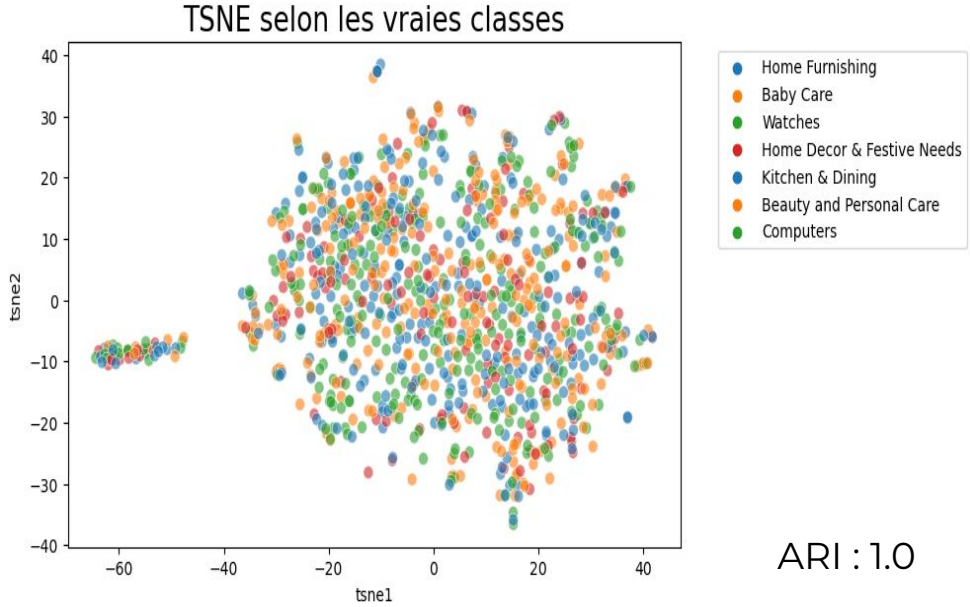
Données Visuelles - **SIFT**

Ces histogrammes basé sur les valeurs d'une image pour visualiser la distribution des niveaux de luminosité dans une image, ce qui peut être important dans le traitement d'images et l'analyse d'images pour diverses applications.

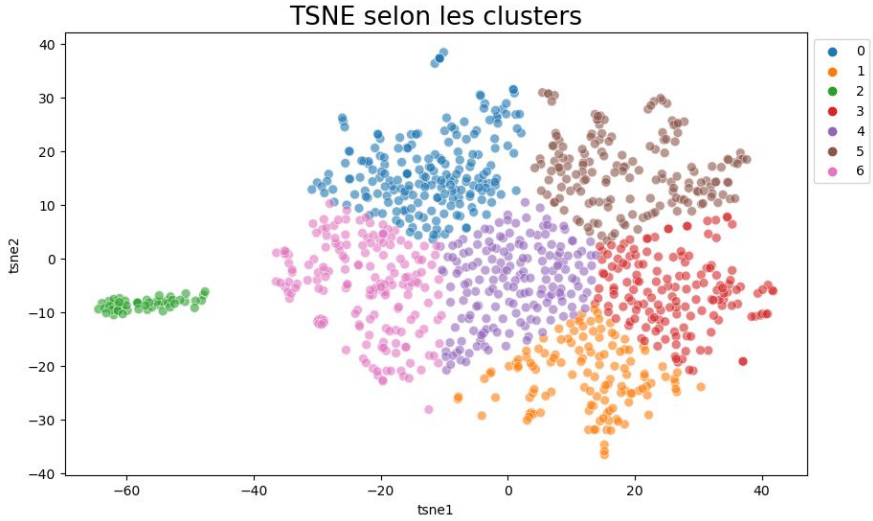




Données Visuelles - **SIFT -** Selon TSNE



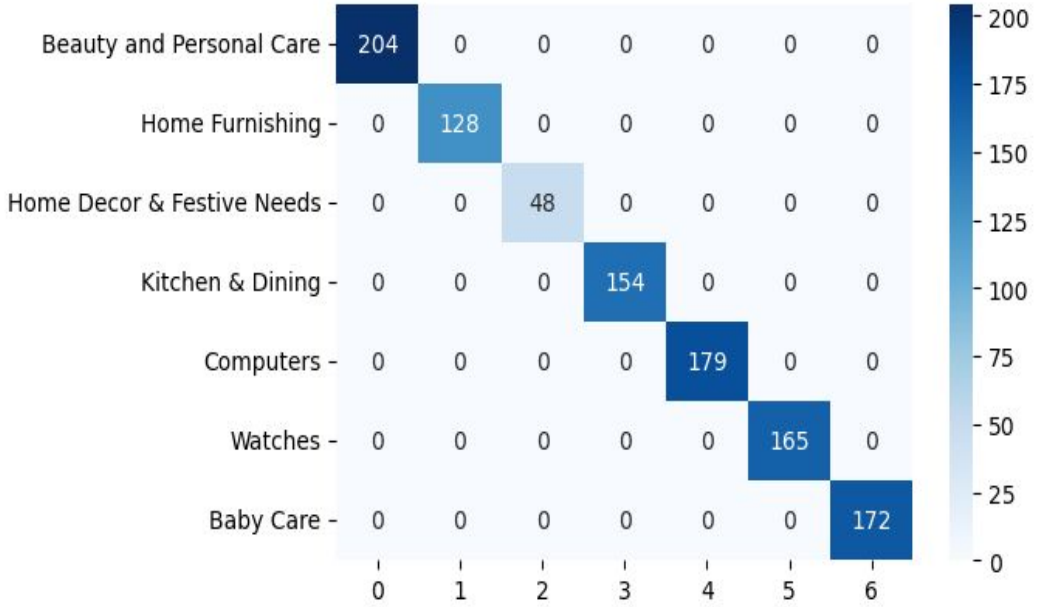
ARI : 1.0



Le score **ARI** est satisfaisant, et les catégories sont regroupées.

Données Visuelles - **SIFT -** Matrice de confusion

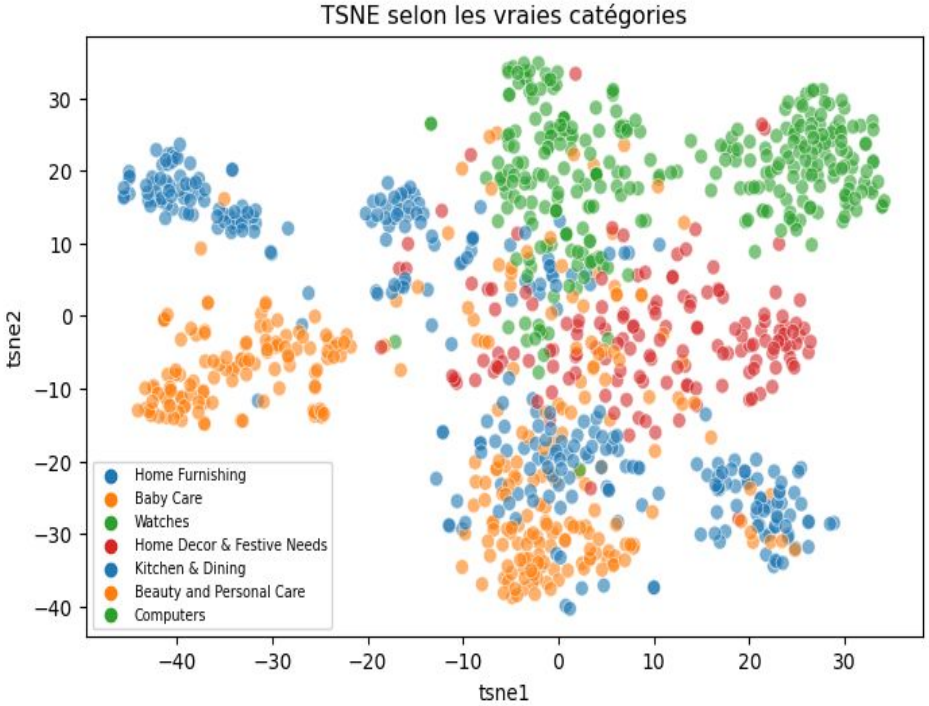
```
Correspondance des clusters :  [0 1 2 3 4 5 6]
[[204   0   0   0   0   0   0]
 [  0 128   0   0   0   0   0]
 [  0   0  48   0   0   0   0]
 [  0   0   0 154   0   0   0]
 [  0   0   0   0 179   0   0]
 [  0   0   0   0   0 165   0]
 [  0   0   0   0   0   0 172]]

              precision    recall  f1-score   support

           0       1.00      1.00      1.00       204
           1       1.00      1.00      1.00       128
           2       1.00      1.00      1.00        48
           3       1.00      1.00      1.00       154
           4       1.00      1.00      1.00       179
           5       1.00      1.00      1.00       165
           6       1.00      1.00      1.00       172

    accuracy                           1.00      1050
   macro avg       1.00      1.00      1.00      1050
weighted avg       1.00      1.00      1.00      1050
```



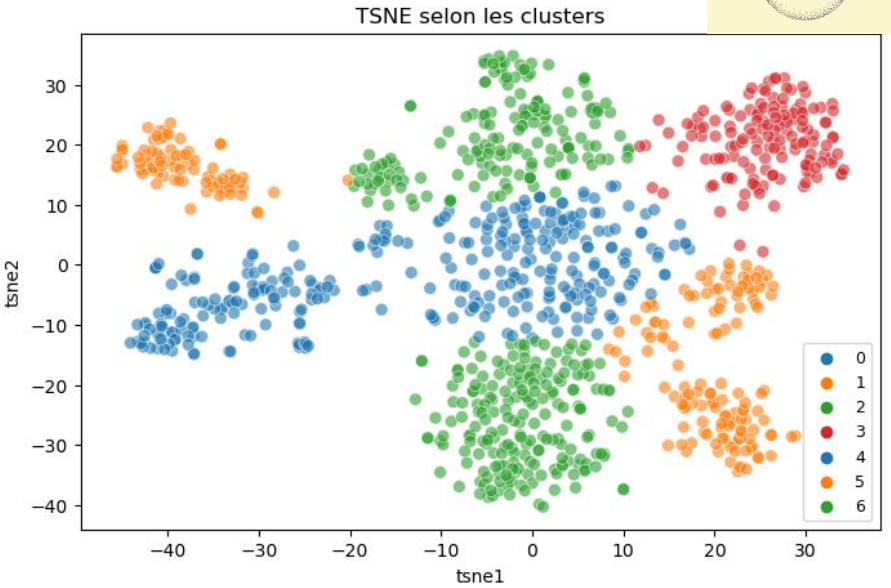
Cette heatmap pour visualiser graphiquement les performances du modèle de classification. Elle affiche le nombre d'observations correctement classées et incorrectement classées pour chaque paire d'étiquettes réelles et prédites.

- F1-score = 1, excellent
- Le score Accuracy est également excellent.

Données Visuelles - **VGG16**



Cette visualisation en 2D des données t-SNE, permet de voir comment les données sont distribuées dans l'espace t-SNE tout en maintenant la référence aux catégories réelles pour une meilleure compréhension.

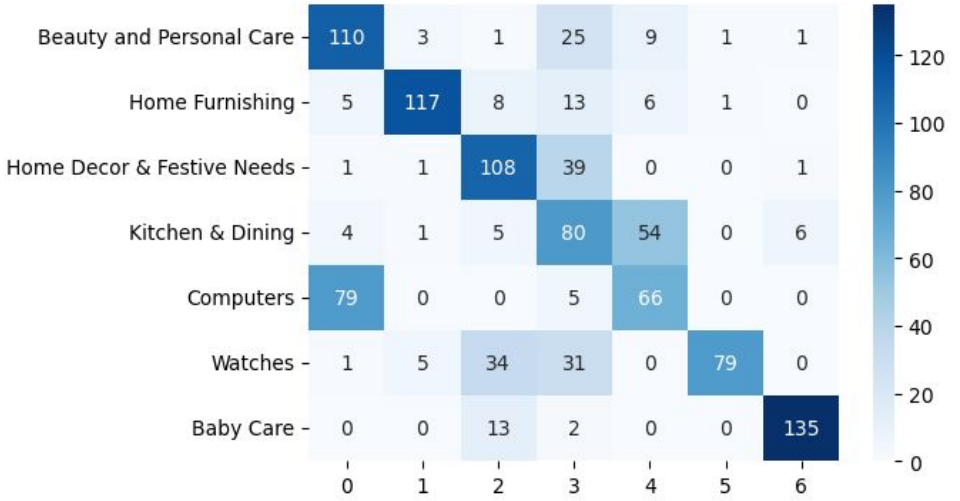


- ARI : 0.4496

L'ARI mesure à quel point les clusters correspondent aux vraies catégories des données.

Données Visuelles - **VGG 16 -** Matrice de confusion

```
Correspondance des clusters :  [1 4 2 6 3 5 0]
[[110   3   1  25   9   1   1]
 [  5 117   8  13   6   1   0]
 [  1   1 108  39   0   0   1]
 [  4   1   5  80  54   0   6]
 [ 79   0   0   5  66   0   0]
 [  1   5  34  31   0  79   0]
 [  0   0  13   2   0   0 135]]
              precision    recall  f1-score   support

           0       0.55      0.73      0.63       150
           1       0.92      0.78      0.84       150
           2       0.64      0.72      0.68       150
           3       0.41      0.53      0.46       150
           4       0.49      0.44      0.46       150
           5       0.98      0.53      0.68       150
           6       0.94      0.90      0.92       150

    accuracy                           0.66      1050
   macro avg       0.70      0.66      0.67      1050
weighted avg       0.70      0.66      0.67      1050
```



Cette heatmap pour visualiser graphiquement les performances du modèle de classification. Elle affiche le nombre d'observations correctement classées et incorrectement classées pour chaque paire d'étiquettes réelles et prédites.

- Score Accuracy n'est pas si mal.
- Score f1 variant de 0,46 à 0,92 indique une certaine variabilité dans les performances, mais cela reste un résultat solide.

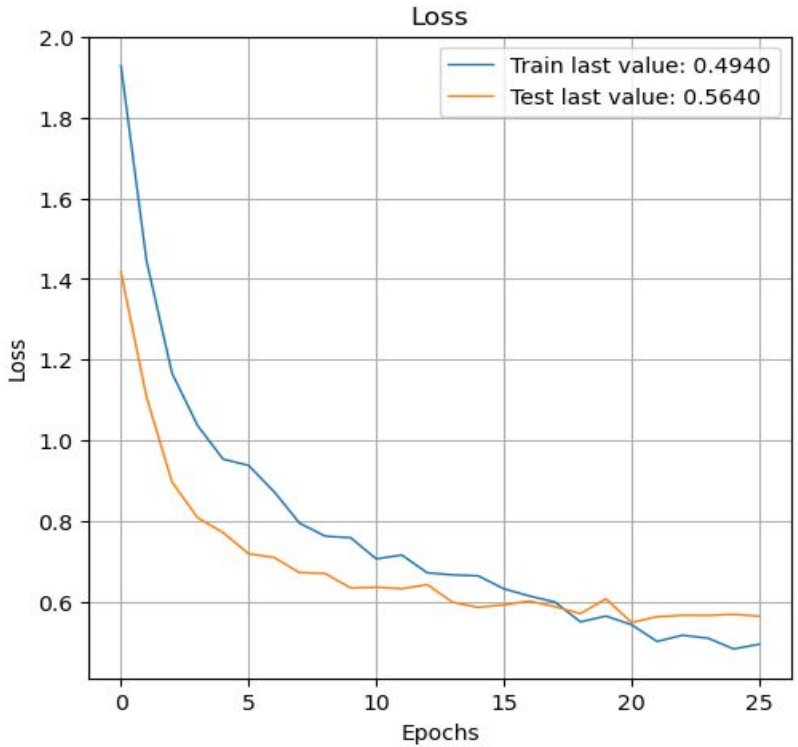
Données Visuelles - **VGG16**

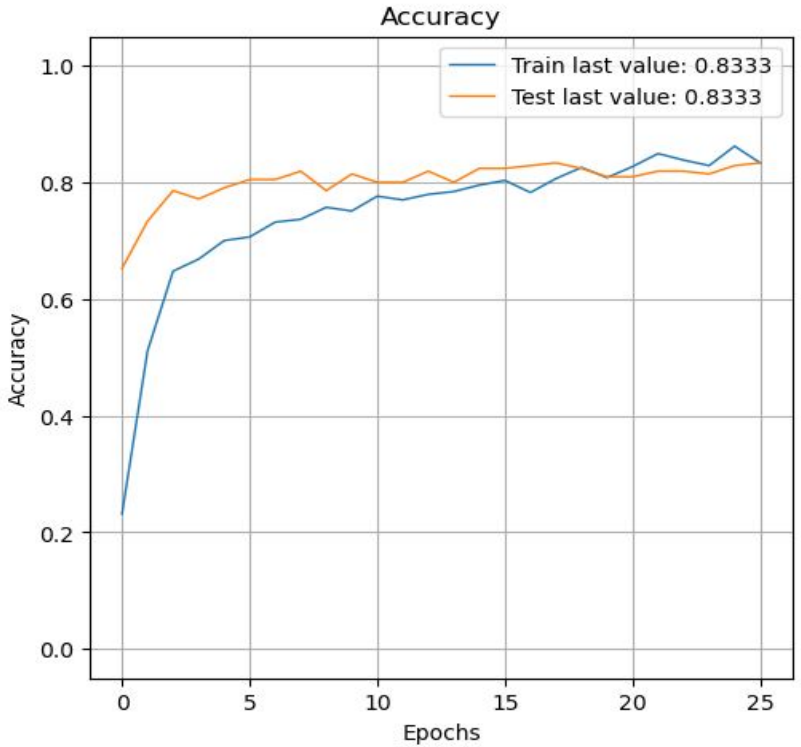
Exemple de confusion

Prédit “ Baby care” au lieu de “ Computers”

Données Visuelles - **VGG16**

VGG16 transfer learning avec data augmentation

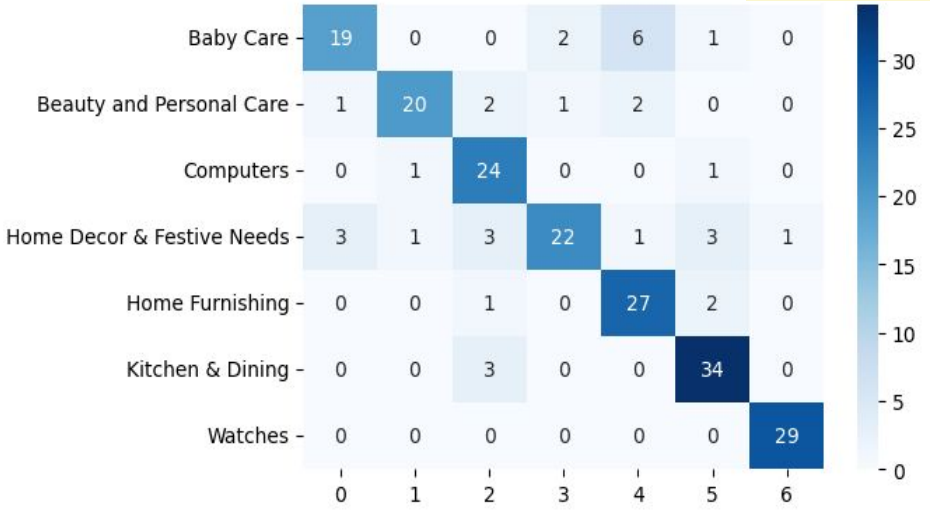




- La perte a diminué et la précision a augmenté, ce qui indique que le modèle en cours d'apprentissage fonctionne correctement.

Données Visuelles - **VGG16**

```
[[19  0  0  2  6  1  0]
 [ 1 20  2  1  2  0  0]
 [ 0  1 24  0  0  1  0]
 [ 3  1  3 22  1  3  1]
 [ 0  0  1  0 27  2  0]
 [ 0  0  3  0  0 34  0]
 [ 0  0  0  0  0  0 29]]
              precision    recall  f1-score   support

           0       0.83      0.68      0.75        28
           1       0.91      0.77      0.83        26
           2       0.73      0.92      0.81        26
           3       0.88      0.65      0.75        34
           4       0.75      0.90      0.82        30
           5       0.83      0.92      0.87        37
           6       0.97      1.00      0.98        29

    accuracy                           0.83       210
   macro avg       0.84      0.83      0.83       210
weighted avg       0.84      0.83      0.83       210
```



Cette heatmap pour visualiser graphiquement les performances du modèle de classification. Les noms de classe sont spécifiés pour les catégories afin d'améliorer la compréhension de la heatmap.

- Une précision de 0,83 est un score satisfaisant.
- Un score F1 variant de 0,75 à 0,98 indique également un bon résultat.

Conclusion

Conclusion

- On a pu obtenir une meilleure représentation des données textuelles en optant pour l'algorithme Bag of Words.

- Pour les données visuelles, on a pu obtenir une caractérisation améliorée en utilisant un algorithme pré-entraîné tel que VGG16.

- Cela démontre notre capacité à prédire les catégories à partir des images fournies par le vendeur.

Merci de votre Attention !