Classifiez automatiquement des biens de consommation

Etudiante : Samira MAHJOUB
Mentor : Nicolas Rangeon
Evaluateur : Zied Jemai
Date : 11/07/2024

Plan de Présentation

- Problématique/ Présentation des données
- Traitement données textuelles
- Traitement données images
- Combinaison textes/images
- Conclusions

« place de marché » souhaite lancer un marketplace e-commerce.

Problématique

L'attribution manuelle des catégories d'articles par les vendeurs est peu fiable et inefficace, surtout avec l'augmentation du volume d'articles.

Objectif

Automatiser la classification des articles en utilisant des descriptions textuelles et des images pour améliorer l'expérience utilisateur.

Mission

Etudier la faisabilité d’une classification automatique à partir des images et des descriptions produites par le vendeur

Présentation des données

- Données issues de la base FlipKart → 1 fichier csv + 1 dossier des images

- 1050 produits

- 15 indicateurs couvrant plusieurs types d'informations:
  - Informations produits
  - Informations tarifaires
  - Notes produits
  - Images produits

- Colonnes très bien renseignées

Ce qui nous intéresse

- ✓ Données textuelles : nom, description
- ✓ Données visuelles : image des produits
- ✓ Les catégories des produits

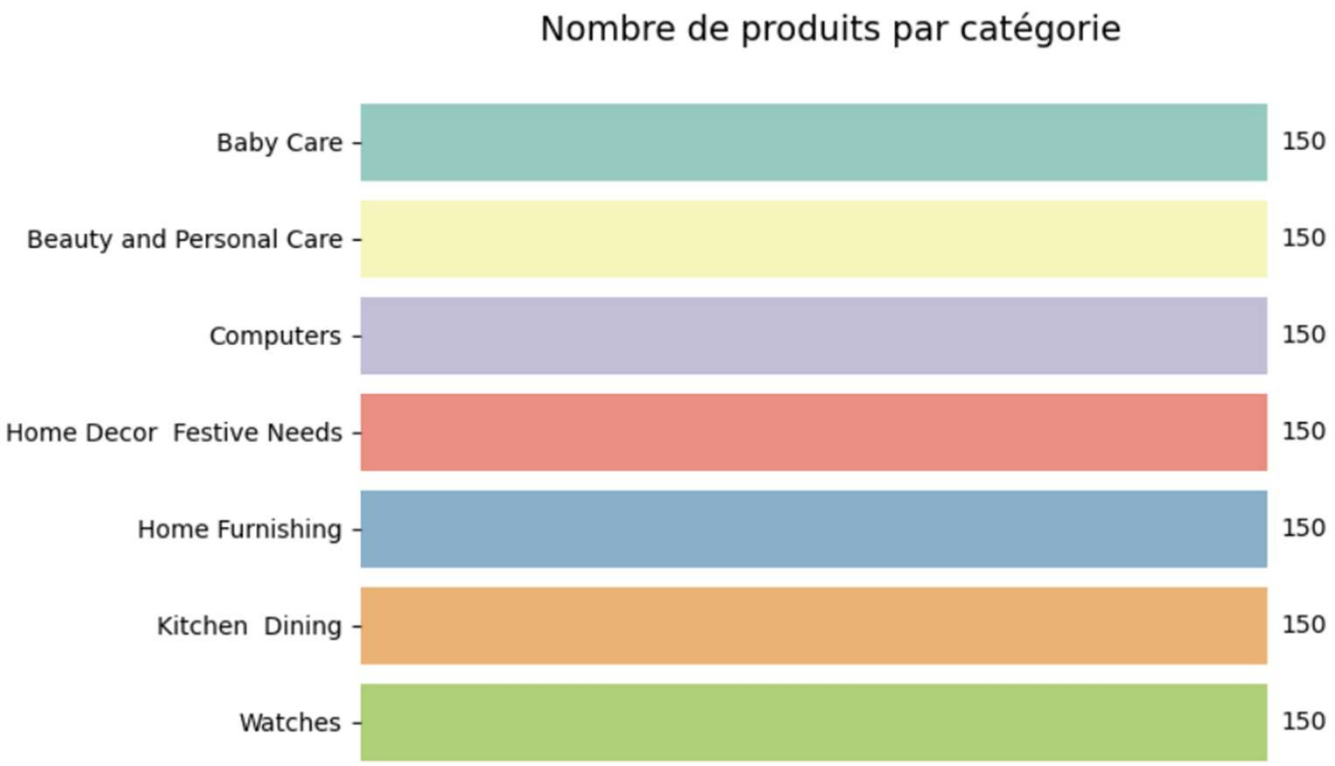
Arbre des catégories des produits

➢ **Nombre de catégories**

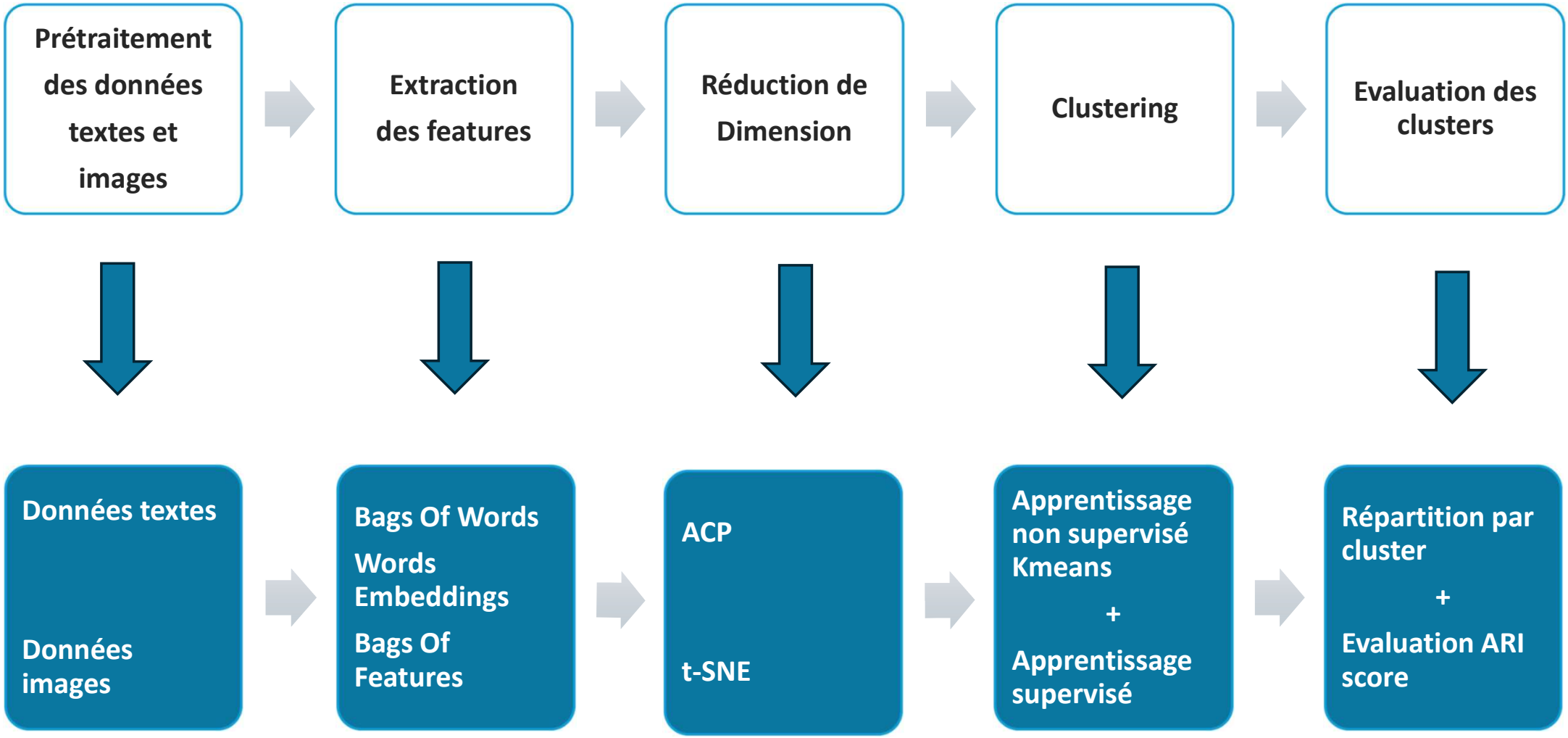
- main_category : 7
- category_1 : 63
- category_2 : 247
- category_3 : 351
- category_4 : 298
- category_5 : 118

➢ **Catégories principales**

- Baby Care
- Beauty and Personal Care
- Computers
- Home Decor Festive Needs
- Home Furnishing
- Kitchen Dining
- Watches



Processus de démarche



Classification des textes

Classification des textes

Prétraitement des données

Natural Langage Preprocessing

3 fonctions utilisées :

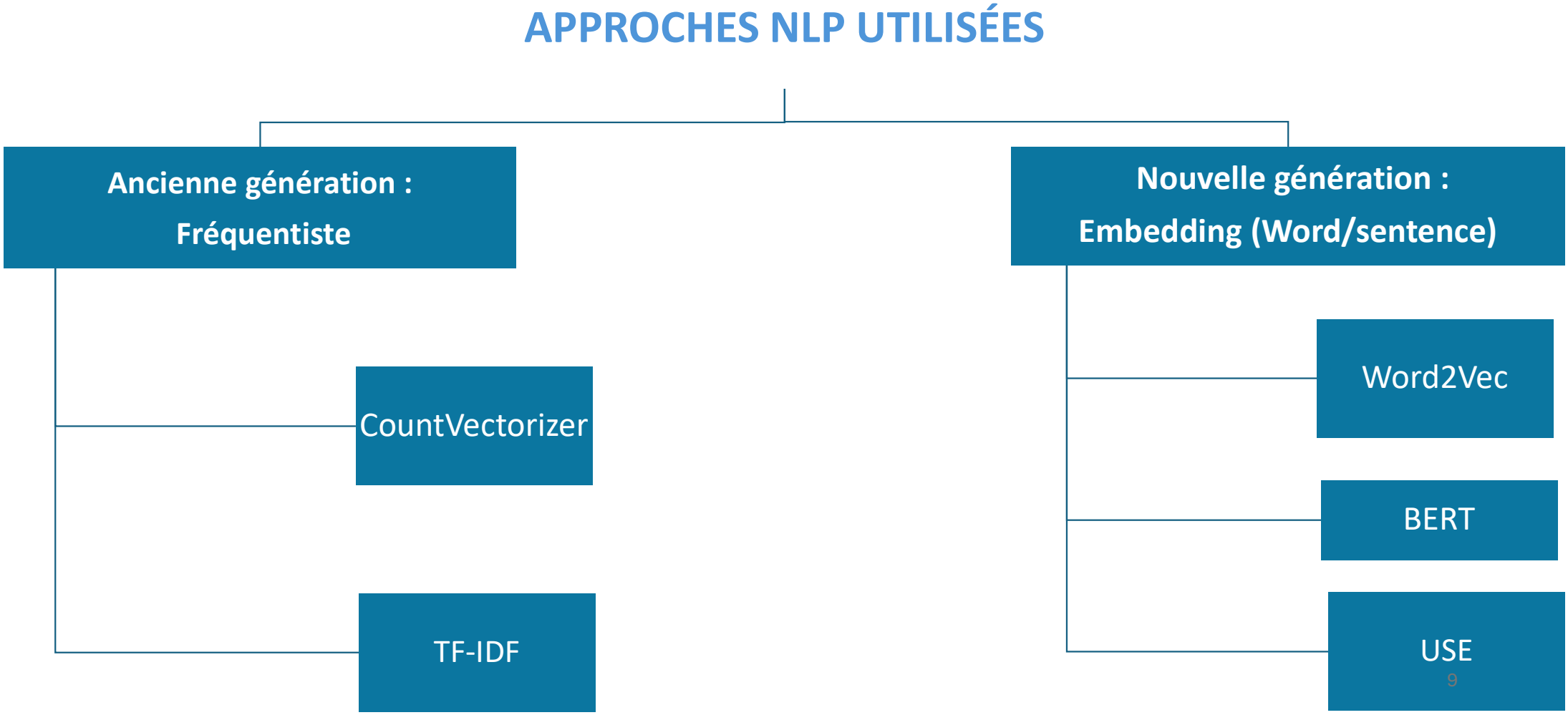
Mapping POS : C'est une étiquette qui indique la catégorie grammaticale (comme nom, verbe, adjectif) d'un mot dans une phrase
Lemmatisation : Transformation des mots à leur forme de base
Stemming : Réduction des mots à leur racine

Exemple de description suite à la lemmatisation :

```
df_category['lemmatizer_cat'].values[0]
```

```
'feature elegance polyester multicolor eyelet door curtain curtain elegance polyester multicolor eyelet door curtain pack
```

Extraction des features



Démarche suivie

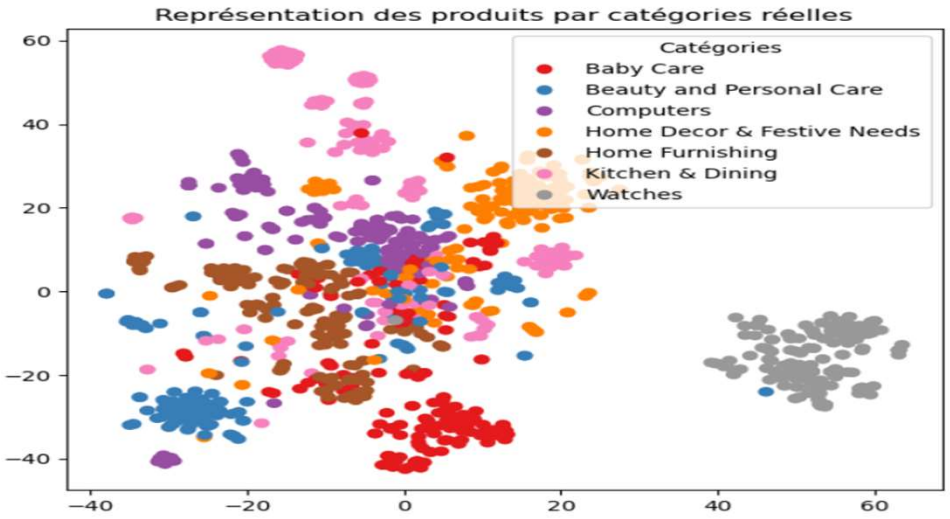
- **DÉMARCHE**
- **Création de caractéristiques** : À partir des noms des produits et des descriptions lemmatisées (sac de mots, vecteurs de mots, etc.).
- **Réduction de dimension** : Utilisation de l'ACP (Test avec LDA), puis T-SNE.
- **Clustering** : Application de l'algorithme K-Means (algorithme non supervisé de clustering qui permet de regrouper les observations du data set en K clusters distincts),**k = 7**
- **Calcul du score ARI** : Pour évaluer la qualité du clustering.
- **Visualisation graphique** : Présentation des résultats sous forme de graphiques.
- **Analyse par classe** : Utilisation d'une matrice de confusion pour évaluer les regroupements.
- **Matrice de confusion :** tableau qui compare les prédictions d'un modèle aux valeurs réelles pour évaluer la performance d'un modèle de classification.

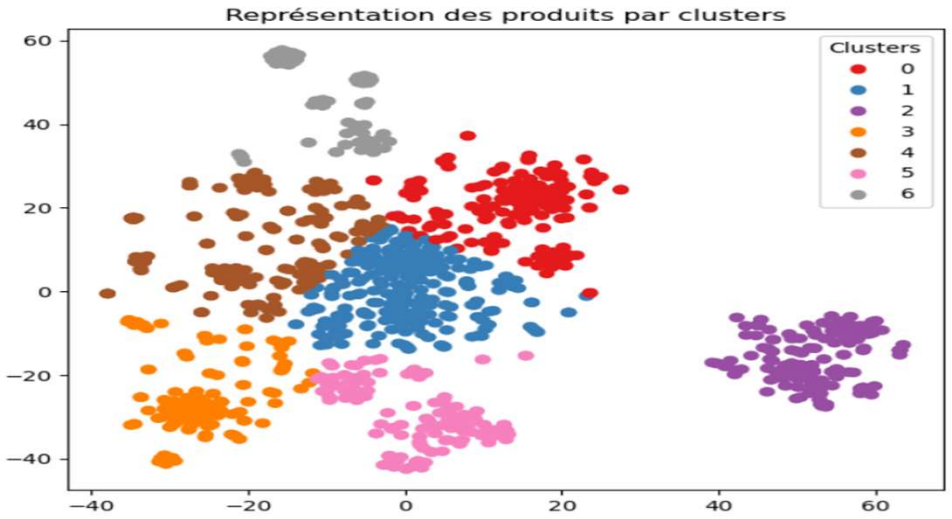
- **MESURE DE PERFORMANCE**
- **Score ARI (Adjusted Rand Index)** :
  - Évalue la concordance entre les regroupements prédits par l'algorithme de clustering et les regroupements de référence.
  - Un score élevé indique une meilleure adéquation des regroupements.

CountVectorizer

CountVectorizer : ***Un outil (classe) de scikit-learn qui implémente la technique Bag of Words pour transformer des textes en vecteurs de comptage de mots.***

Bag of Words : ***Une technique pour transformer des textes en vecteurs de fréquences de mots (compte le nombre de fois qu'un mot apparait dans un document)***





ARI : 0,37 ⟶ **Faible concordance clustering**

Les clusters créés par l'algorithme ne correspondent pas bien aux vraies catégories.

CountVectorizer

Le f1-score macro avg : 0,11

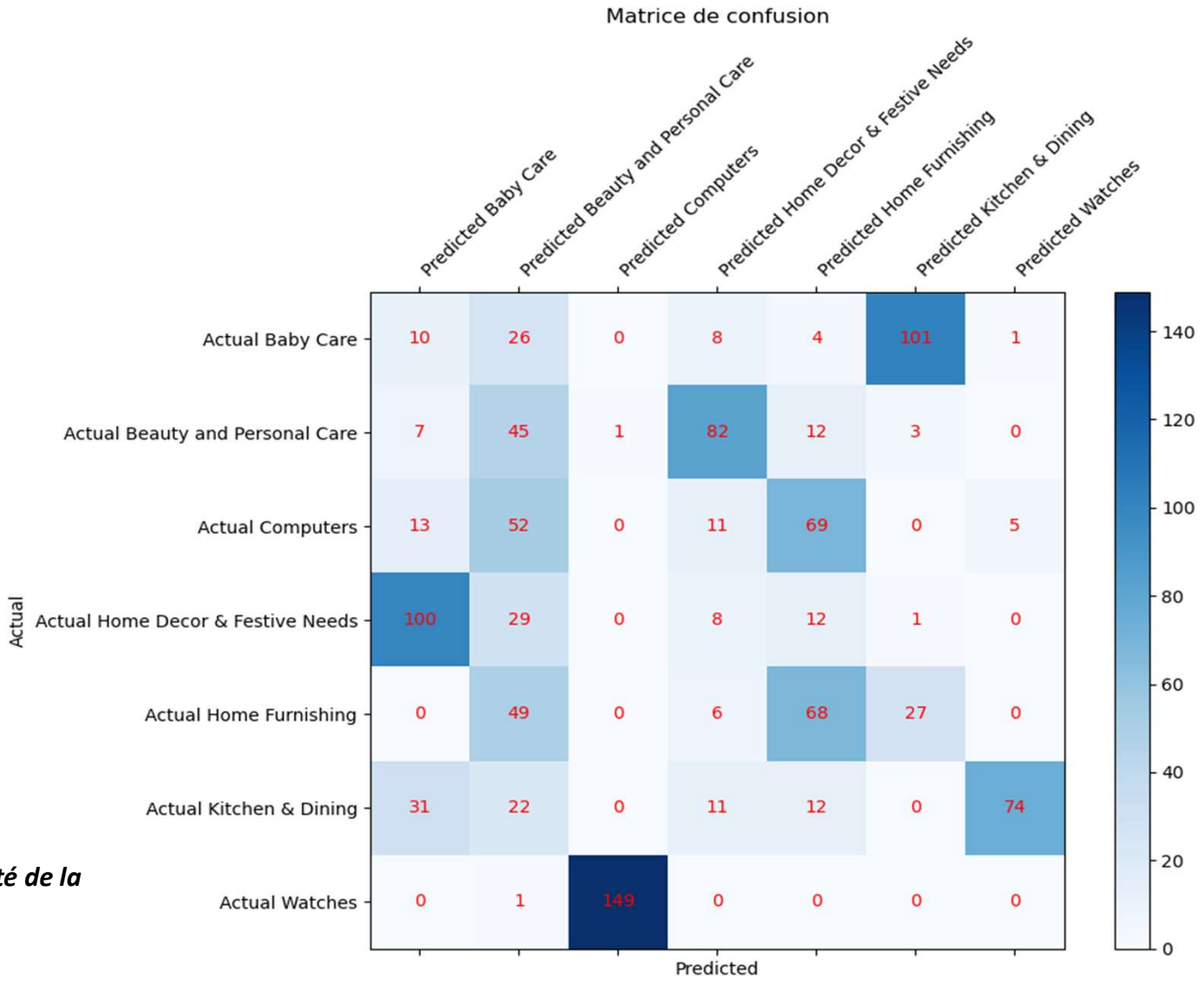
↓

Faible performance globale

Le f1-score macro avg : ***indique la moyenne non pondérée du f1-score sur toutes les classes.***

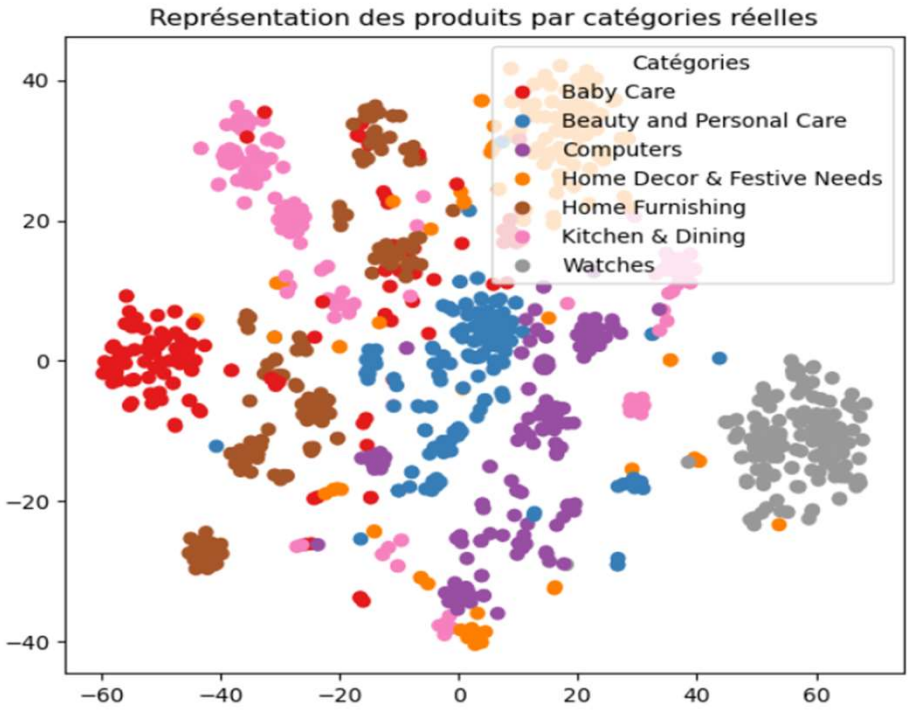
Un score proche de 0 signifie que le modèle a des performances faibles, ce qui est cohérent avec la matrice de confusion qui montre des erreurs de classification significatives.

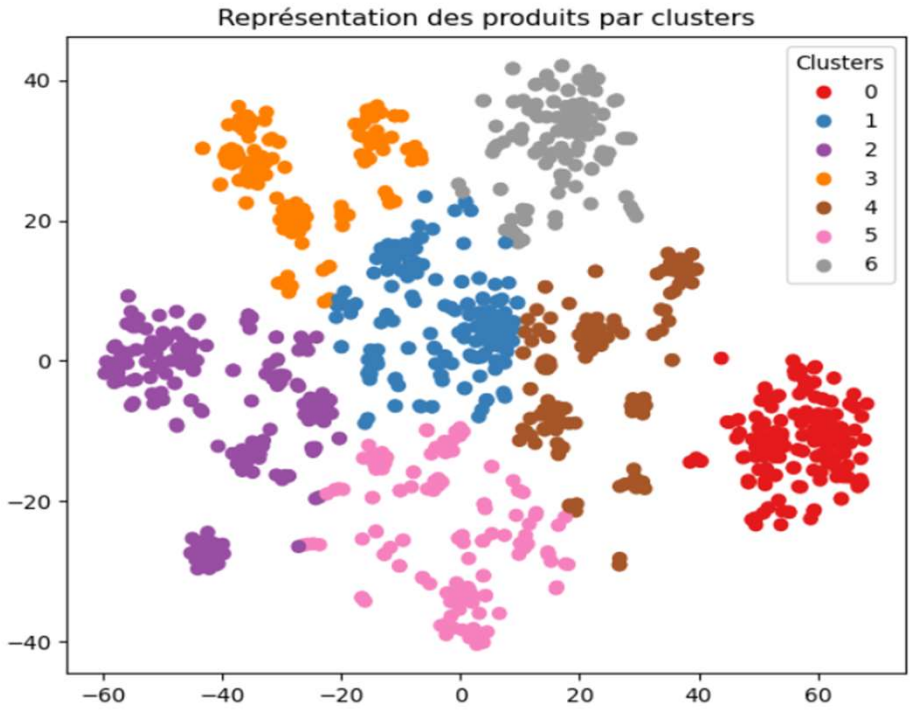
Le faible f1-score macro reflète la mauvaise qualité de la classification observée dans les graphiques.



TF-IDF (term frequency-inverse document frequency) :

les fréquences des mots sont remplacées par des scores TF-IDF



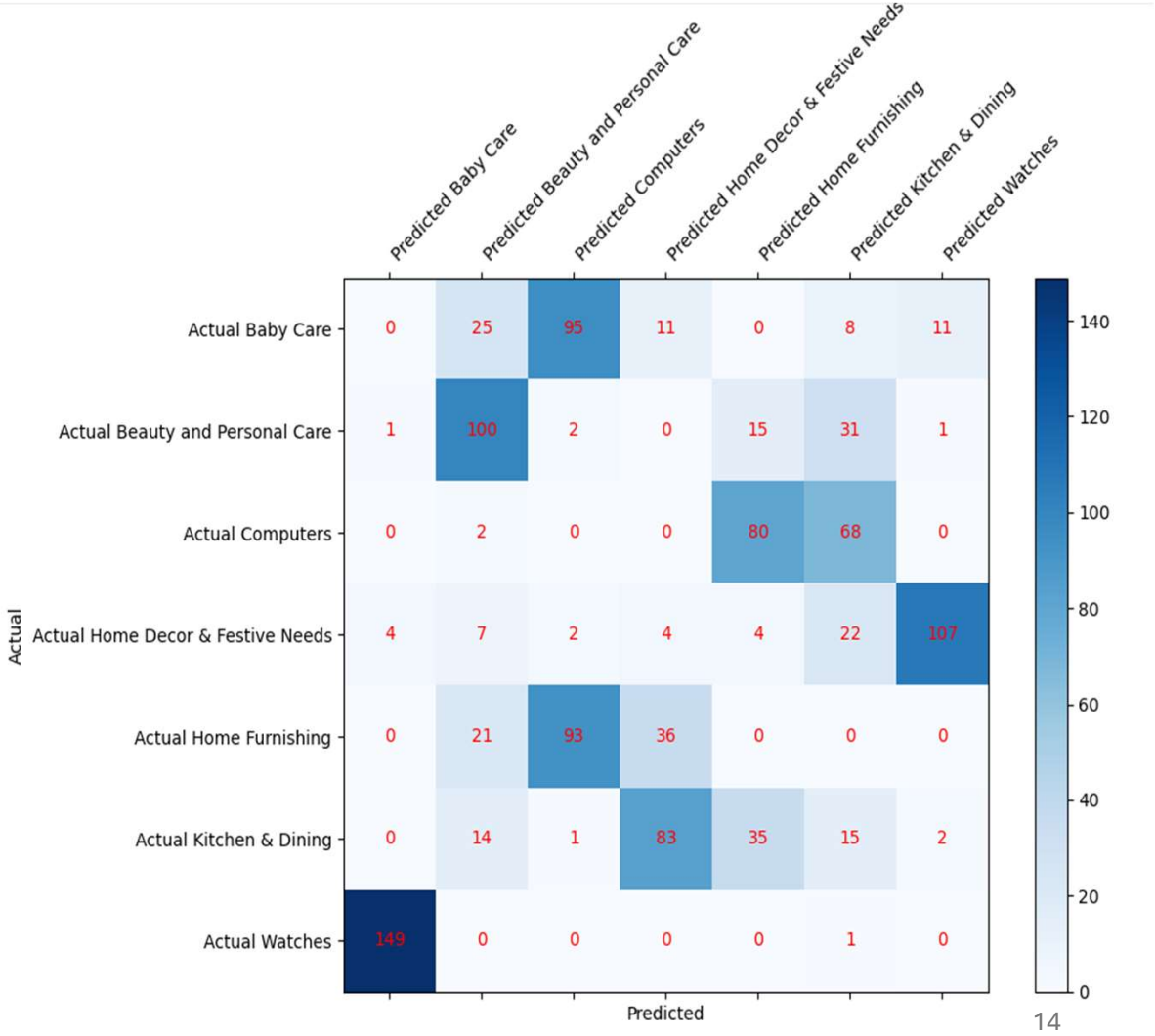


ARI : 0,45 = correspondance modérée clustering .

TF-IDF (term frequency-inverse document frequency) :

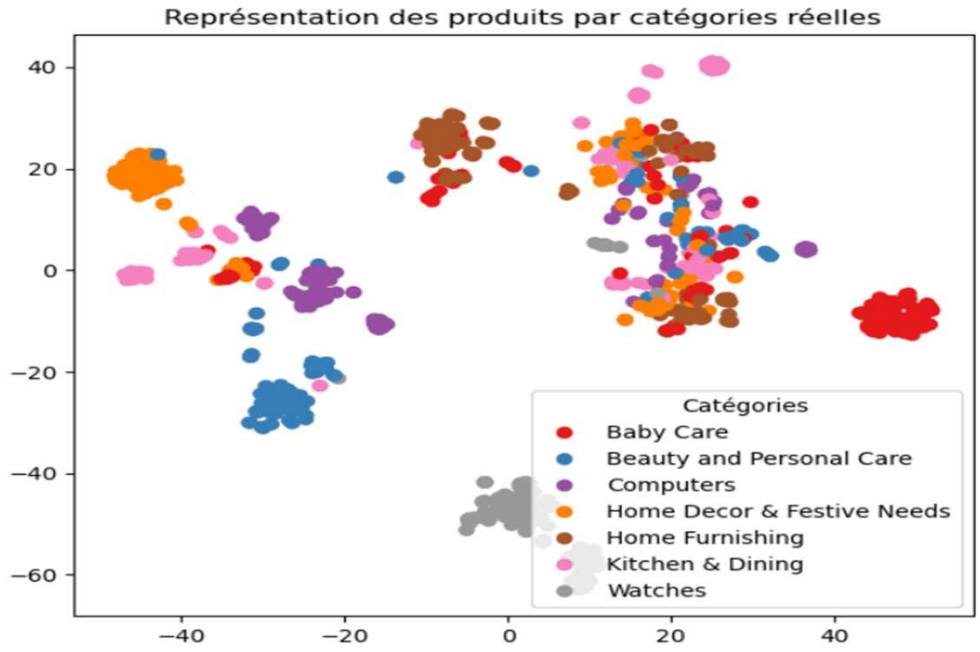
- ***La matrice de confusion montre la performance de classification du modèle pour différentes catégories de produits.***
- ***Par exemple ,Actual Baby Care : Faible performance, avec la majorité des produits mal classés en "Computers".***
- ***Actual Beauty and Personal Care : Bonne performance, avec la plupart des produits correctement classés dans leur catégorie.***

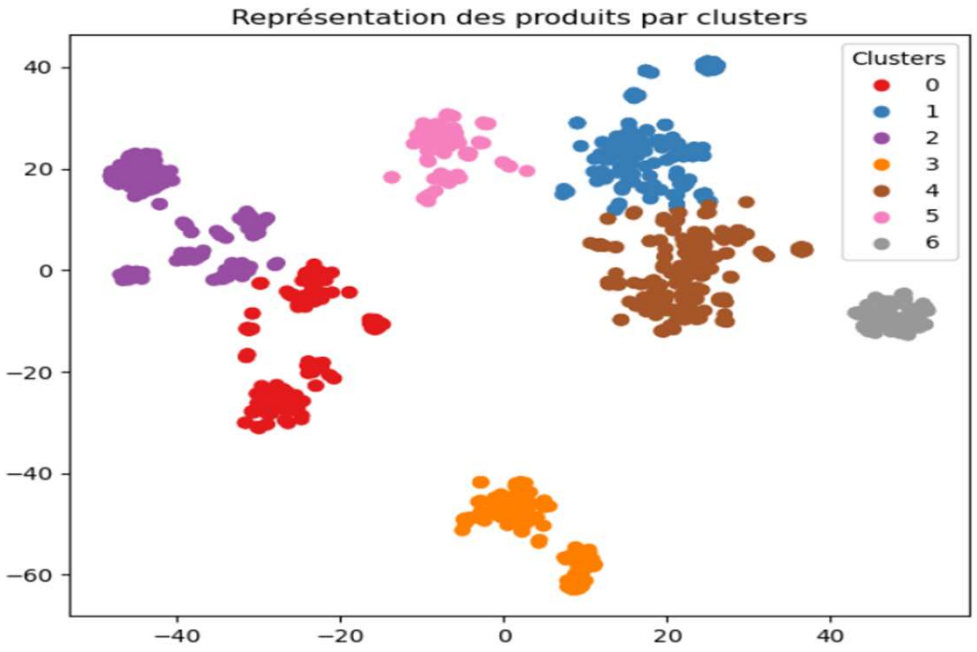
Le f1-score macro avg : 0,11
=
Faible performance globale



Word2Vec : modèle d'Embedding

Technique par plongement de mot. Chaque mot est représenté par un vecteur





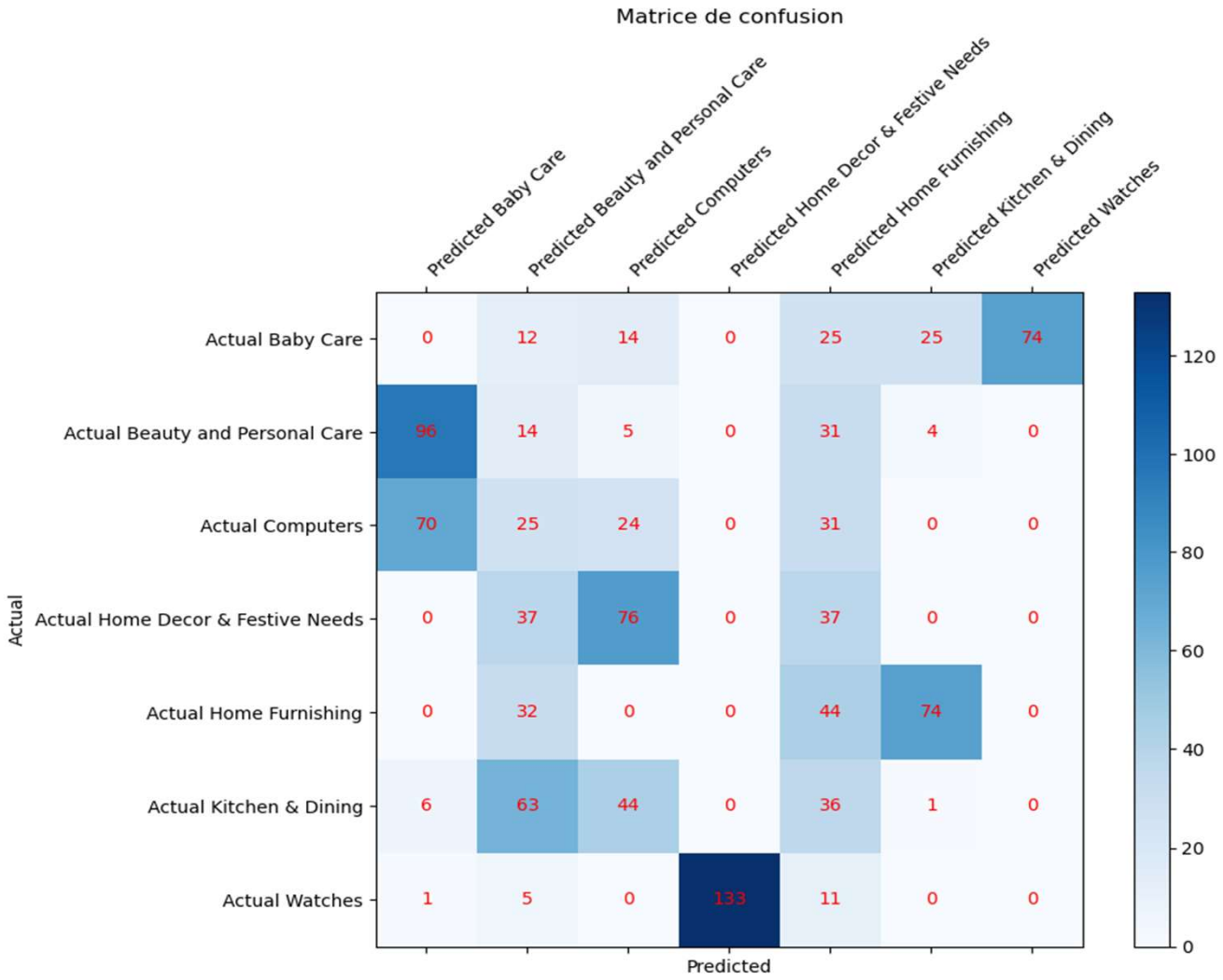
ARI : 0,30 = correspondance faible clustering .

Cohérence : ***Les clusters formés par l'algorithme de clustering ne correspondent pas toujours bien aux catégories réelles des produits.***

Clustering : ***Certaines catégories de produits sont bien séparées (comme les montres), mais d'autres sont mal regroupées, ce qui montre une performance variable du modèle de clustering.***

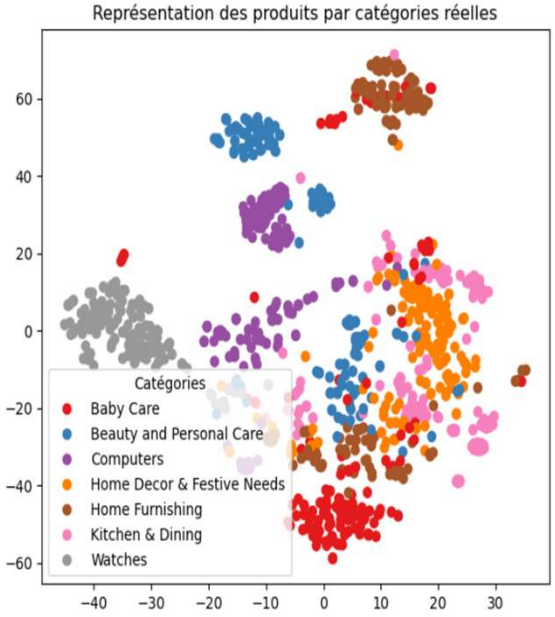
Word2Vec : modèle d'Embedding

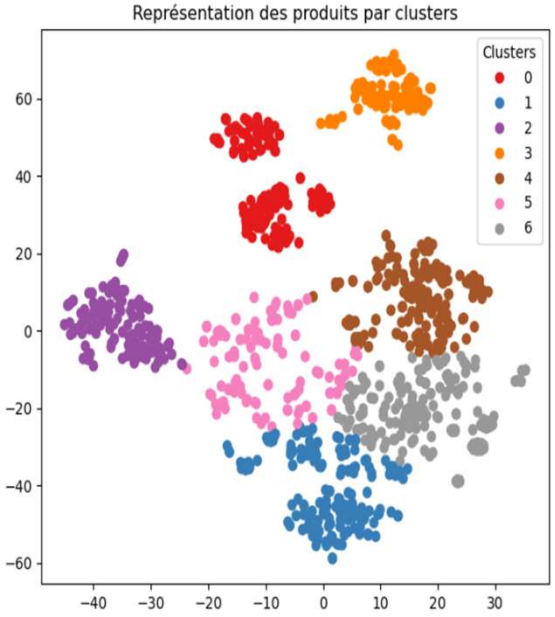
Le f1-score macro avg : 0,11
=
Faible performance générale

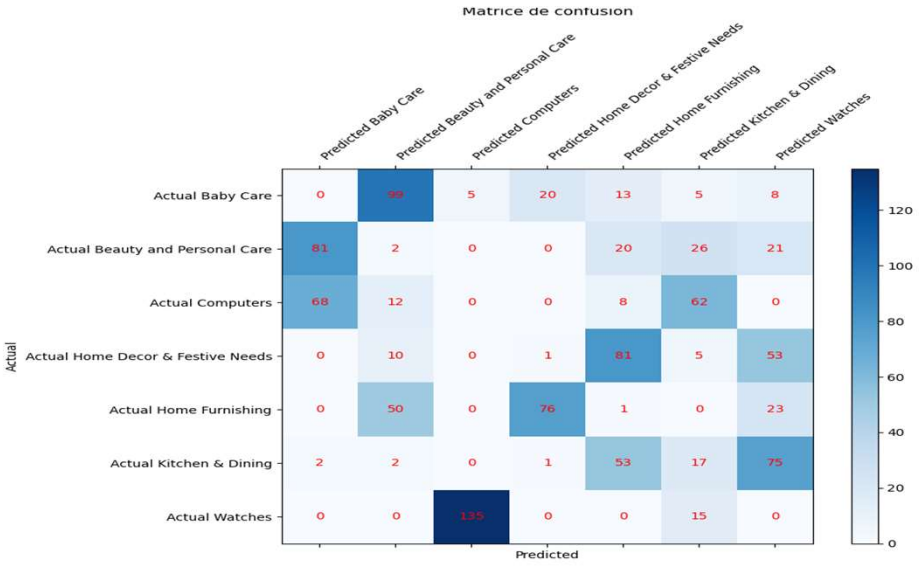


BERT: Bidirectional Encoder Représentations from Transformers

BERT est un modèle de langage pré-entraîné qui comprend le contexte de chaque mot dans une phrase en examinant simultanément ses parties gauches et droites.





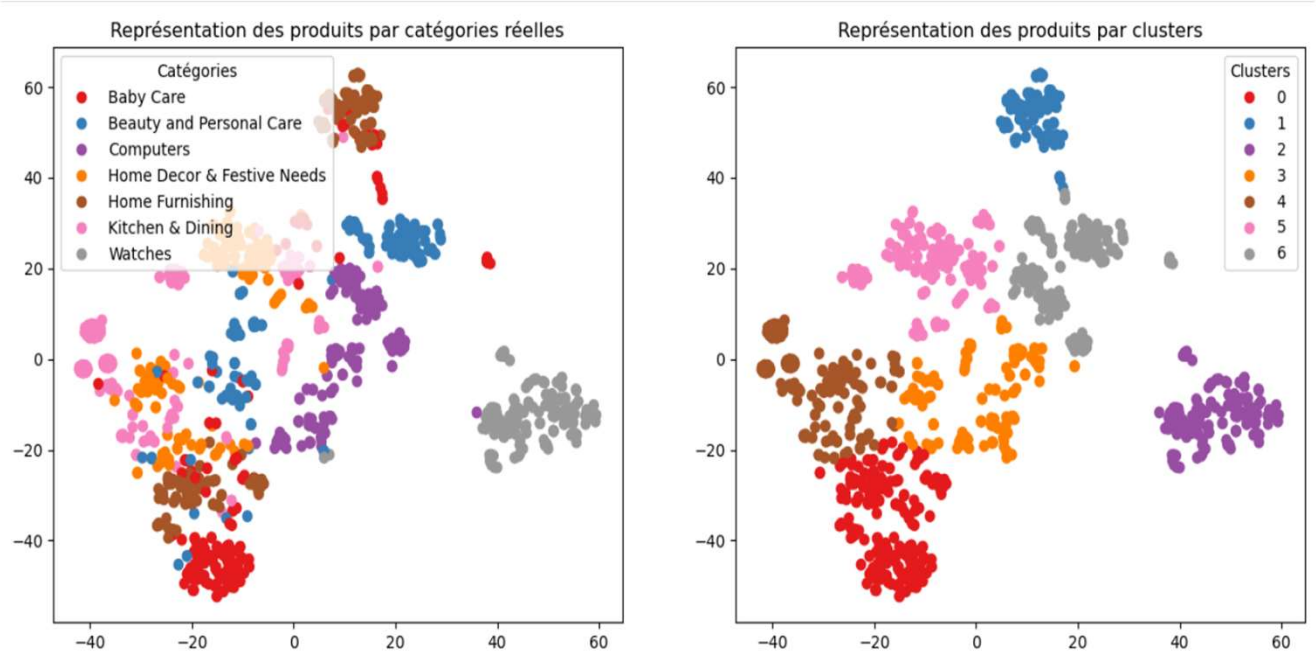


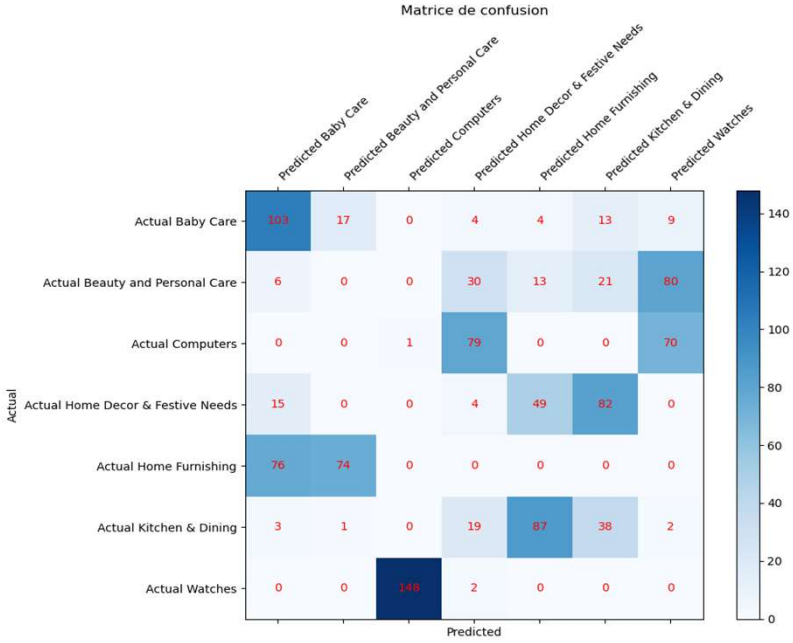
ARI : 0,36 = correspondance faible clustering .

Le f1-score macro avg : 0,021
=
Faible performance générale

USE - Universal Sentence Encoder

modèle qui convertit des phrases en vecteurs numériques pour comprendre leur signification.





ARI : 0,42 = correspondance modéré clustering .

Le f1-score macro avg : 0,12
=
Faible performance générale

Comparaison des performances : ARI

CountVectorizer	TD-IDF	Word2Vec	BERT	USE
0,37	0,45	0,30	0,36	0,42

Classification des images

Classification des images

Etapes de prétraitement des images :

Création des datasets d'images	Démarche suivie
•**Séparation des données** : 90% pour l'entraînement, 10% pour le test. •**Organisation des images** : Répertoire structuré avec des images triées par catégorie. •**Création des datasets** : •Deux datasets distincts (train et test) en renommant les images selon leur catégorie. •Ajout d'un label pour chaque image : •Nom de la catégorie. •Numéro de la catégorie.	•**Extraire et réunir l'ensemble des descripteurs SIFT** •**Création de caractéristiques** : Génération de descripteurs d'images ("bag-of-images" ou Transfer Learning). *(Utilisation de MiniBatchKMeans)* •**Réduction de dimension** : ACP puis T-SNE. •**Clustering** : Algorithme K-Means. •**Score ARI** : Calcul pour évaluer le clustering. •**Visualisation** : Présentation graphique.

Classification des images

APPROCHES DE TRAITEMENT D'IMAGES

Génération de descripteurs

SIFT

Transfer Learning basé sur les réseaux de neurones

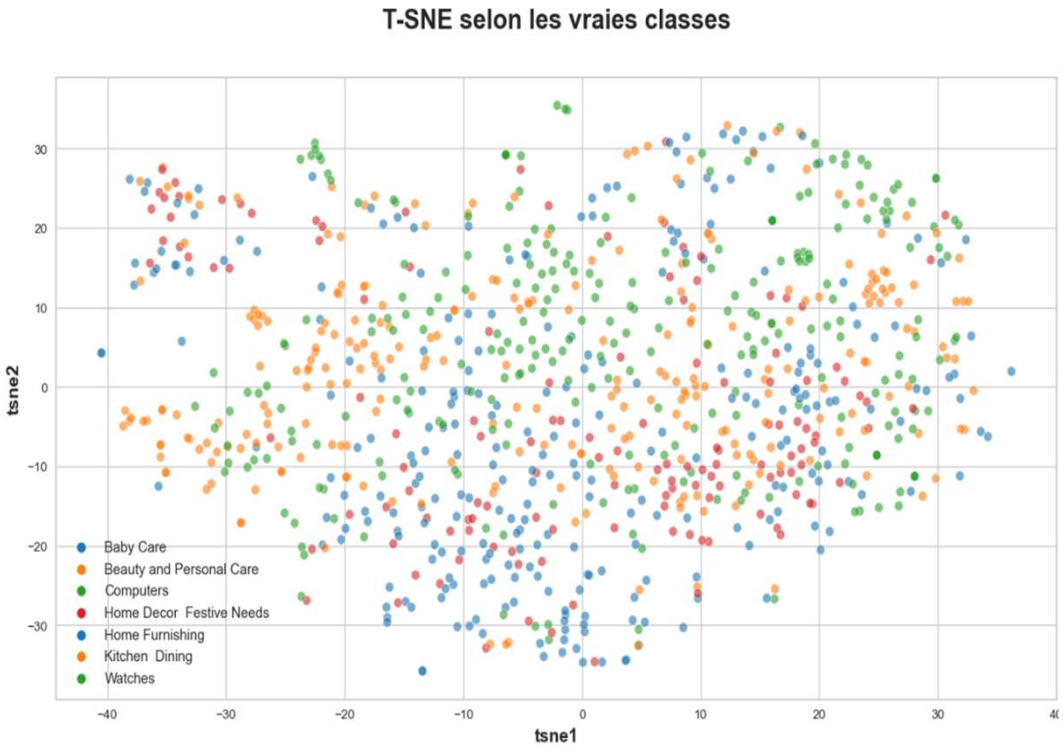
CNN Transfer Learning

SIFT (Scale Invariant Feature Transform): **permet d' identifier les éléments similaires entre différentes images**

Transfer Learning (VGG-16): **réseau de neurones convolutif pré-entrainé sur ImageNet 17**

SIFT : Scale-Invariant Feature Transform

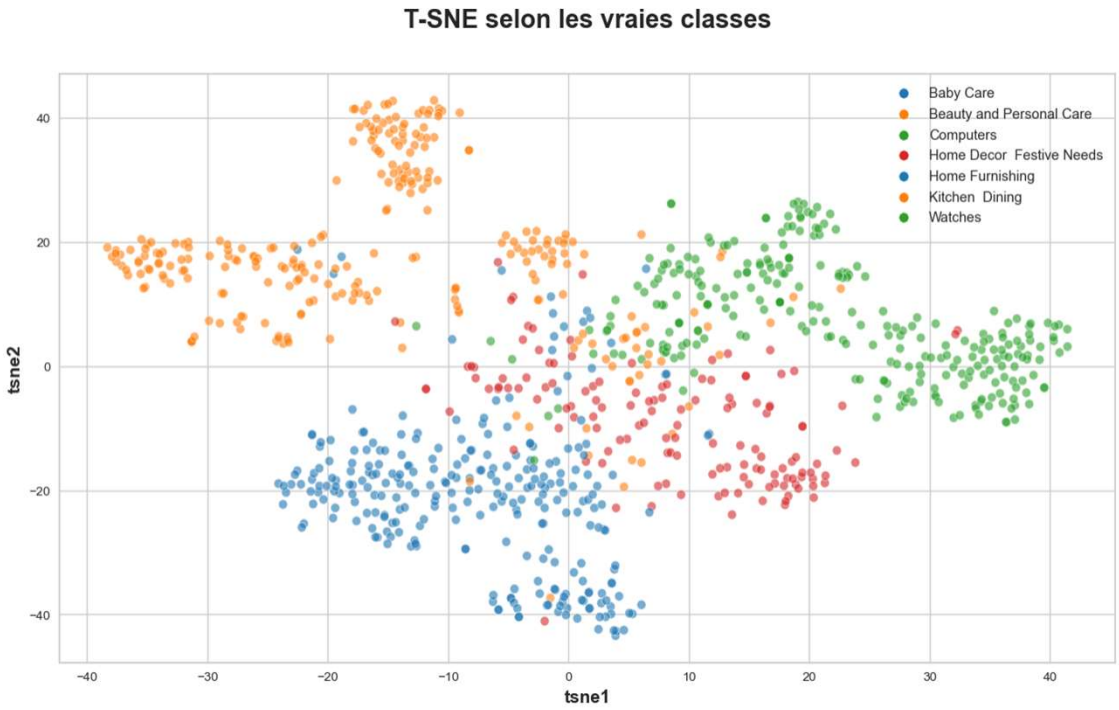
SIFT est un algorithme qui détecte et décrit des descripteurs locaux dans des images, indépendamment de l'échelle et de la rotation.



ARI : 0,58 = correspondance modéré clustering .

CNN Transfer Learning

- ***le Transfer Learning consiste à utiliser le modèle VGG16 pré-entraîné de réseaux de neurones convolutifs (CNN)sur ImageNet (ImageNet est un projet de recherche visant à développer une grande base de données d'images avec des annotations, c'est-à-dire des images et leurs descriptions).***



étapes à suivre

- Chargement des images, extraction des features
- Réduction de dimension
- Clustering

ARI : 0,52 = correspondance modéré clustering .

Classification supervisée

Classification Supervisée d'Images en Utilisant l'Apprentissage par Transfert avec CNN

Apprentissage par Transfert avec CNN

L'apprentissage par transfert est une technique très puissante pour la classification d'images, surtout lorsqu'on dispose de peu de données d'entraînement. L'idée est d'utiliser les connaissances acquises par un modèle CNN pré-entraîné sur un grand jeu de données génériques (comme ImageNet) pour résoudre un problème spécifique avec peu de données.

Etapes clés pour la classification d'images par apprentissage par transfert avec CNN

1. ***Charger un modèle CNN pré-entraîné (ex : VGG-16, ResNet)***
2. ***Séparer le jeu de données en entraînement, validation et test***
3. ***Augmenter artificiellement les données d'entraînement (rotation, zoom, etc.)***
4. ***Créer un modèle en utilisant le CNN pré-entraîné comme extracteur de caractéristiques et ajouter un classifieur***
5. ***Entraîner le modèle sur les données augmentées***
6. ***Évaluer les performances sur les jeux de validation et test***

Mesure de Performance : Accuracy

L'accuracy mesure la précision globale d'un modèle de classification supervisée.
Elle indique la proportion d'exemples correctement classés.

Un score d'accuracy plus élevé signifie de meilleures performances de classification globales du modèle.

Classification Supervisée d'Images

Approches CNN utilisées

1. Classification supervisée simplifiée

   Modèle VGG16

2. ImageDatagenerator avec augmentation des données

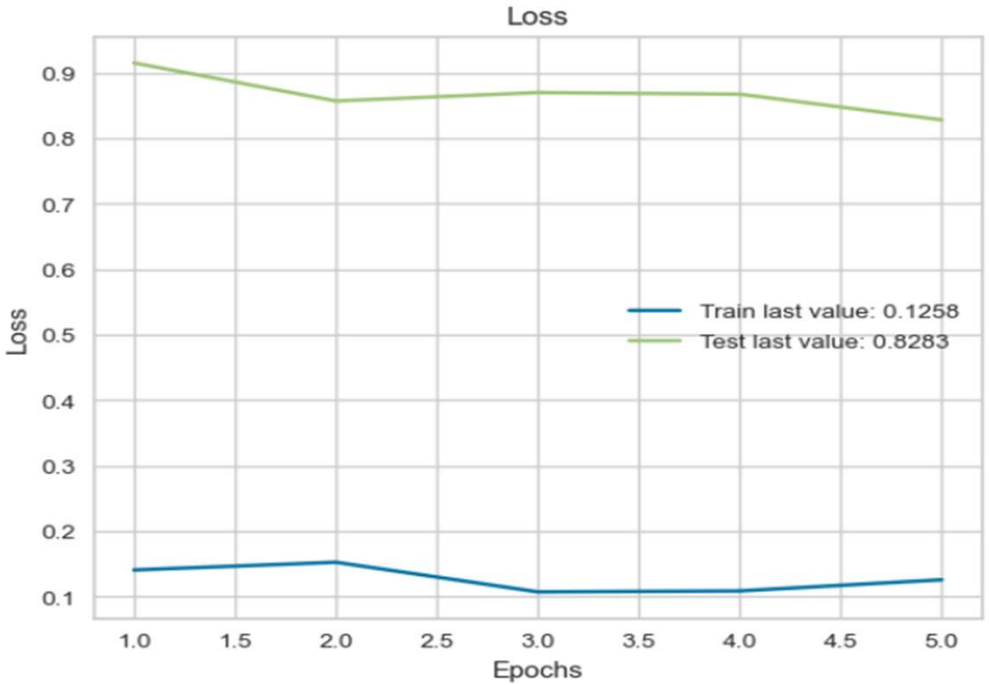
   Modèle VGG16

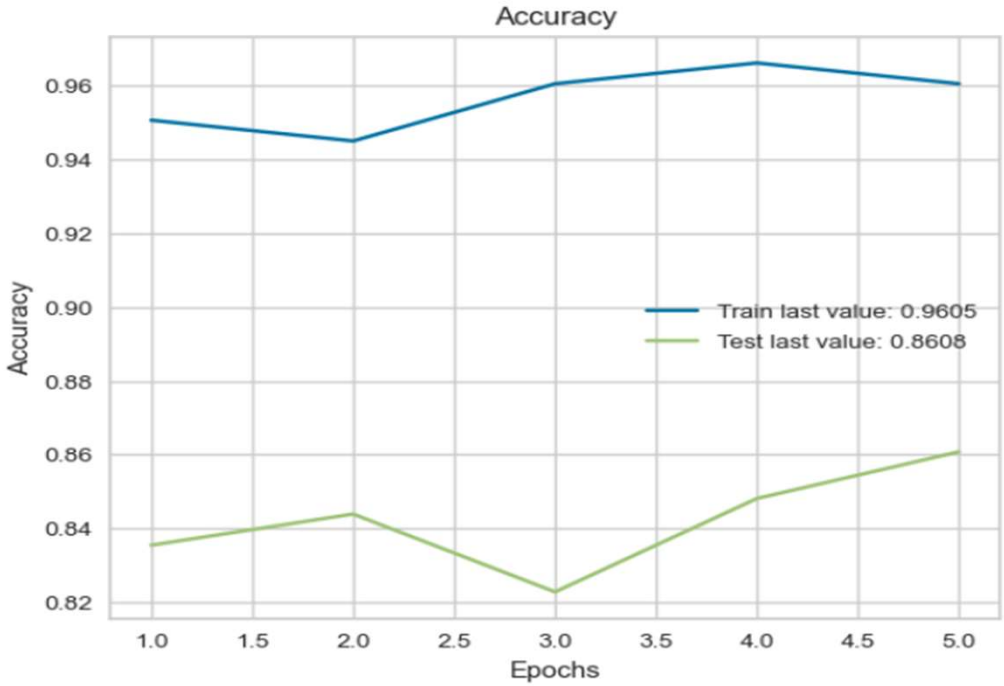
3. Nouvelle approche avec augmentation intégrée des données dans l' ensemble de données pour l' entraînement du modèle

   Modèle VGG16 ,Modèle VGG19, Modèle ResNet50

Classification supervisée simplifiée (VGG16)

La classification supervisée d'images avec VGG16 consiste à utiliser un modèle de réseau de neurones convolutif pré-entraîné sur de grandes quantités d'images pour extraire automatiquement les caractéristiques visuelles d'une nouvelle image, puis à entraîner un classifieur simple sur ces caractéristiques pour prédire la classe de l'image, permettant d'obtenir de bonnes performances avec peu de données d'entraînement grâce à l'apprentissage par transfert.

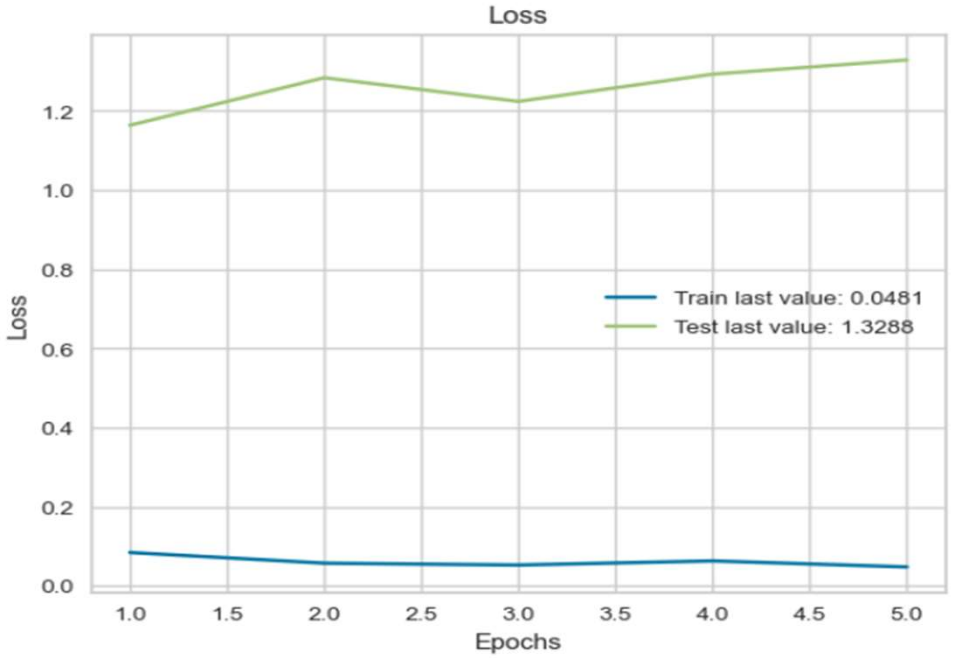


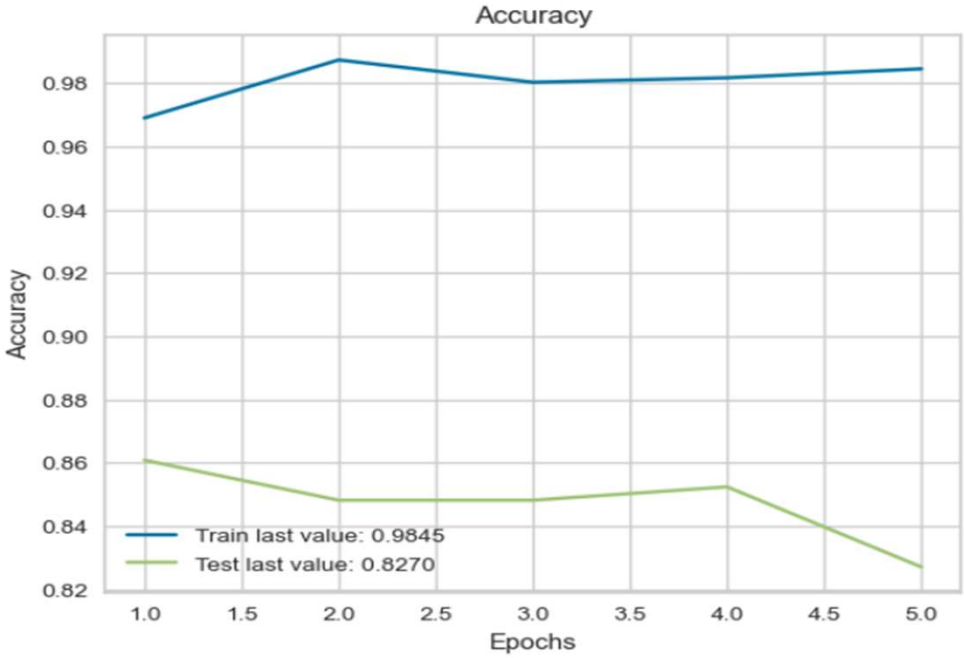


- ***Le modèle montre un bon ajustement sur les données d'entraînement, mais un écart de performance sur les données de test, indiquant un potentiel surapprentissage.***

ImageDatagenerator avec augmentation des données (VGG16)

VGG16 est un modèle de classification d'images qui utilise des couches de convolution et de pooling pour extraire des caractéristiques visuelles, puis utilise des couches de neurones pour prédire la classe de l'image.





Le modèle montre une bonne performance sur les données d'entraînement avec une loss faible et une précision élevée. Cependant, la performance sur les données de test est significativement inférieure, indiquant un surapprentissage.

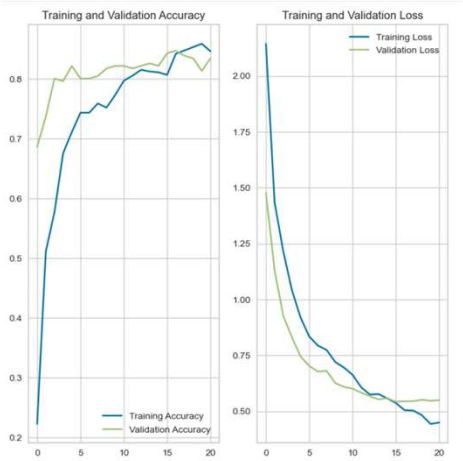
Nouvelle approche avec augmentation intégrée des données (VGG16,VGG19,RESNET50)

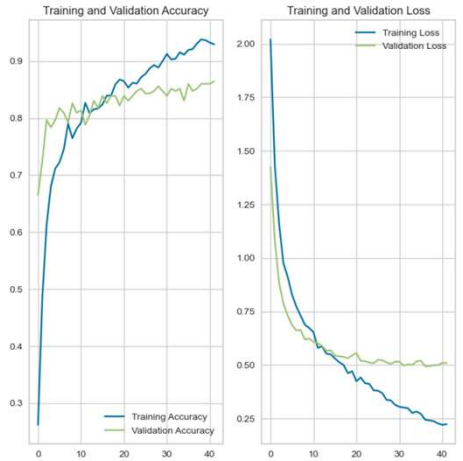
- **VGG16 :** L'approche utilisant VGG16 avec augmentation intégrée des données montre des performances améliorées par rapport aux approches sans augmentation.

- **VGG19 :** La méthode utilisant VGG19 se distingue comme **la plus performante**. Elle offre une meilleure généralisation et stabilité grâce à ses capacités avancées de traitement des caractéristiques.

- **ResNet50 :** L'approche avec ResNet50 est également performante, offre une bonne capacité de généralisation avec des temps de traitement efficaces, ce qui est avantageux pour des applications nécessitant une exécution rapide.

Conclusion : Parmi les trois approches, **VGG19** est la plus performante en termes de capacité à généraliser et à classifier correctement les produits, ce qui en fait le choix optimal pour une application en environnement e-commerce.





Conclusion

Faisabilité validée pour Texte et Images

Résultats satisfaisants de classification supervisée d’images

- **MERCI ! Des questions ?**