PLACE DE MARCHÉ

Classification automatique de biens de consommation

Marion Dedieu

01/2024

SOMMAIRE

01.

PROBLÉMATIQUE & EXPLORATION DES DONNÉES

CONTEXTE

Lancement d’une marketplace

PROBLÉMATIQUE

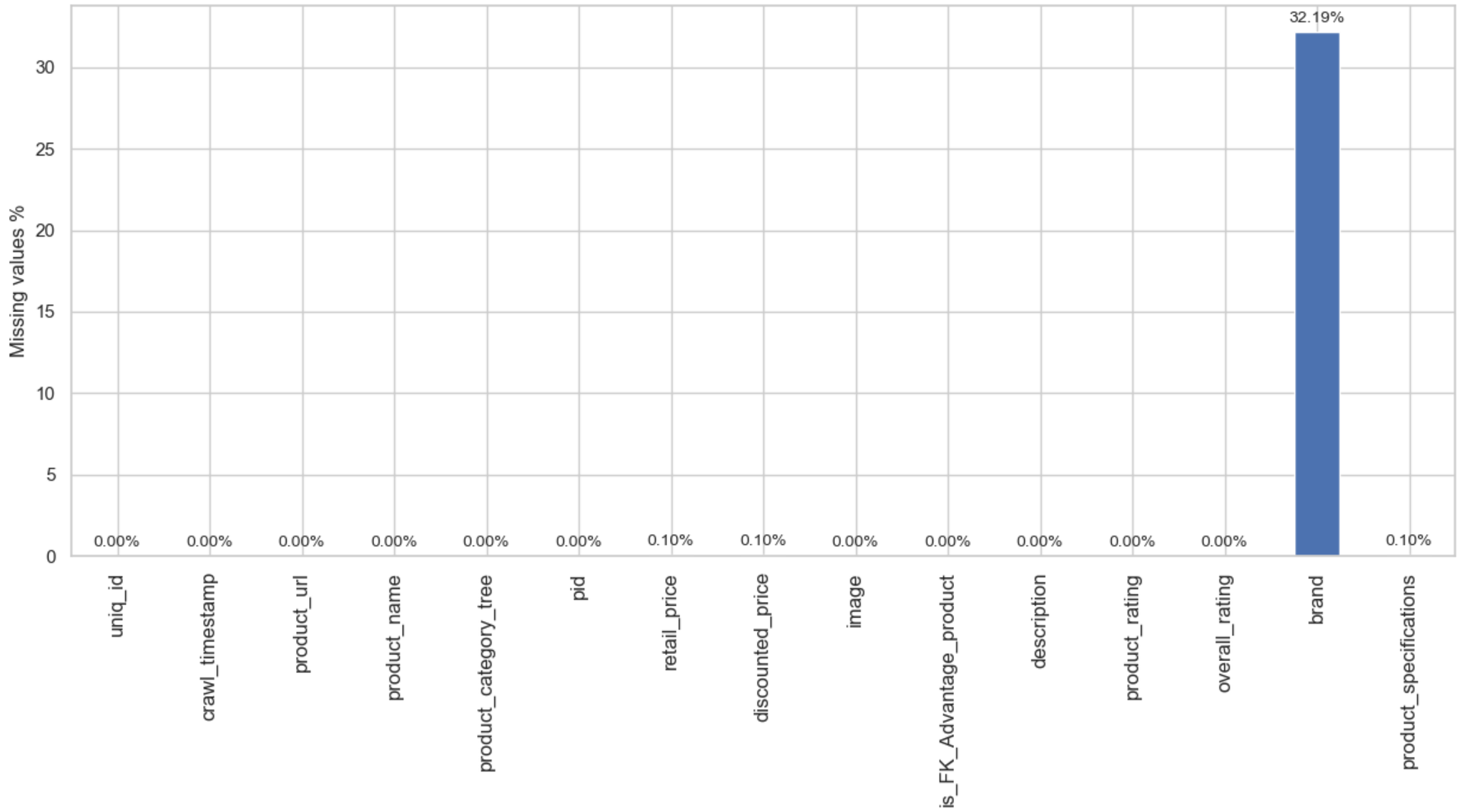
- Automatiser la tâche de catégorisation des articles, à partir :
  - Du texte de la description
  - De l’image des produits
- Tester une API pour la collecte de données

JEU DE DONNÉES

- Fichier de données : flipkart_com-ecommerce_sample_1050.csv :
  - 15 colonnes :
    - nom des produits,
    - catégorie,
    - description,
    - nom de l’image associée,
    - prix, etc.
  - 1050 lignes

- Dossier "Images" de 1050 éléments

VALEURS MANQUANTES



TRAITEMENT DES CATÉGORIES

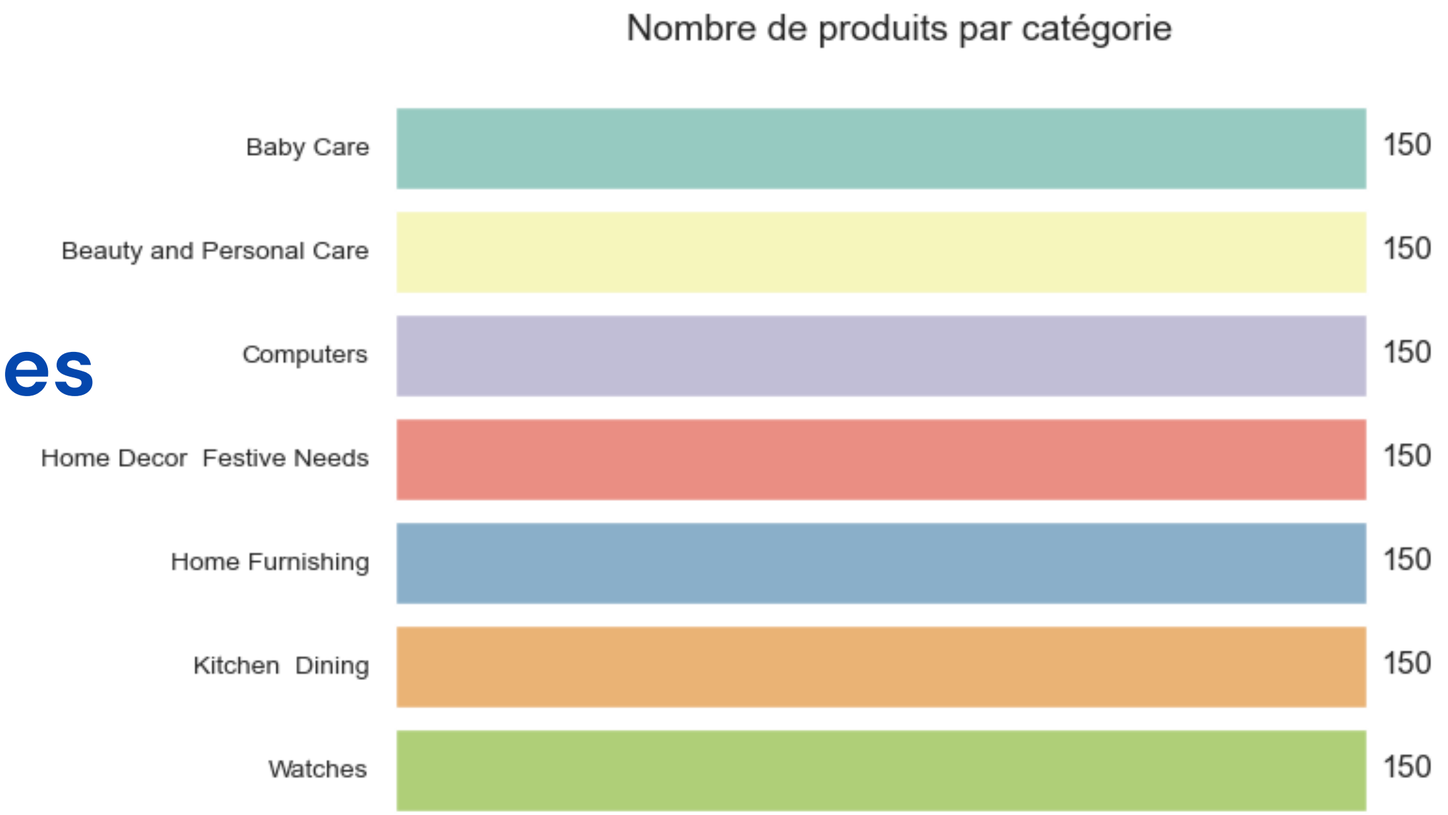
SÉPARATION DES CATÉGORIES PRINCIPALES ET DES SOUS-CATÉGORIES

Nombre de catégories

main_category : 7
sub_category_1 : 63
sub_category_2 : 247
sub_category_3 : 351
sub_category_4 : 298
sub_category_5 : 118

Catégories principales

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



02.

ETUDE DE FAISABILITÉ

NATURAL LANGUAGE PROCESSING (NLP)

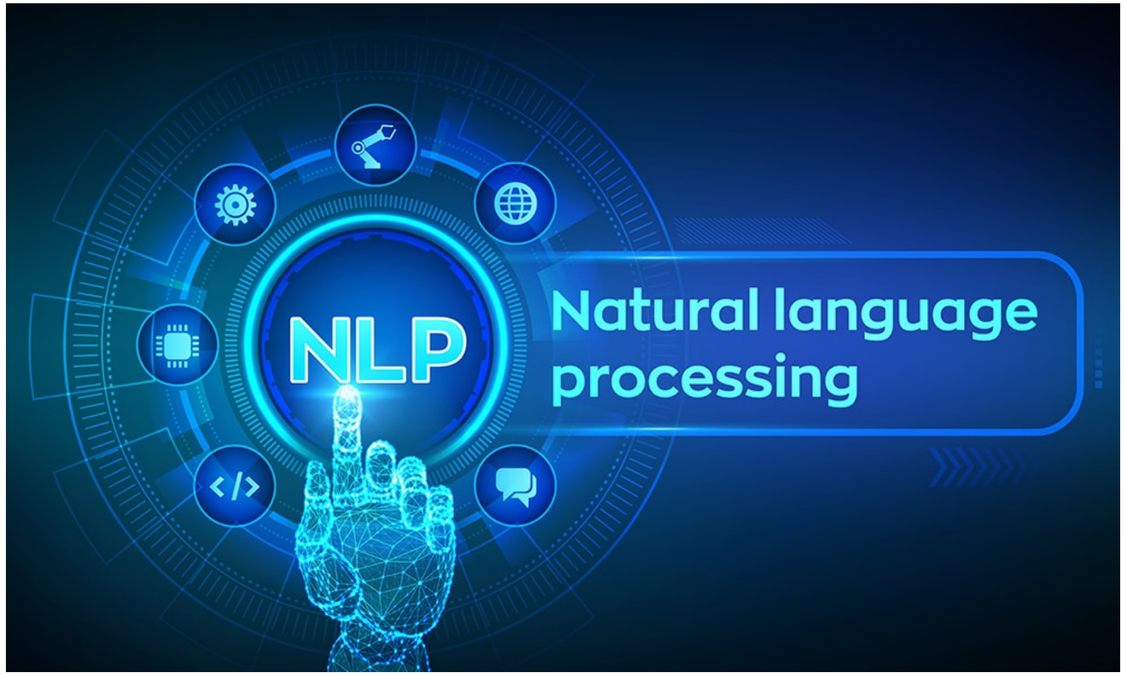
CRÉATION DES FONCTIONS

- **3 fonctions :**
  - Mapping POS
  - Lemmatisation : Transformation des mots à leur forme de base
  - Stemming : Réduction des mots à leur racine

- Choix de la lemmatisation

Exemple de description suite à la lemmatisation :

feature elegance polyester multicolor eyelet door curtain curtain elegance polyester multicolor eyelet door ....



APPROCHES NLP UTILISÉES

Ancienne génération : Fréquentiste

- CountVectorizer
- TF-IDF

Nouvelle génération : Embedding

- Word2Vec
- BERT
- USE

DÉMARCHE

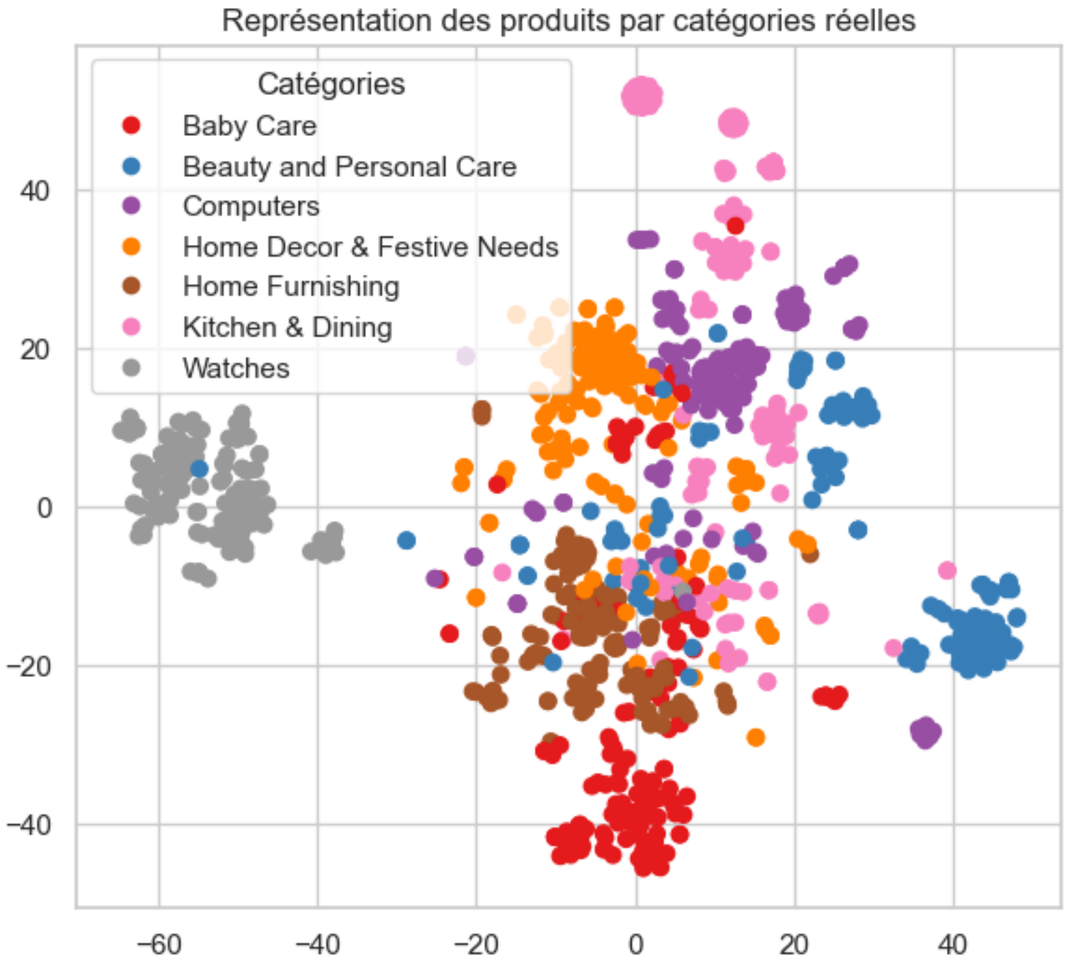
- Création de features à partir des noms des produits et des descriptions lemmatisées (bag-of-words, vecteurs de mots, etc.)
- Réduction de dimension avec une ACP (Test avec LDA), puis T-SNE
- Clustering avec l'algorithme K-Means
- Calcul du score ARI
- Visualisation graphique
- Analyse par classe avec une matrice de confusion

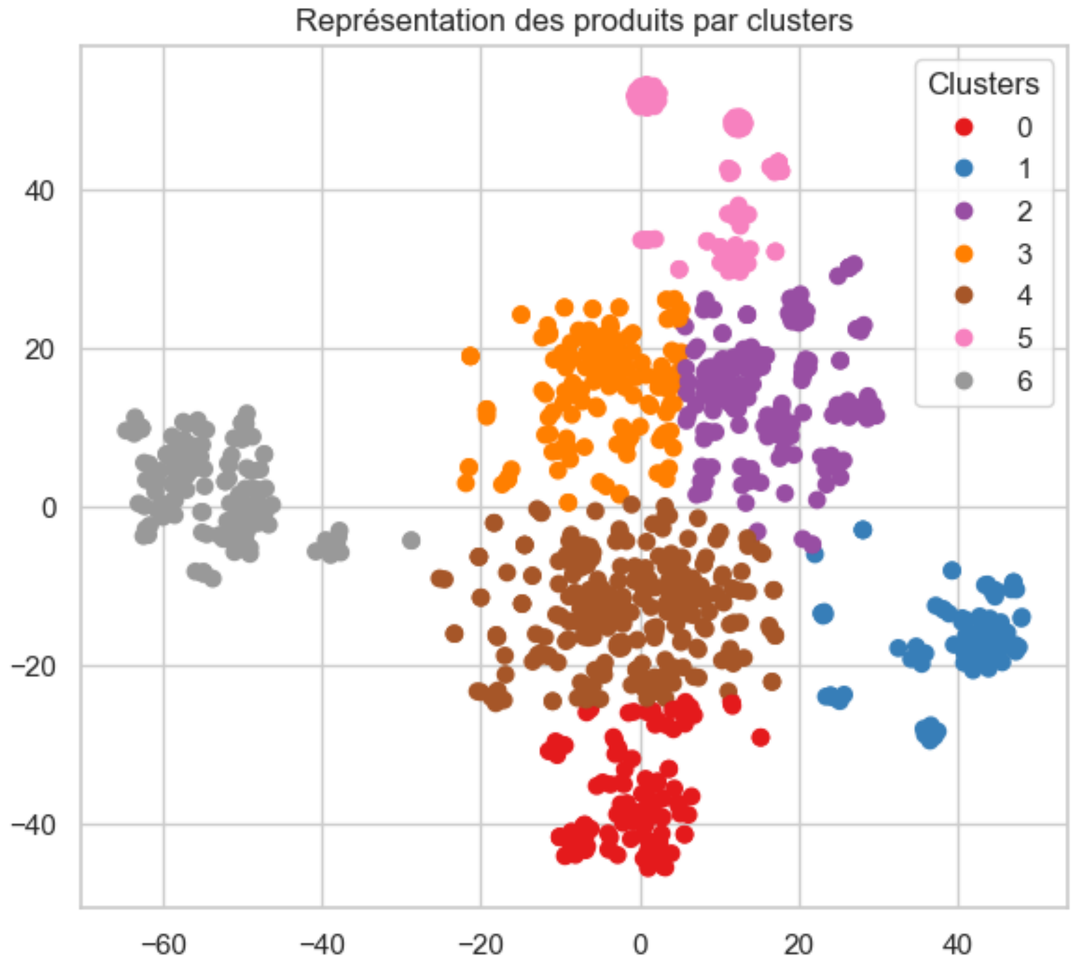
MESURE DE PERFORMANCE

- **Score ARI** (Adjusted Rand Index) :
  - Mesure la concordance entre les regroupements prédits par un algorithme de clustering et les regroupements de référence
  - Un score plus élevé indique une meilleure adéquation des regroupements

COUNTVECTORIZER

ARI = 0.4110

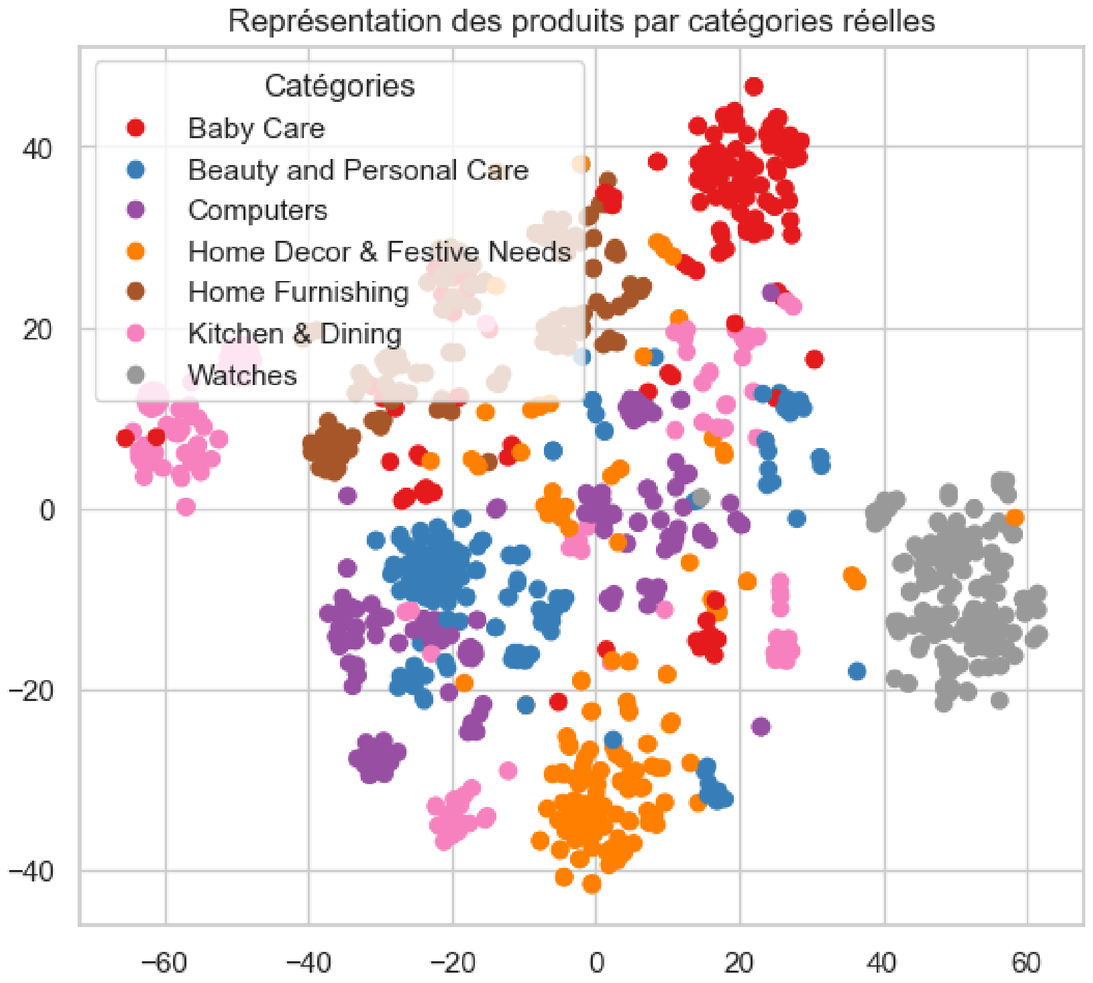


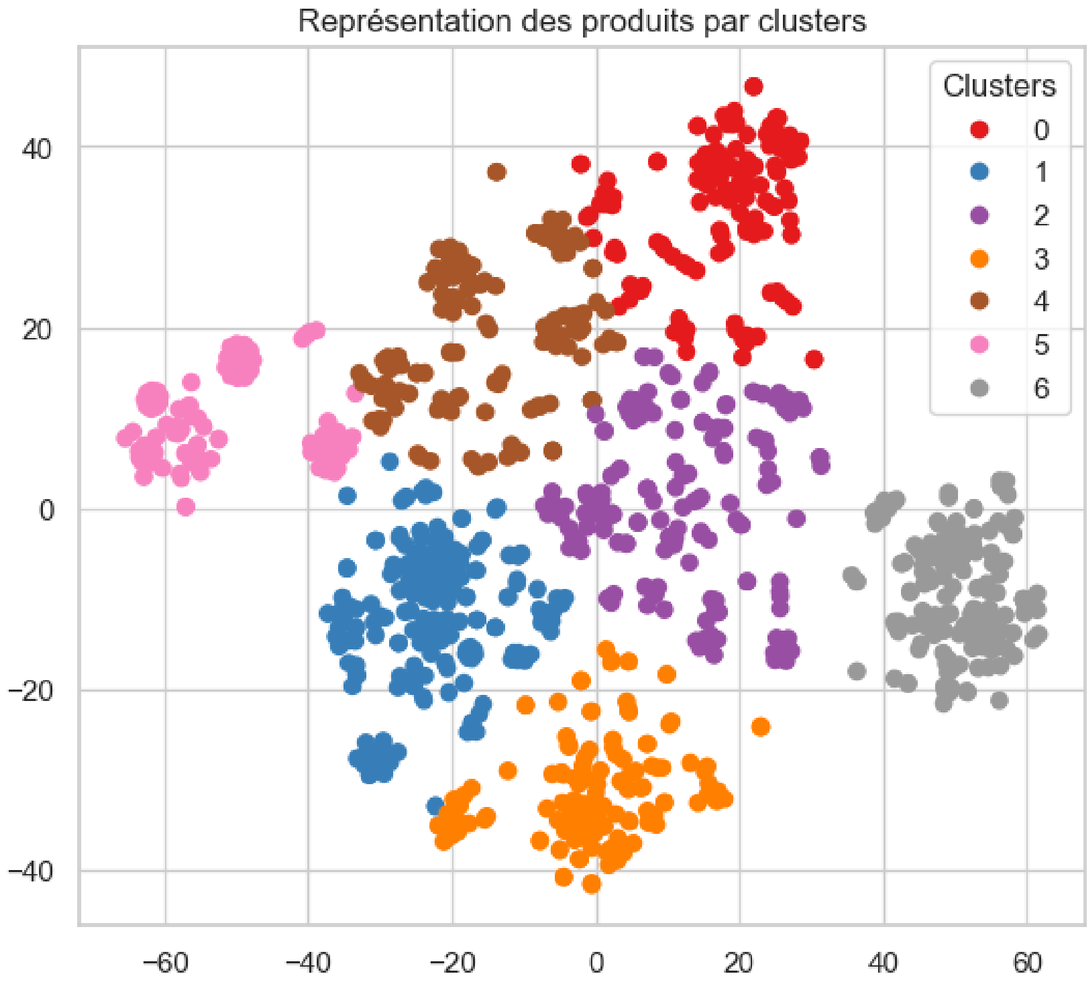


COUNTVECTORIZER

TF-IDF

ARI = 0.4549



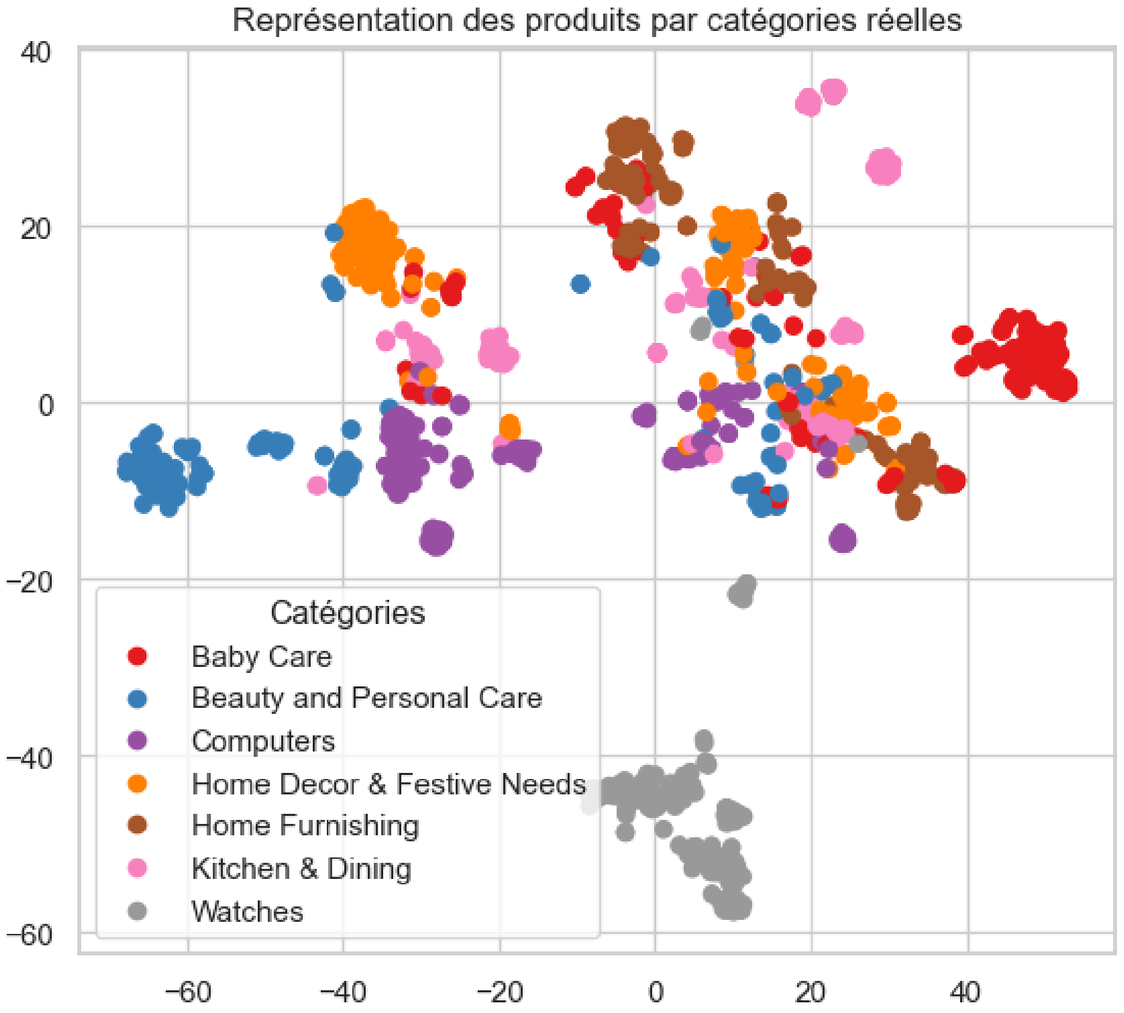


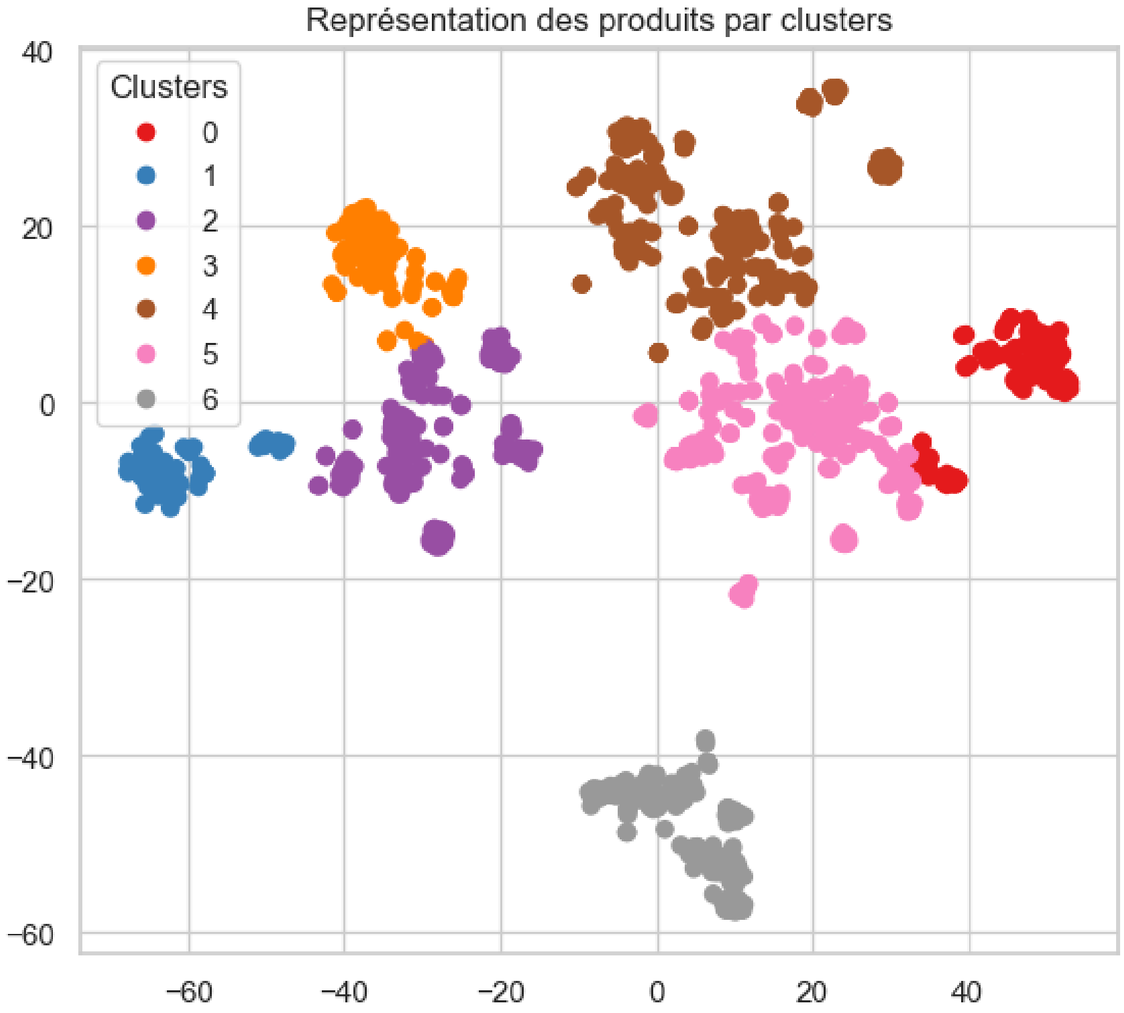
TF-IDF


Matrice de confusion
Actual Baby Care
Actual Beauty and Personal Care
Actual Computers
Actual Home Decor & Festive Needs
Actual Home Furnishing
Actual Kitchen & Dining
Actual Watches
101 8 15 2 22 2 0
1 106 24 11 7 0 1
1 78 69 2 0 0 0
9 1 26 97 13 0 4
18 0 0 0 105 27 0
14 5 35 20 1 75 0
0 0 1 0 0 0 149
Predicted Baby Care
Predicted Beauty and Personal Care
Predicted Computers
Predicted Home Decor & Festive Needs
Predicted Home Furnishing
Predicted Kitchen & Dining
Predicted Watches
140
120
100
80
60
40
20
0

WORD2VEC

ARI = 0.3190



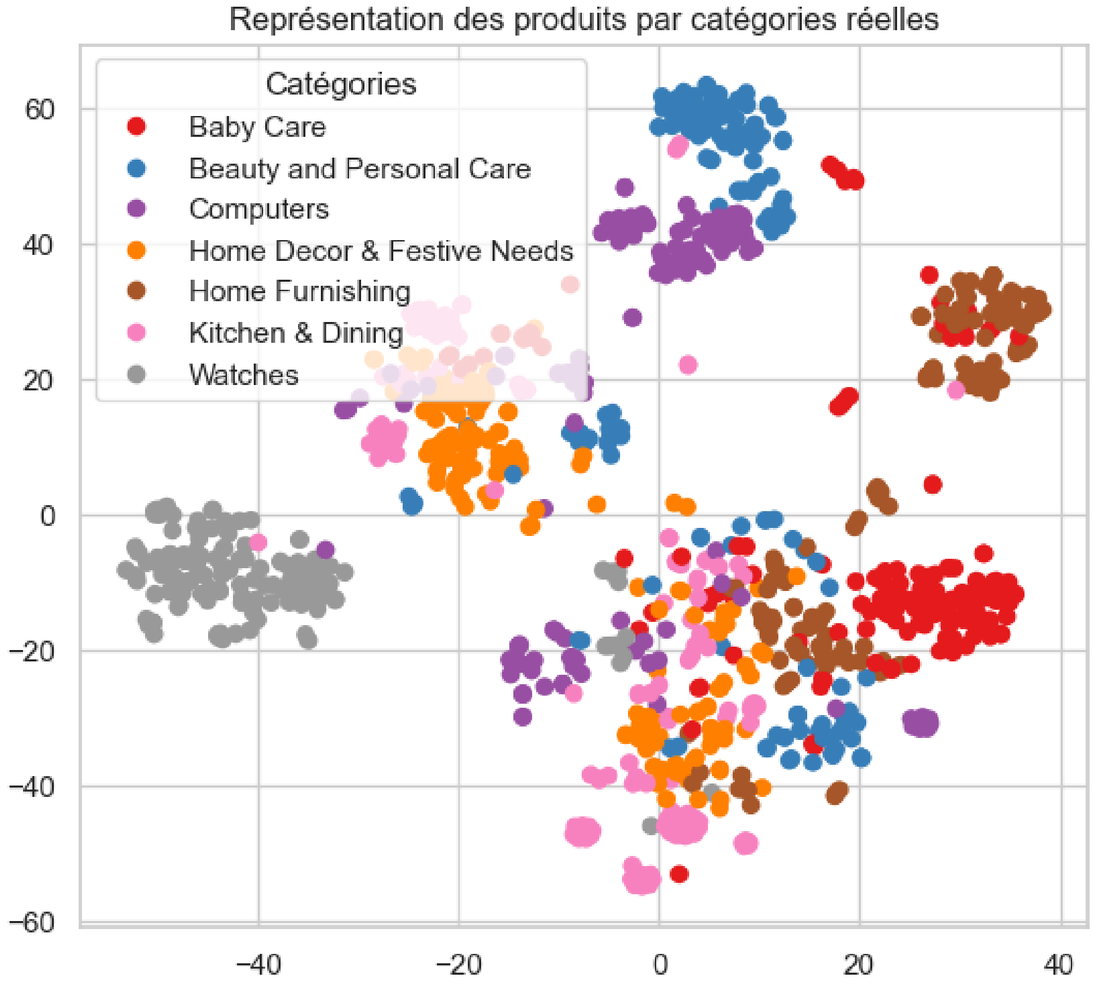


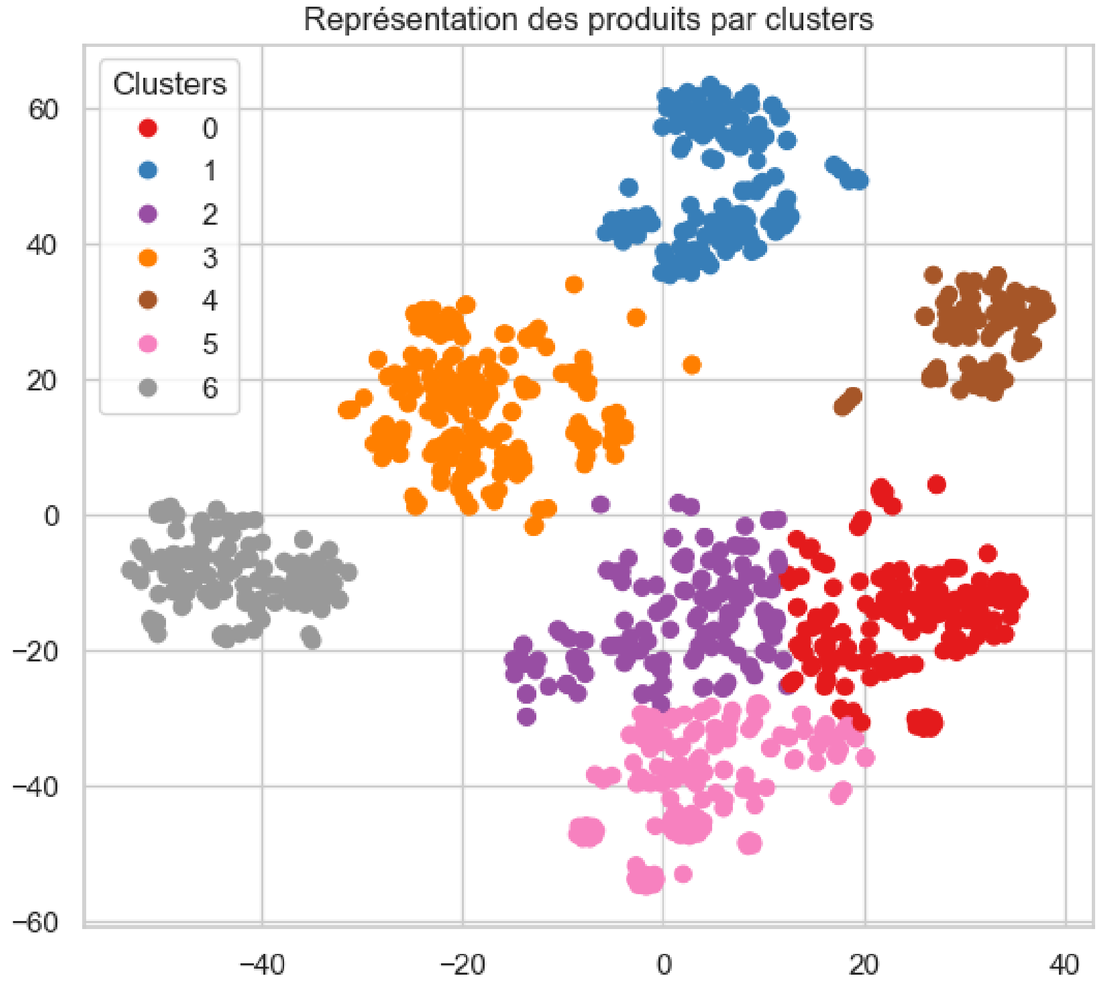
WORD2VEC


Matrice de confusion
Actual Baby Care
Actual Beauty and Personal Care
Actual Computers
Actual Home Decor & Festive Needs
Actual Home Furnishing
Actual Kitchen & Dining
Actual Watches
83 0 4 10 33 20 0
0 76 20 5 11 38 0
0 0 94 0 1 55 0
0 0 8 68 33 41 0
19 0 0 0 104 27 0
0 0 41 9 61 39 0
0 0 0 0 3 11 136
Predicted Baby Care
Predicted Beauty and Personal Care
Predicted Computers
Predicted Home Decor & Festive Needs
Predicted Home Furnishing
Predicted Kitchen & Dining
Predicted Watches
120
100
80
60
40
20
0

BERT

ARI = 0.3263



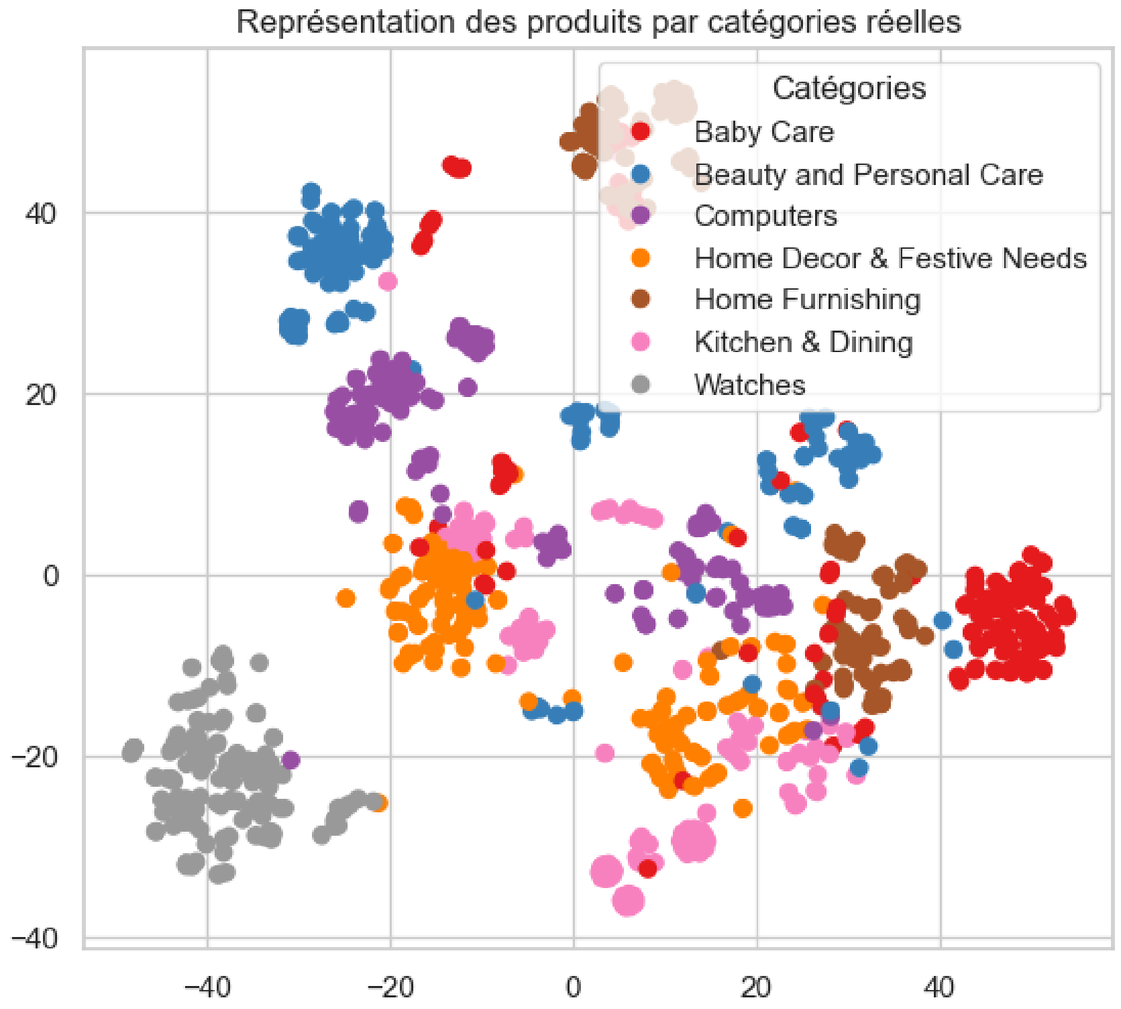


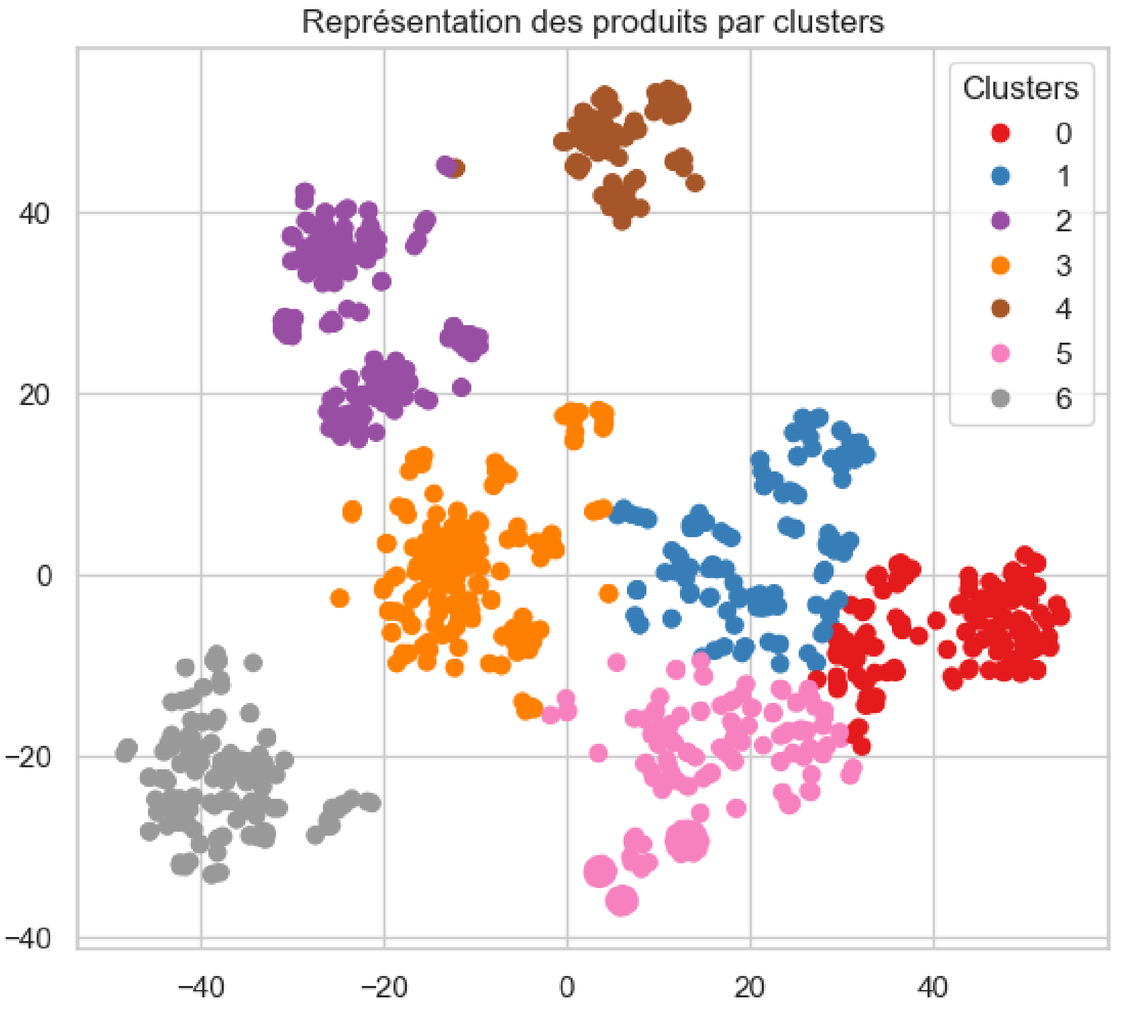
BERT


Matrice de confusion
Actual Baby Care
Actual Beauty and Personal Care
Actual Computers
Actual Home Decor & Festive Needs
Actual Home Furnishing
Actual Kitchen & Dining
Actual Watches
93 7 14 14 18 4 0
10 80 11 22 0 27 0
11 70 43 25 0 0 1
1 0 25 84 0 40 0
45 0 19 0 74 12 0
0 2 26 50 1 70 1
0 0 13 0 0 2 135
Predicted Baby Care
Predicted Beauty and Personal Care
Predicted Computers
Predicted Home Decor & Festive Needs
Predicted Home Furnishing
Predicted Kitchen & Dining
Predicted Watches
120
100
80
60
40
20
0

USE

ARI = 0.3263





USE


Matrice de confusion
Actual Baby Care
Actual Beauty and Personal Care
Actual Computers
Actual Home Decor & Festive Needs
Actual Home Furnishing
Actual Kitchen & Dining
Actual Watches
91 14 10 14 15 6 0
3 40 80 21 0 6 0
0 52 72 23 0 2 1
0 13 0 79 0 57 1
57 17 0 0 75 1 0
0 8 2 47 1 92 0
0 0 0 0 0 0 150
Predicted Baby Care
Predicted Beauty and Personal Care
Predicted Computers
Predicted Home Decor & Festive Needs
Predicted Home Furnishing
Predicted Kitchen & Dining
Predicted Watches
140
120
100
80
60
40
20
0

COMPARAISON DES PERFORMANCES (ARI)

CountVectorizer	TD-IDF	Word2Vec	BERT	USE
0.4110	**0.4549**	0.3190	0.3263	0.3970

TRAITEMENT DES IMAGES

CRÉATION DES DATASETS D'IMAGES

- Séparation en train (90%) et test (10%)
- Création d'un répertoire : images triées par catégorie
- Création de 2 datasets train et test en renommant les images par catégorie et en ajoutant un label :
  - Nom de la catégorie
  - Numéro de la catégorie

```
test_data.sample(5)
```

	image_path	label_name	label
3	Datacat/test/Computers/Computers148.jpg	Computers	2
70	Datacat/test/Home Decor Festive Needs/Home De...	Home Decor Festive Needs	3
15	Datacat/test/Home Furnishing/Home Furnishing13...	Home Furnishing	4
59	Datacat/test/Watches/Watches135.jpg	Watches	6
37	Datacat/test/Kitchen Dining/Kitchen Dining14...	Kitchen Dining	5

APPROCHES DE TRAITEMENT D'IMAGES

Génération de descripteurs

- SIFT

Transfer Learning basé sur les réseaux de neurones

- CNN Transfer Learning

DÉMARCHE

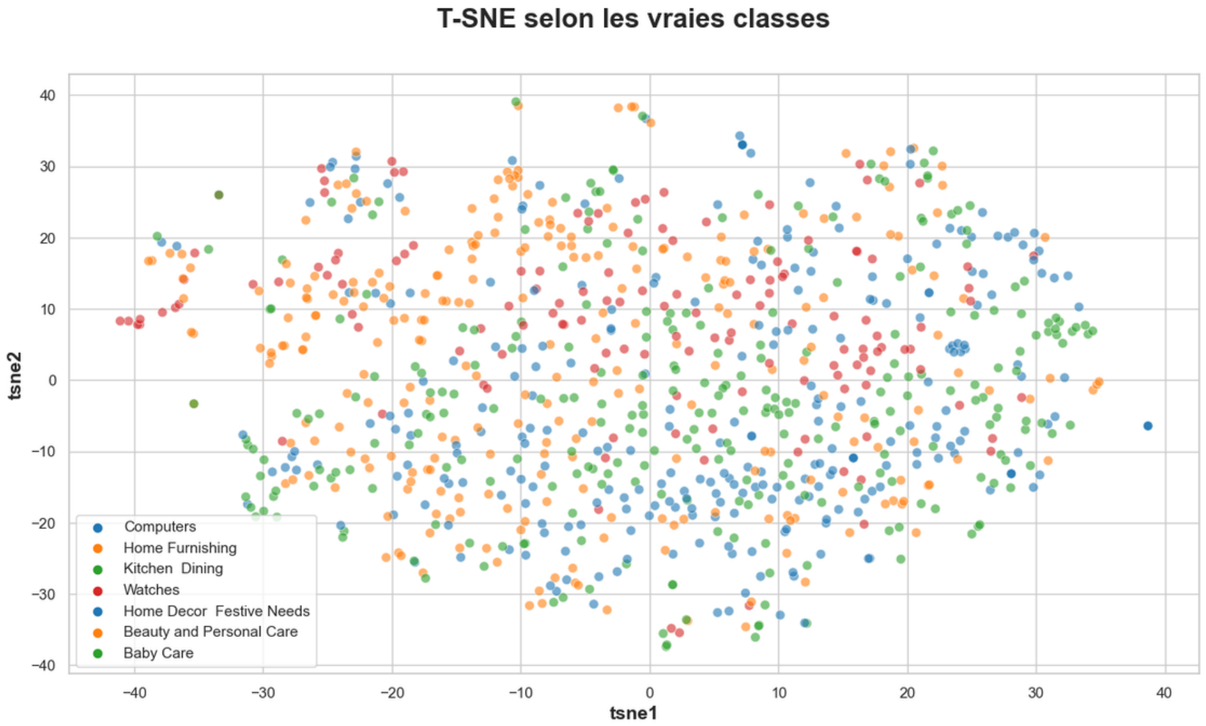
- Création de features à partir des images ("bag-of-images" pour la génération de descripteurs ou via un algorithme de Transfer Learning)
- Réduction de dimension avec une ACP, puis T-SNE
- Clustering avec l'algorithme K-Means
- Calcul du score ARI
- Visualisation graphique

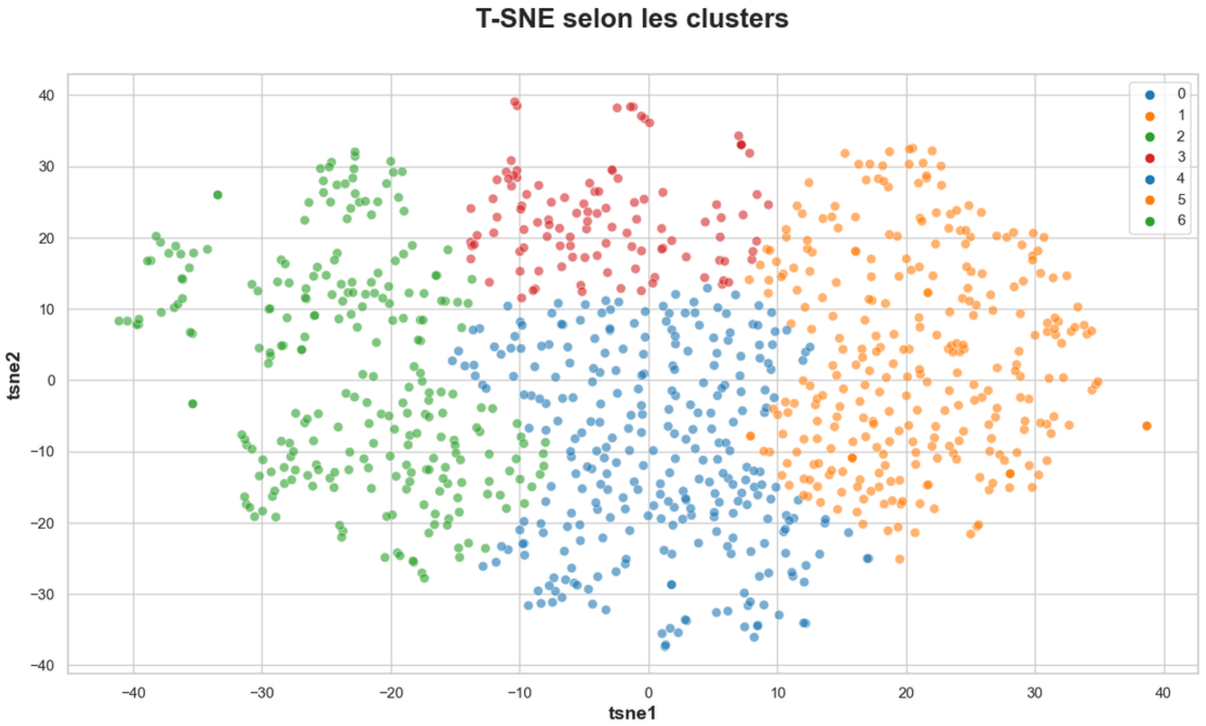
MESURE DE PERFORMANCE

- **Score ARI** (Adjusted Rand Index) :
  - Mesure la concordance entre les regroupements prédits par un algorithme de clustering et les regroupements de référence
  - Un score plus élevé indique une meilleure adéquation des regroupements

SIFT

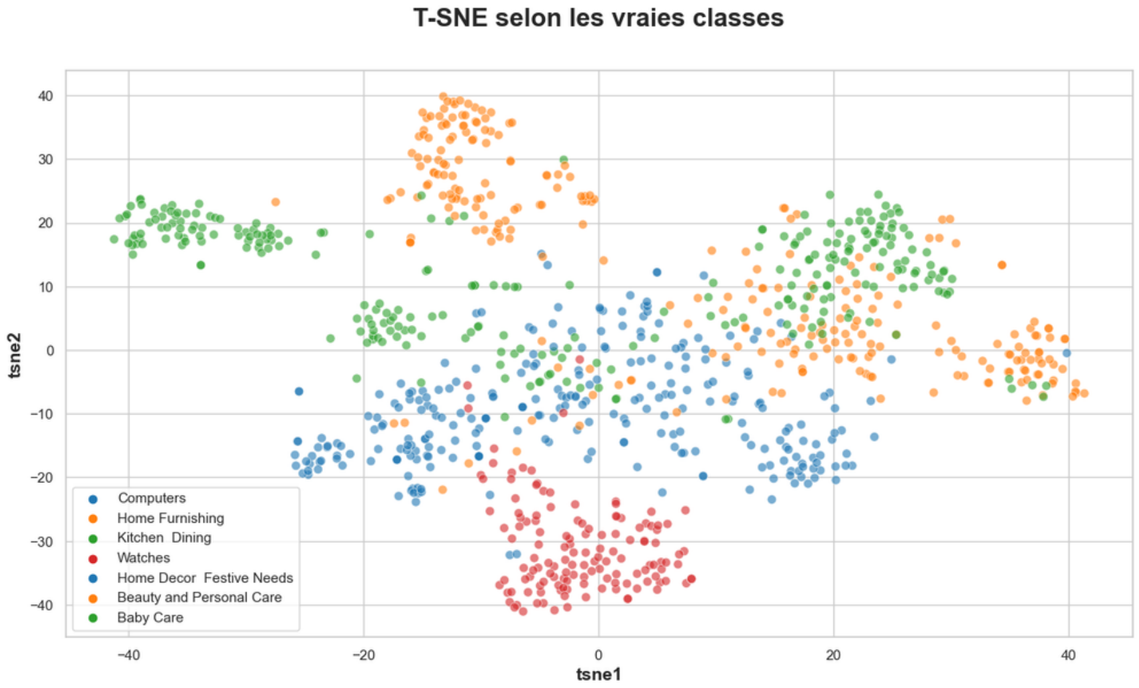
ARI = 0.0609

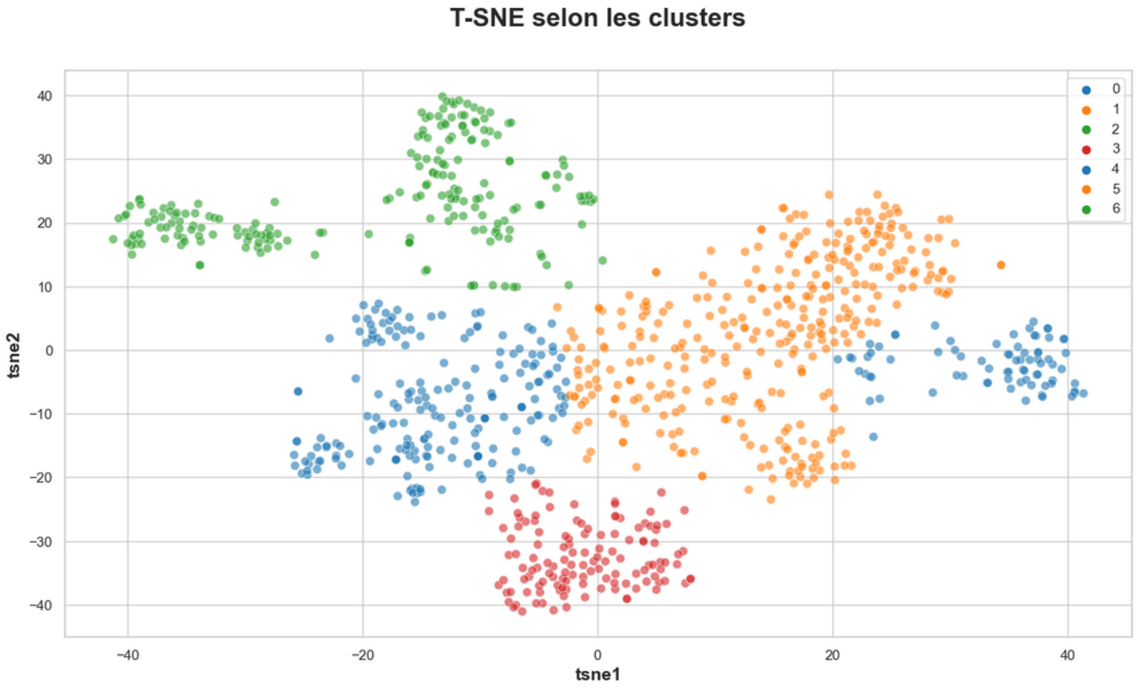




CNN TRANSFER LEARNING

ARI = 0.5406





COMPARAISON DES PERFORMANCES (ARI)

SIFT	CNN Transfer Learning (VGG16)
0.0609	**0.5406**

03.

CLASSIFICATION SUPERVISÉE

DÉMARCHE

- Création de fonctions pour créer les modèles pré-entraînés
- Séparation du jeu de données en jeu d'entraînement, de validation et de test
- Préparation des images par diverses transformations ou techniques d'augmentation des données (ex : rotation, changement d'échelle, ajout de bruit, etc.)
- Création du modèle de réseaux neuronaux convolutifs (CNN)
- Création du callback
- Entraînement du modèle
- Calcul des scores

MESURE DE PERFORMANCE

- **Accuracy** :
  - Mesure la précision globale d'un modèle de classification supervisée en indiquant la proportion d'exemples correctement classés.
  - Un score plus élevé signifie une meilleure classification.

APPROCHES CNN UTILISÉES

— **Classification supervisée simplifiée**

- Modèle VGG16

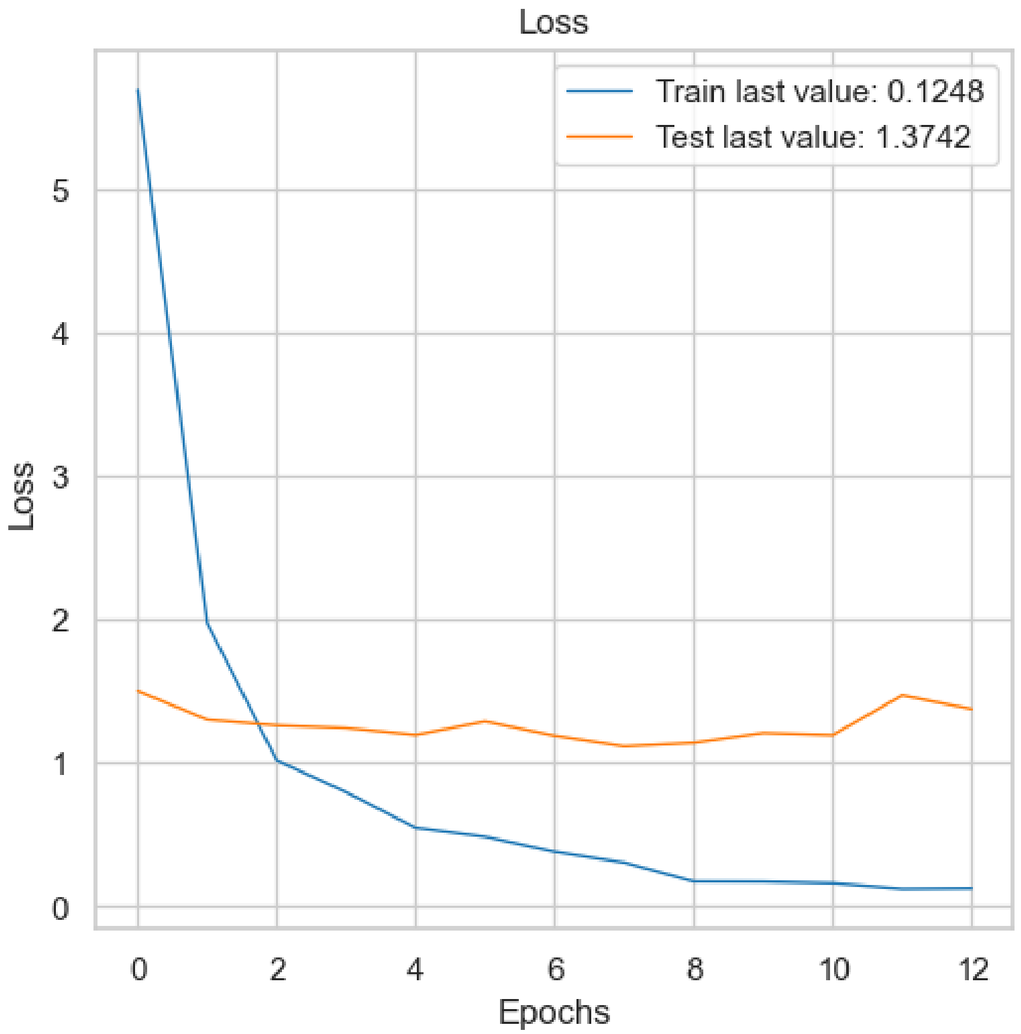
— **ImageDatagenerator avec augmentation des données**

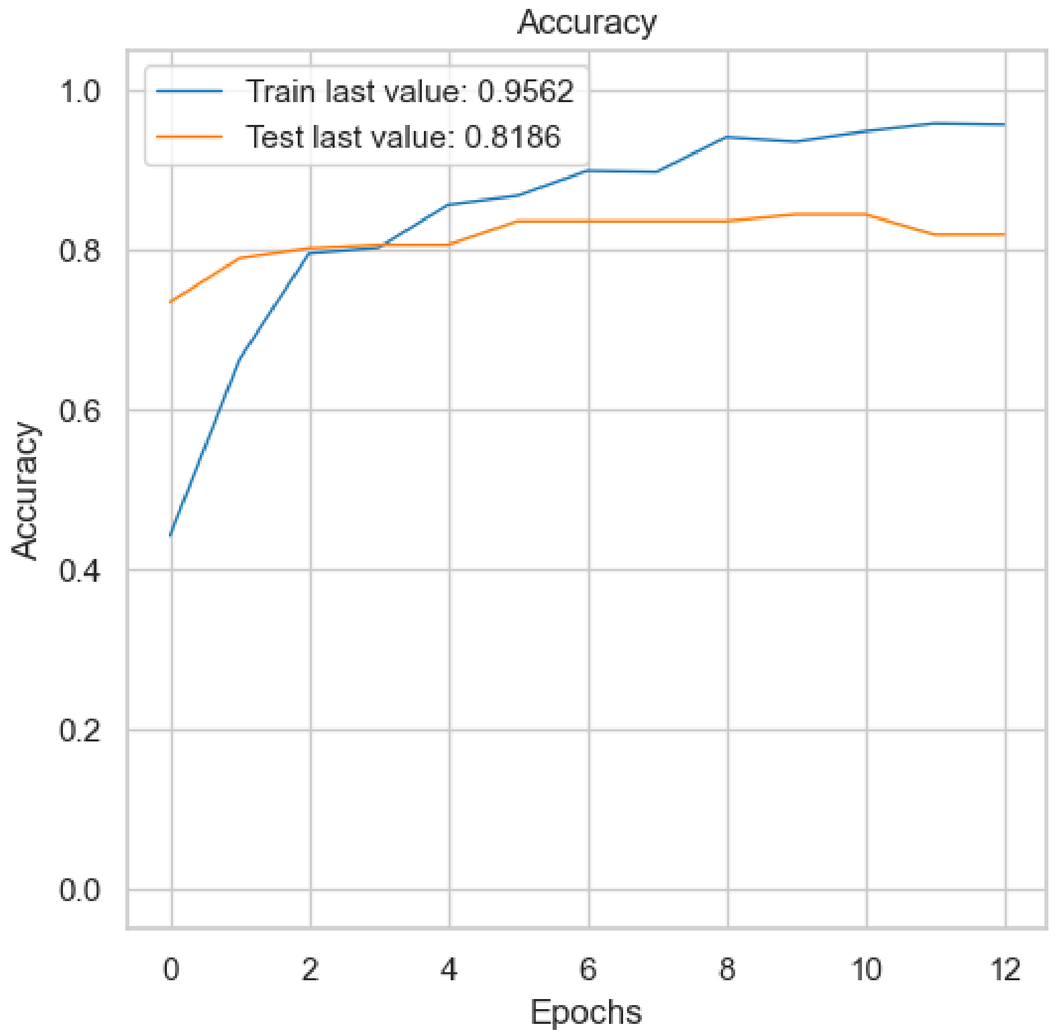
- Modèle VGG16

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

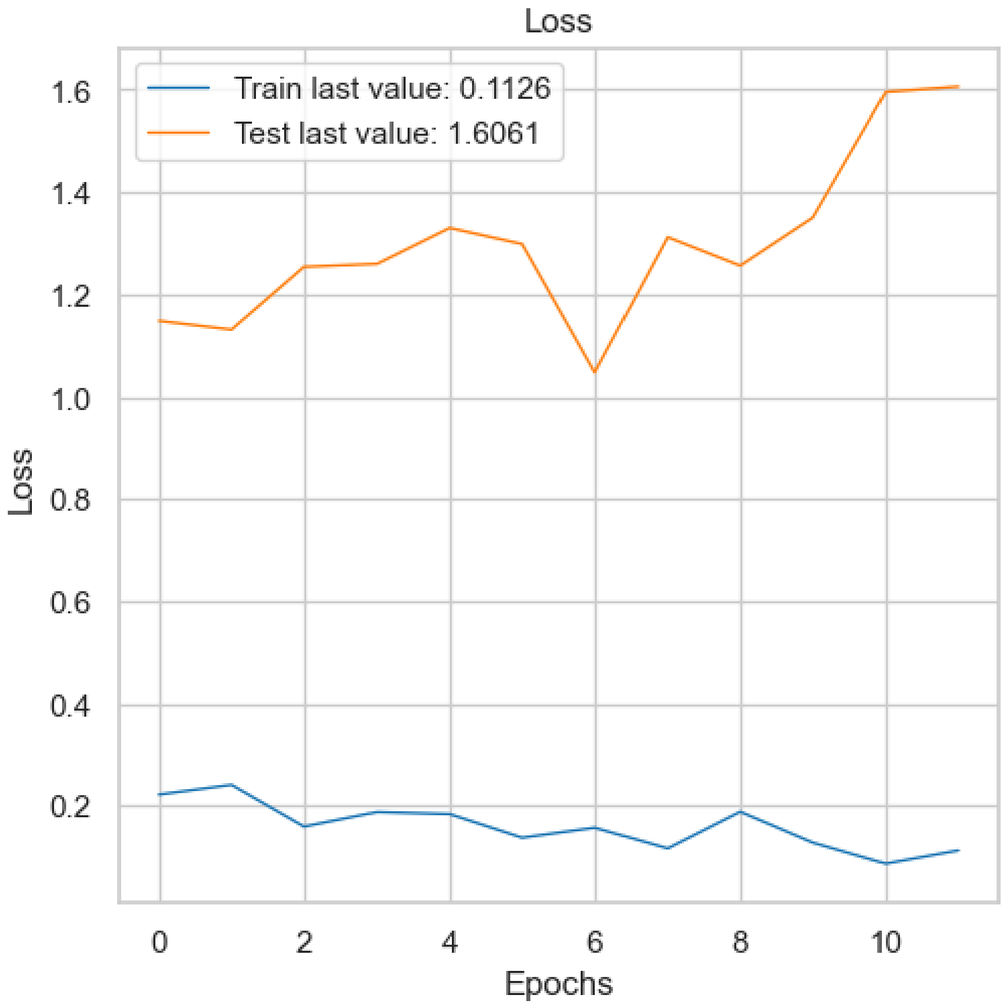
- Modèle VGG16
- Modèle VGG19
- Modèle ResNet50

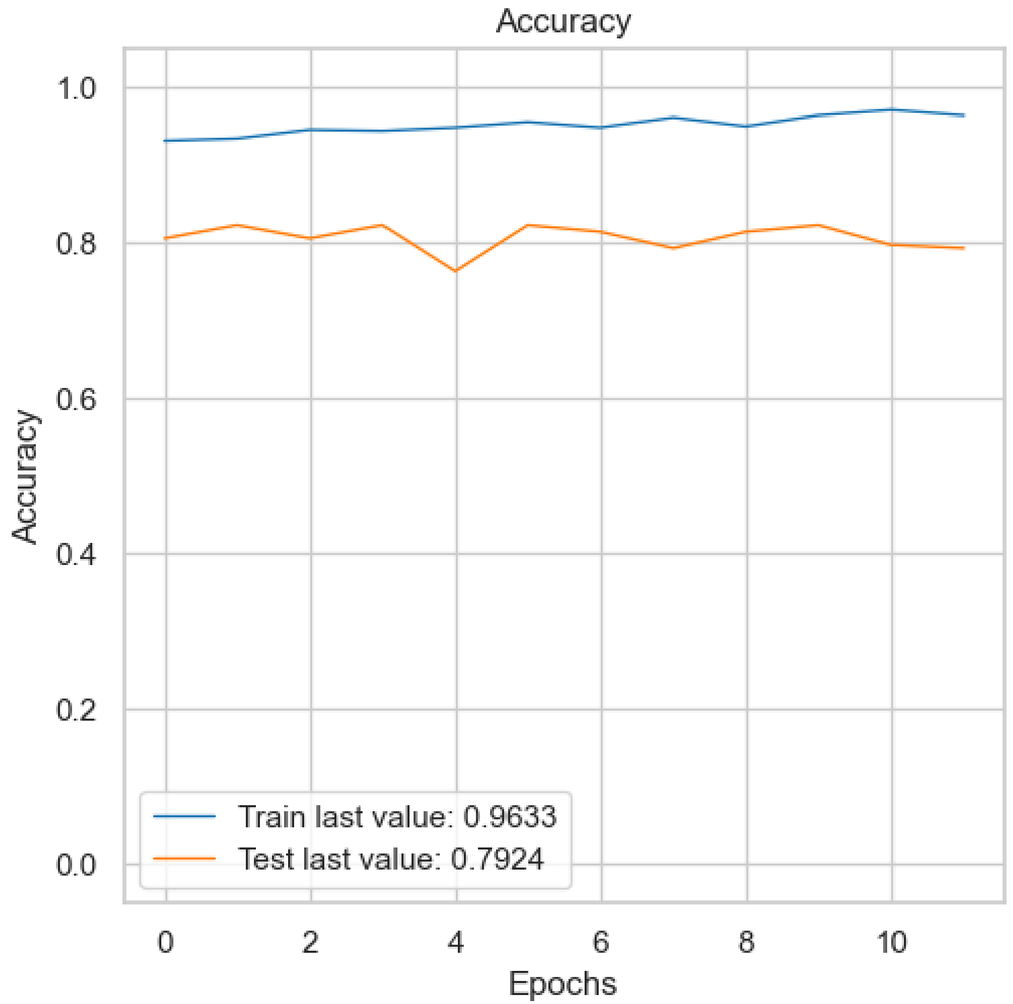
CLASSIFICATION SUPERVISÉE SIMPLIFIÉE (VGG16)



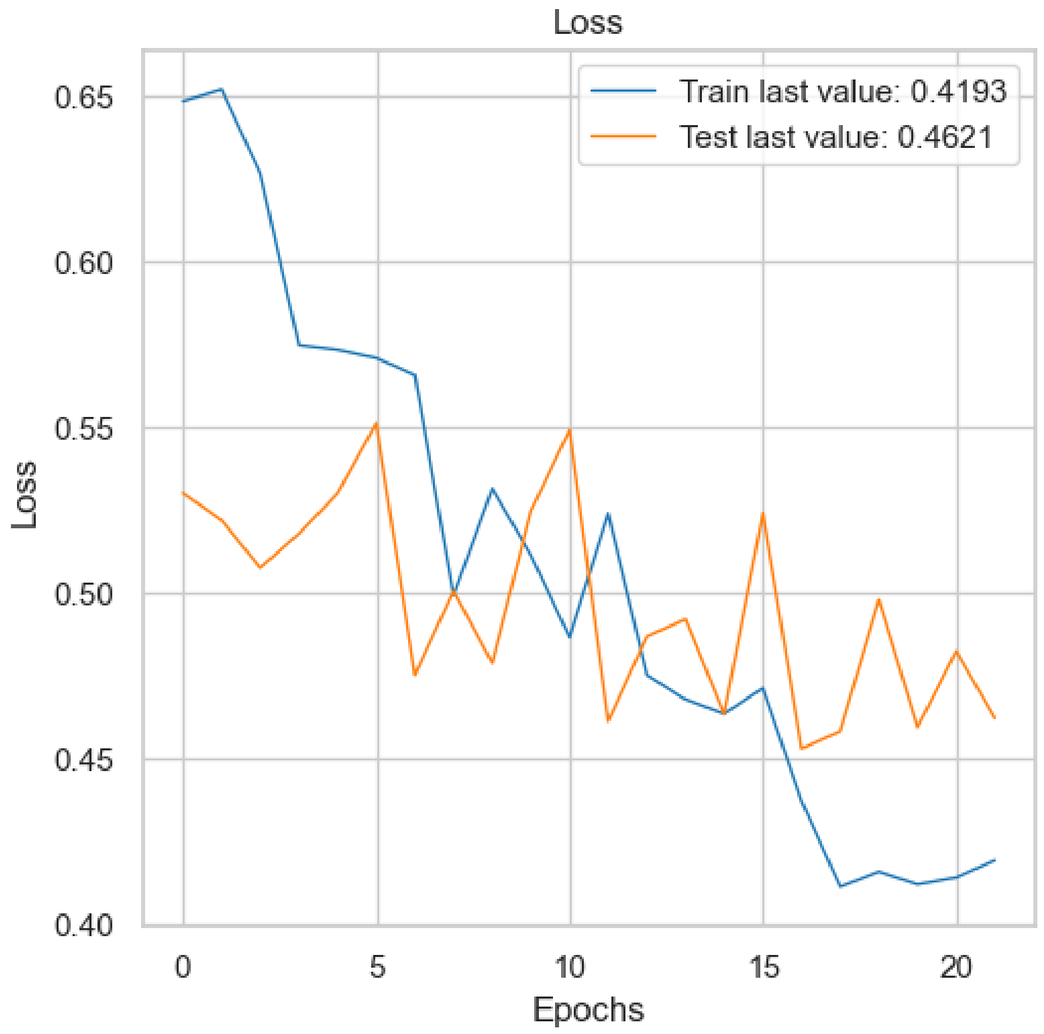


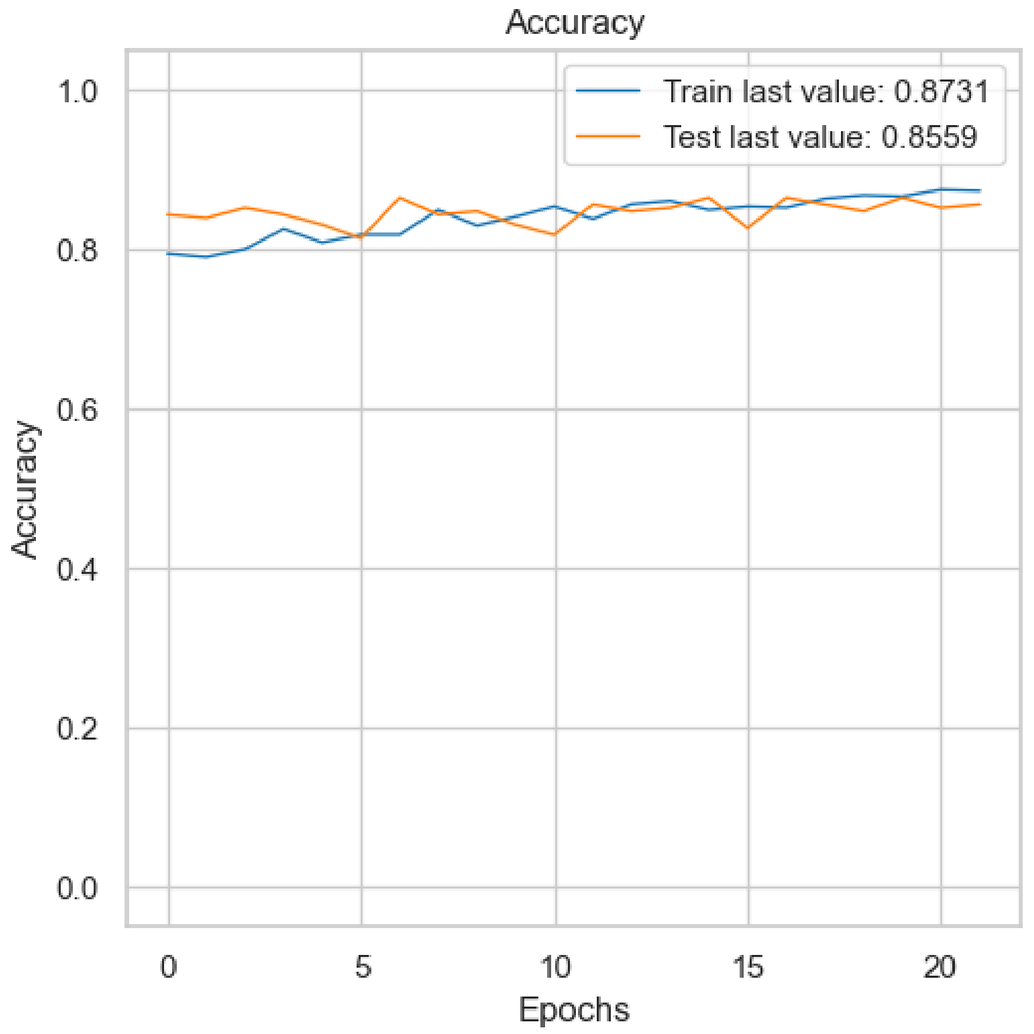
IMAGEDATAGENERATOR AVEC AUGMENTATION DES DONNÉES (VGG16)



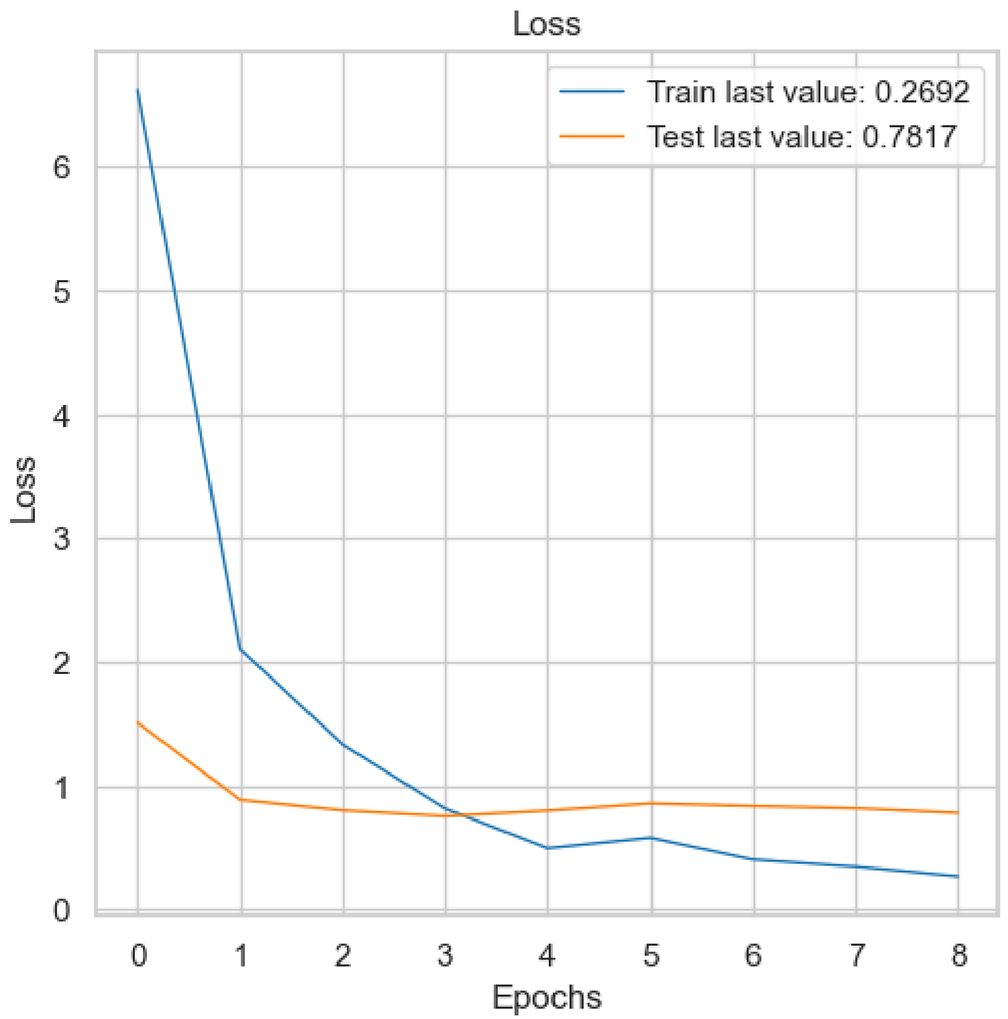


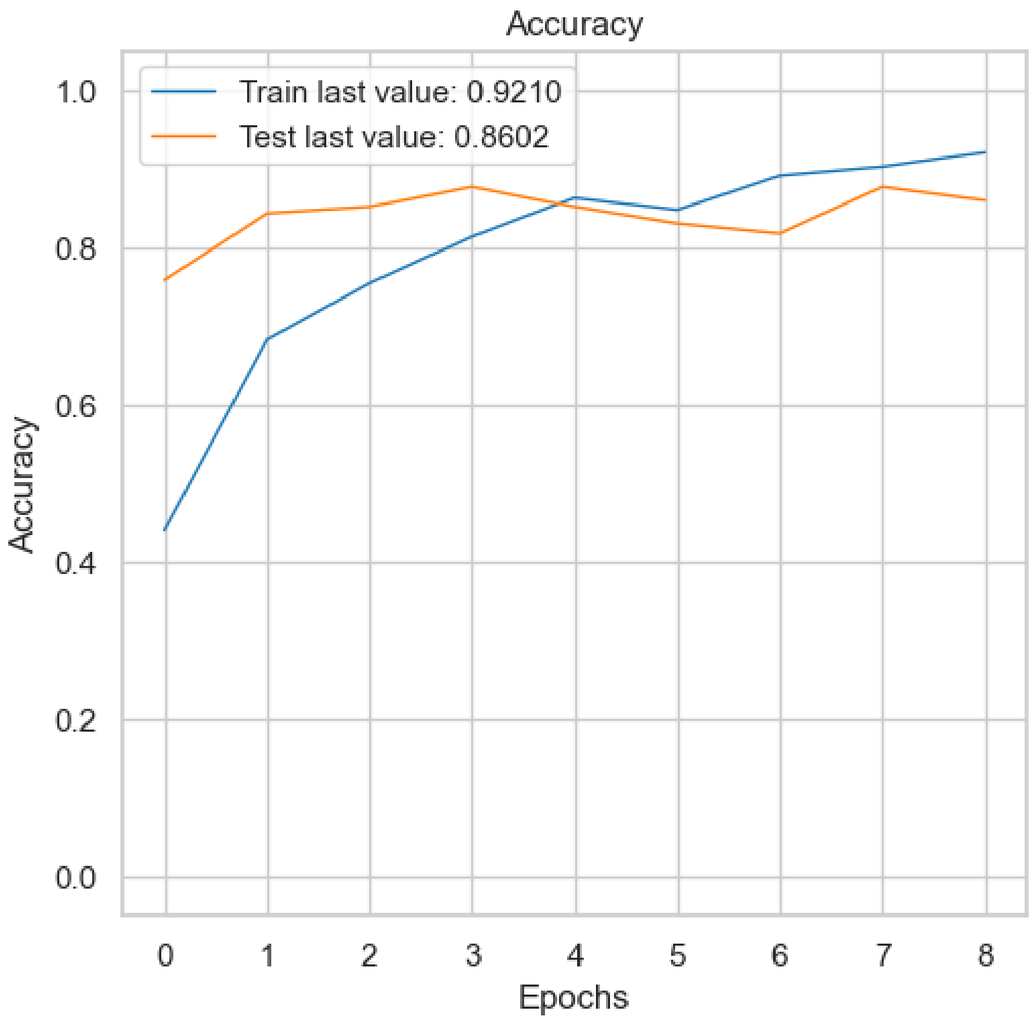
NOUVELLE APPROCHE AVEC AUGMENTATION INTÉGRÉE DES DONNÉES (VGG16)



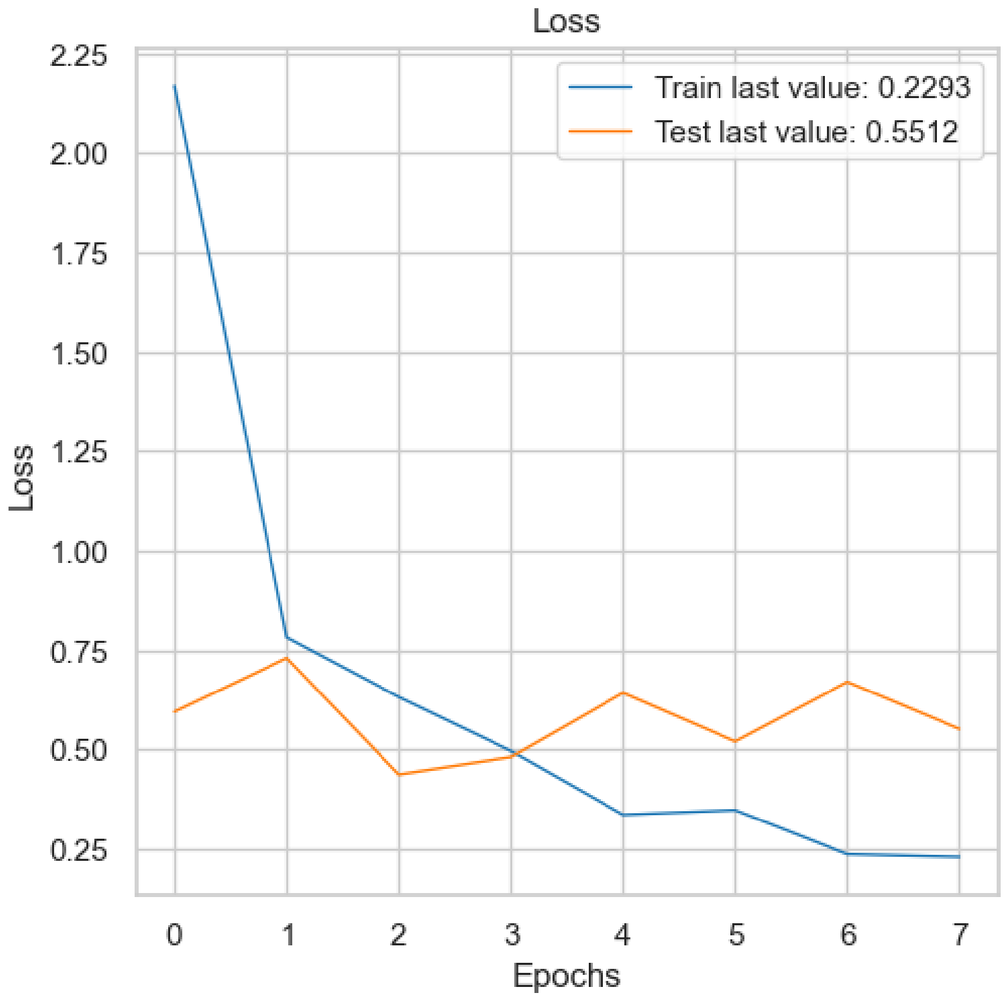


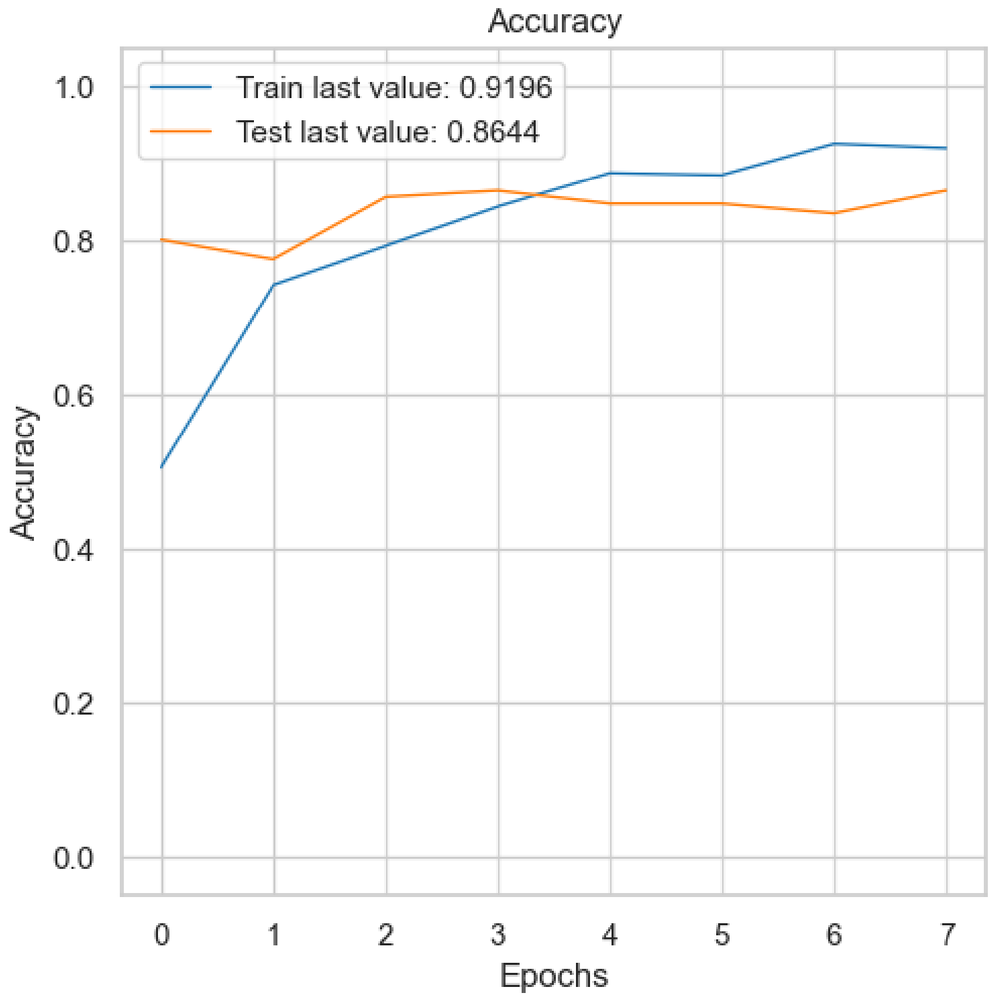
NOUVELLE APPROCHE AVEC AUGMENTATION INTÉGRÉE DES DONNÉES (VGG19)





NOUVELLE APPROCHE AVEC AUGMENTATION INTÉGRÉE DES DONNÉES (RESNET50)





COMPARAISON DES PERFORMANCES

	Classification supervisée simplifiée (VGG16)	ImageDatagenerator avec augmentation des données (VGG16)	Nouvelle approche avec augmentation intégrée des données (VGG16)	Nouvelle approche avec augmentation intégrée des données (VGG19)	Nouvelle approche avec augmentation intégrée des données (ResNet50)
Validation Accuracy (epoch optimal)	0.8354	0.8517	0.8644	**0.8771**	0.8559
Test Accuracy (epoch optimal)	0.6000	0.5333	0.5714	**0.6381**	0.5524
Durée Training (dernier epoch)	336s	348s	312s	311s	**96s**
Durée Validation (dernier epoch)	97s	116s	105s	103s	**35s**

04.

TEST DE L’API

DÉMARCHE

- Compréhension de l'utilité de l'API
- Recherches sur différentes API
- Utilisation de RapidAPI avec API Edamam
- Utilisation d'une requête filtrée sur l'ingrédient "Champagne"
- Extraction au format csv contenant les 10 premiers produits contenant pour chaque produit les données suivantes : foodId, label, category, foodContentsLabel, image

RESPECT DES 5 PRINCIPES DU RGPD

- Licéité, Loyauté et Transparence
- Limitation des Finalités
- Minimisation des Données
- Exactitude des Données
- Conservation des Données

04.

CONCLUSION

- FAISABILITÉ VALIDÉE POUR TEXTE ET IMAGES
- RÉSULTATS SATISFAISANTS DE CLASSIFICATION SUPERVISÉE D’IMAGES
- TEST API VALIDÉ

- COMBINAISON DE PLUSIEURS MODÈLES EN FONCTION DES RÉSULTATS PAR CLASSE

- UTILISATION DE MODÈLES MULTI-MODAUX : ENTRÉE TEXTE ET IMAGES

AXES D'AMÉLIORATION

MERCI !

Des questions ?