

Classification automatique des biens de consommation

Projet #6

Formation : Data Scientist

Oumou Faye

Mentor : Medina Hadjem



Agenda

Introduction
&
Mission

Étude de faisabilité
Texte | Image



Classification supervisée
avec CNN

Test de l'API

Introduction & Mission

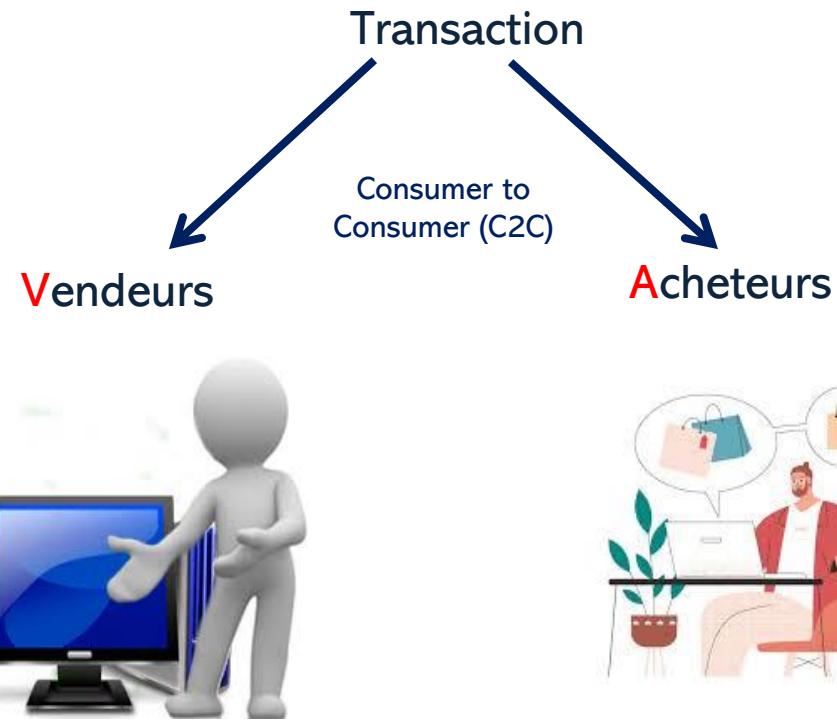


CONTEXTE

Site internet de vente d'articles
(biens de consommation)
en ligne



Place de Marché



Les **vendeurs** proposent des articles à des **acheteurs**

PROCESSUS DE VENTE

1

Le vendeur publie une **photo** de l'article (données visuelles)

&

2

Le vendeur fournit une **description** de l'article (données textuelles)

3

Le vendeur attribue la **catégorie** du produit à vendre
(Sélection d'une catégorie à l'aide d'un menu déroulant)



La catégorie attribuée par le vendeur est souvent **peu fiable**

POSITION



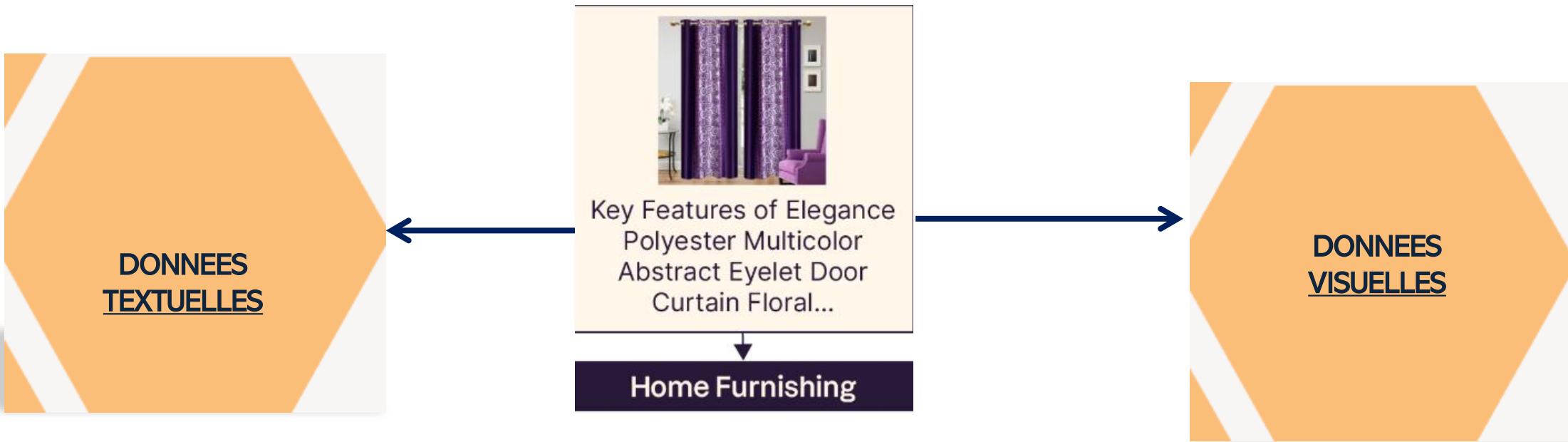
DATA SCIENTIST

MISSION



Étudier la faisabilité d'un moteur de classification des articles mis en vente en se basant sur les données textuelles et visuelles

Étude de classification multimodale



Dataset : **1.050 lignes réparties en 7 catégories**:
150 enregistrements pour chaque catégorie

1. Home Furnishing (#150)
2. Baby Care (#150)
3. Watches (#150)
4. Home Decor & Festive Needs (#150)
5. Kitchen & Dining (#150)
6. Beauty and Personal Care (#150)
7. Computer (#150)





Pipeline de faisabilité de la classification du **texte** par catégorie



TRANSFORMATION

[Collecte des données]

- 1. Nom du produit
- 2. Description du produit
- 1. Catégorie du produit (#7)



[Nettoyage des données]

- 1. Suppression des caractères spéciaux
- 1. Conversion des textes en minuscules



[Prétraitement textuel]

- 1. Tokenisation : Découpage en mots individuels
- 2. Stop Words: Suppression des mots vides
- 3. Lemmatisation: réduction des mots à leur forme canonique

[Extraction des Features]

(Basiques)

- 1. Word Count : Comptage simple de mot (nb. de mots)
- 2. BoW : Matrice de comptage de nombre des mots pour chaque document
- 3. Tf-IDF : Pondération des mots par leur importance

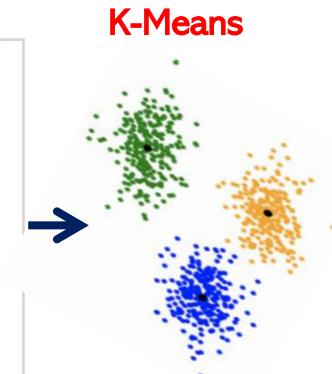
(Avancées)

- 1. Word2vec : Vecteur denses pour chaque mot
- 2. BERT : Embeddings riches et contextuels pour les phrases
- 3. USE : Production d'Embeddings globaux (capture du sens de la phrase)

[Réduction de dimension 2D] → [Clustering + ARI]

- 1. ACP : Composante principale #2
- 2. T-SNE : t-distributed Stochastic Neighbor Embedding)
- 3. SVD : Singular Value Decomposition
- 4. Autoencodeur : réseau de neurones non supervisé
- 5. UMAP : (Uniform Manifold Approximation and Projection)

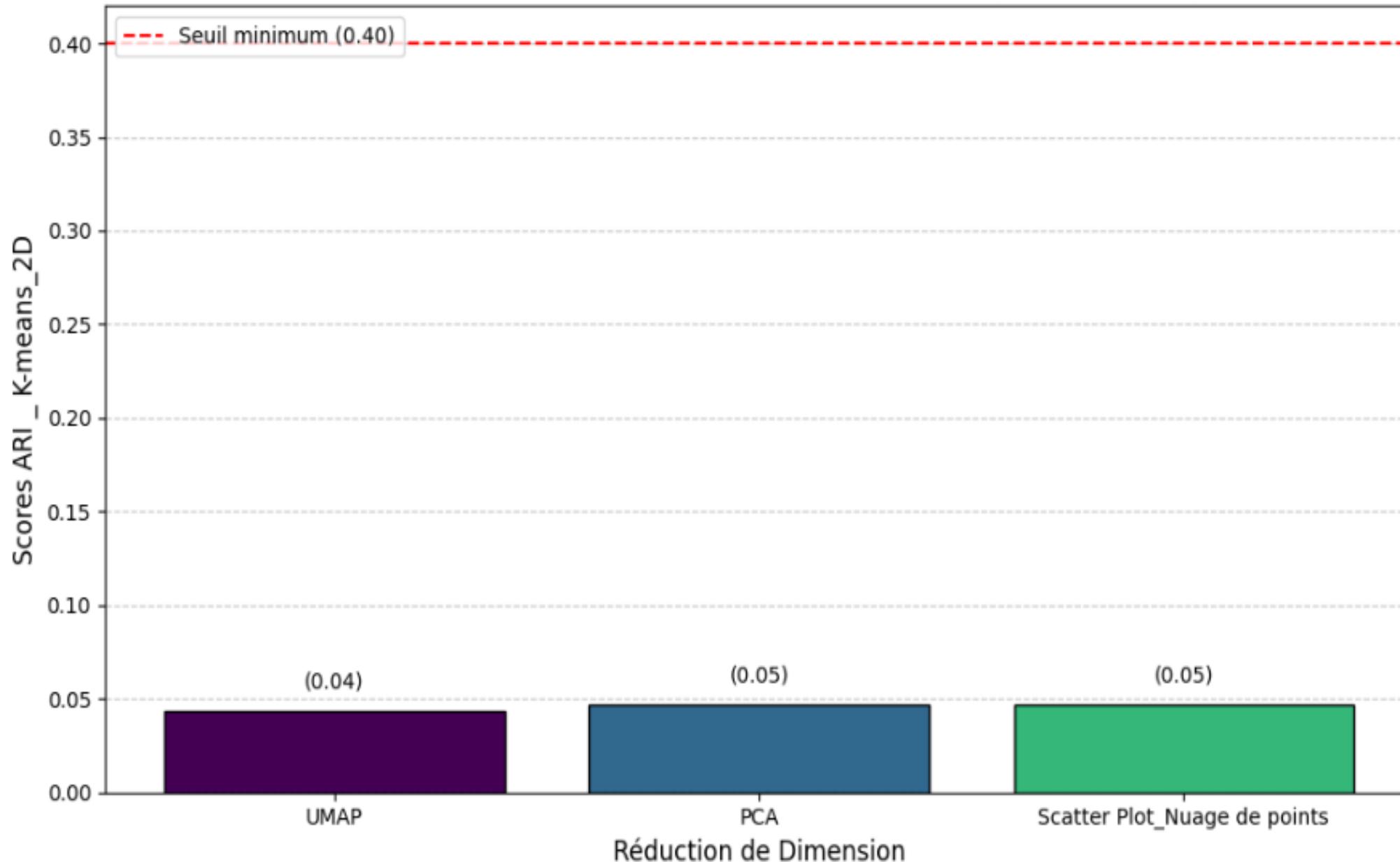
K-Means



Performance K-means (Réd.2D) – Méthodes **basiques**



Scores ARI K-Means sur feature ****Word Count**** (Réduction 2D)

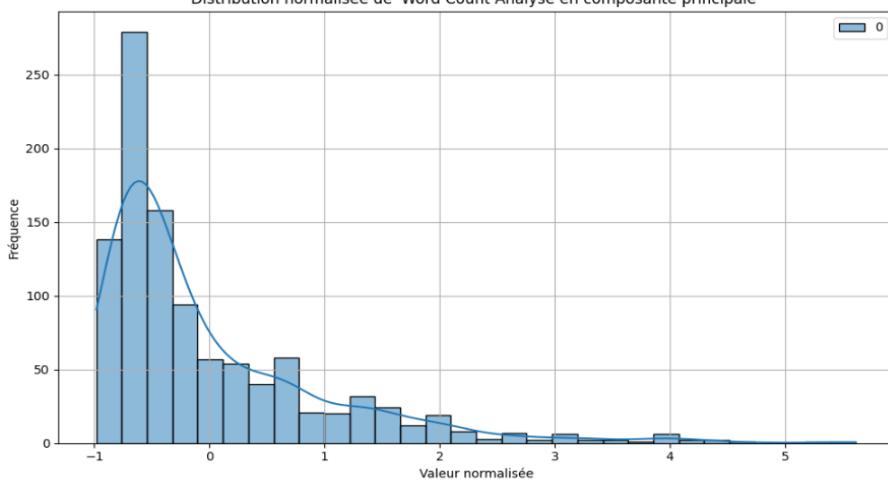


Projection (Réd.2D) – Méthodes Basique – Word Count

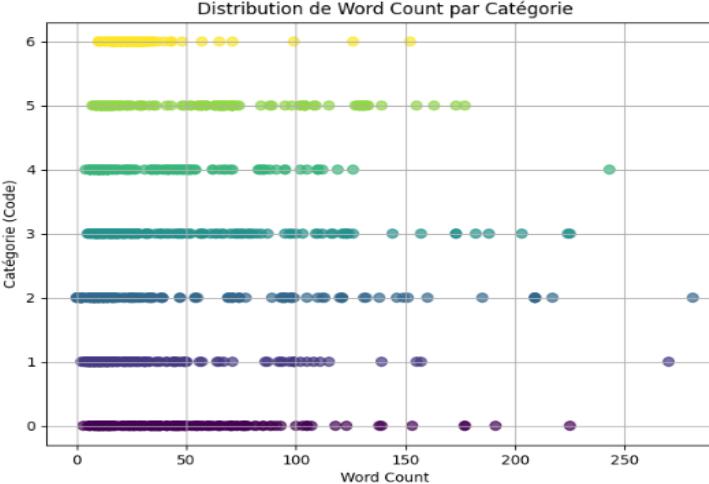


Projection (Réd.2D's)

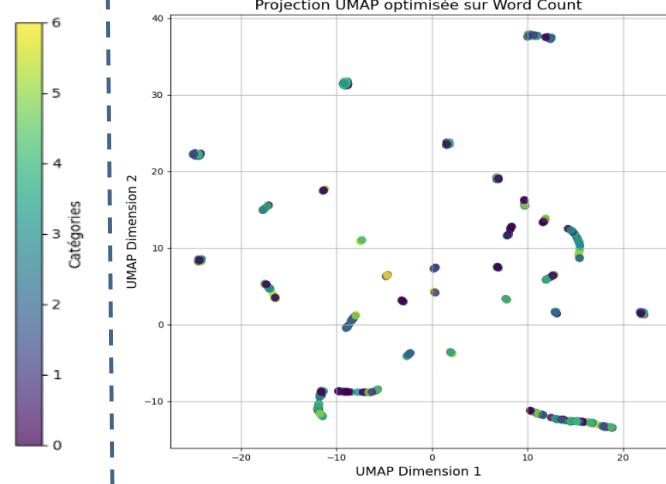
Distribution normalisée de 'Word Count Analyse en composante principale'



Distribution de Word Count par Catégorie

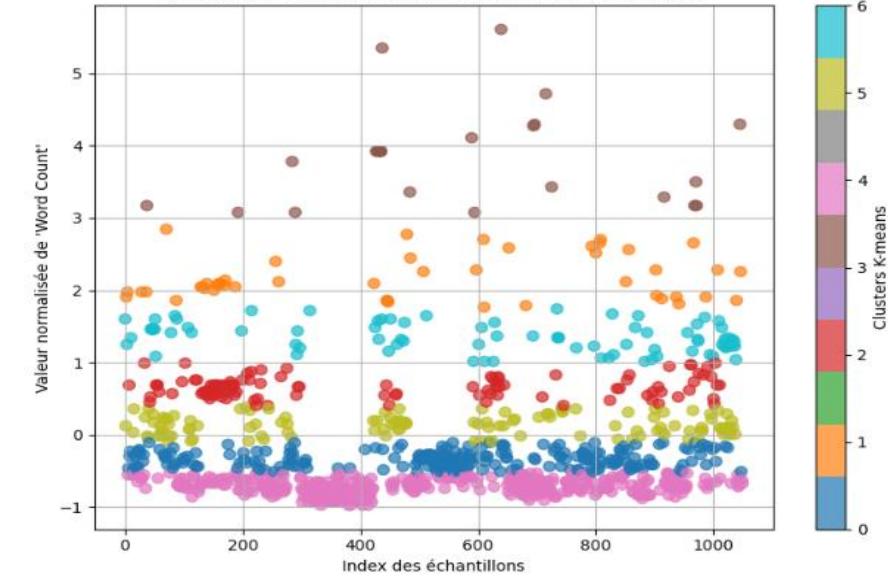


Projection UMAP optimisée sur Word Count



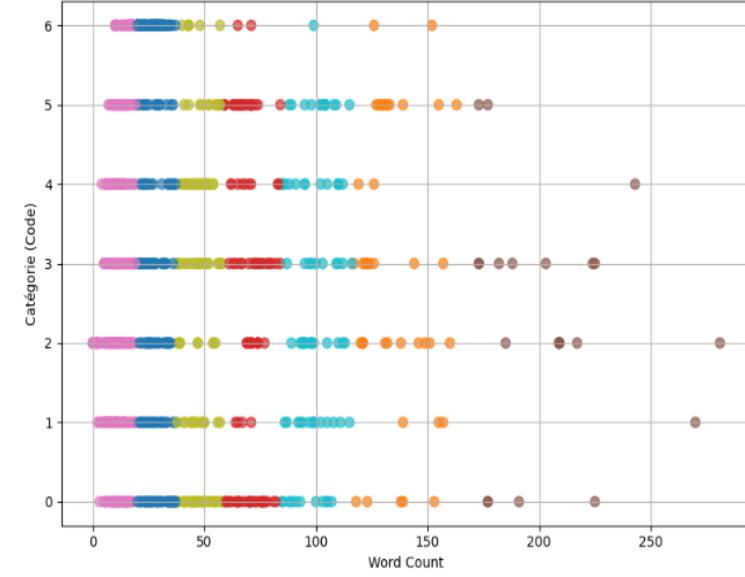
Clustering [ACP]

K-means sur les données ACP normalisées 'Word Count'



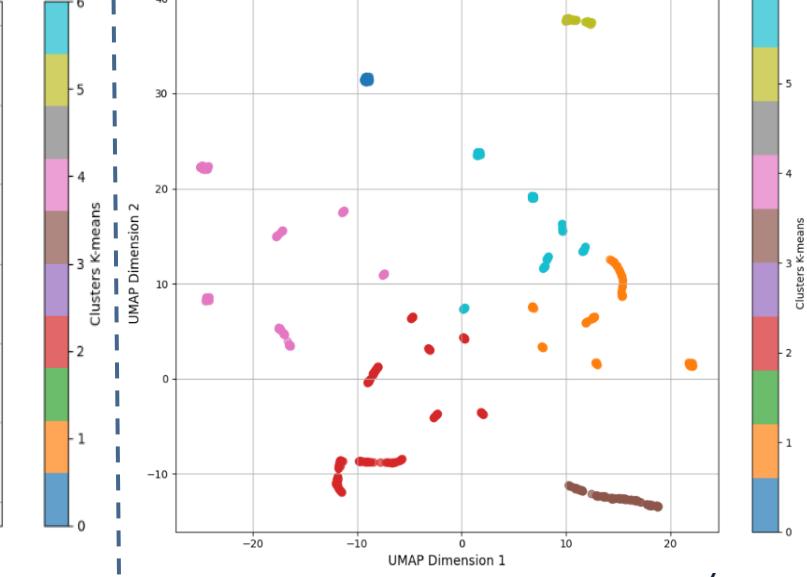
Clustering [Word Count]

K-means sur Word Count



Clustering [UMAP]

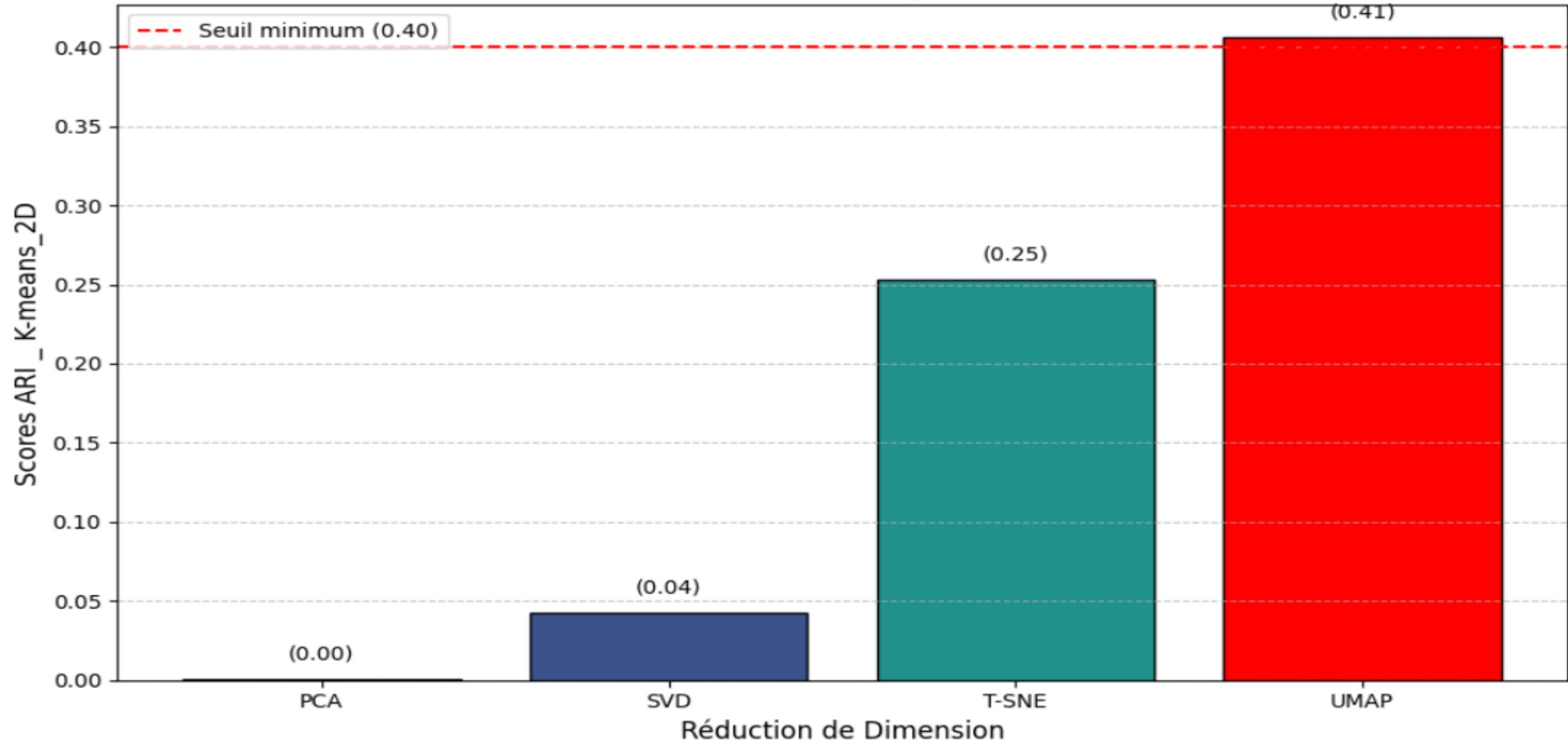
K-means avec 7 clusters sur les dimensions UMAP (Word_Count)



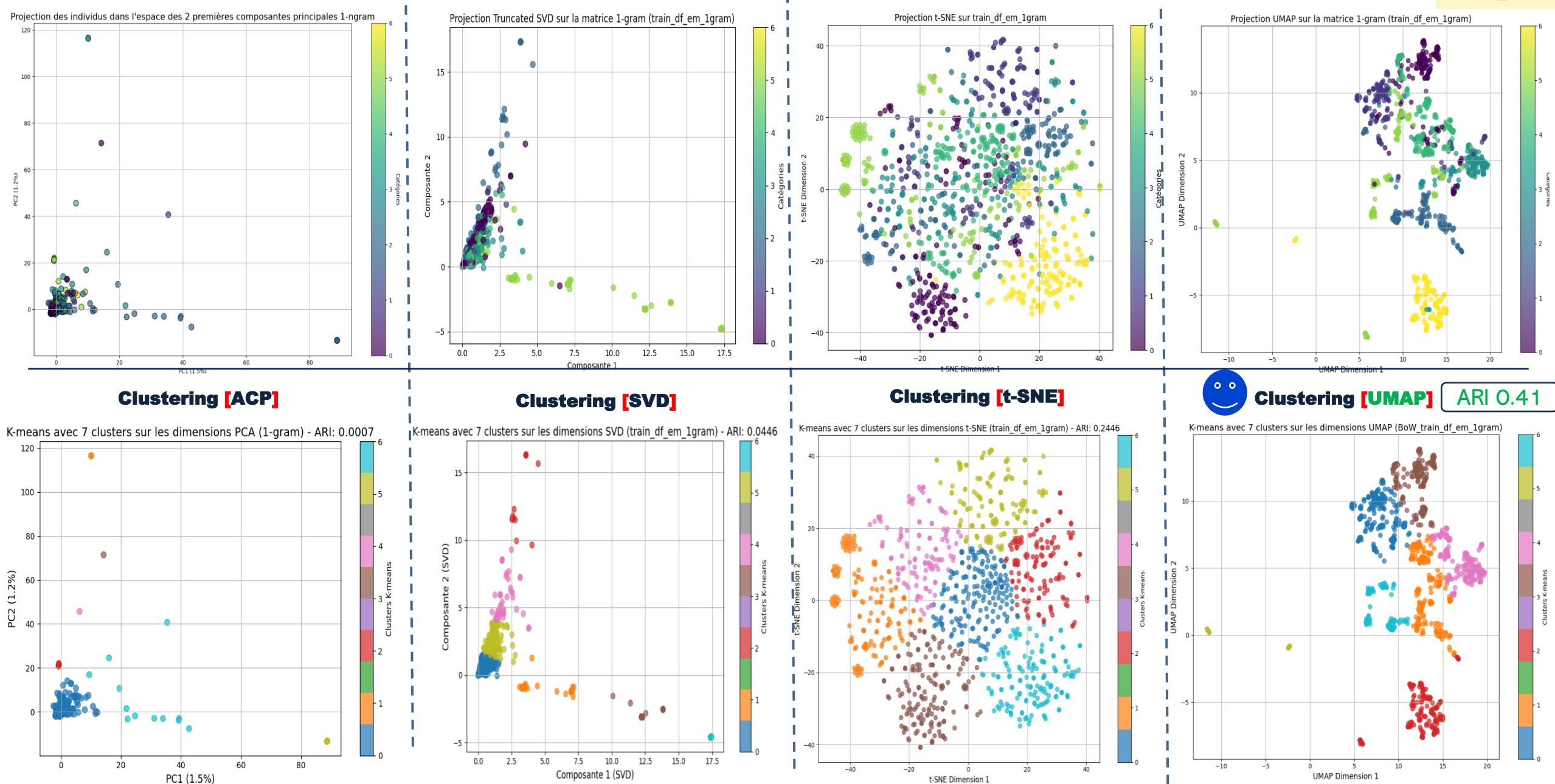
Performance K-means (Réd.2D) – Méthodes basiques



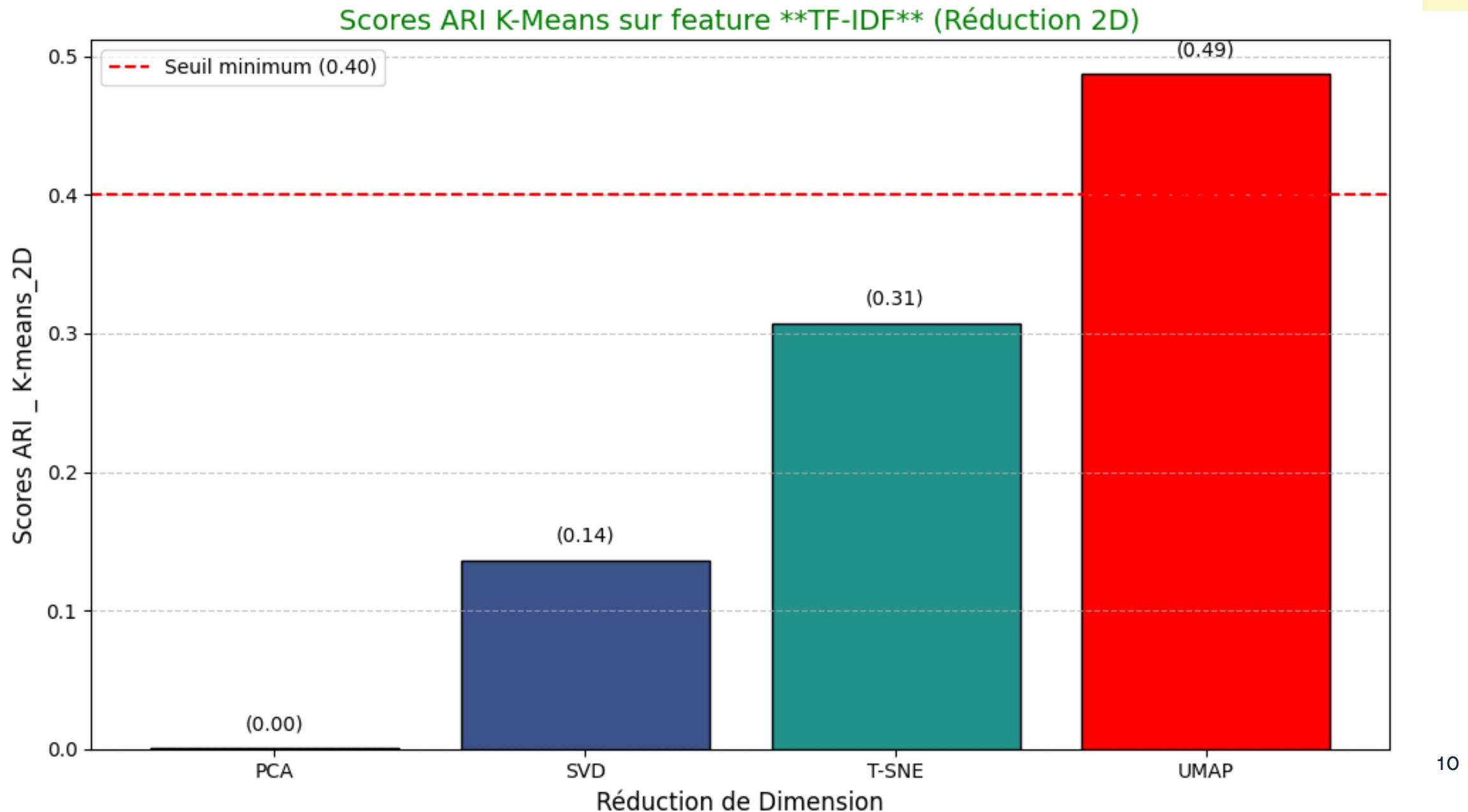
Scores ARI K-Means sur feature ****Bag of Words**** (Réduction 2D)



Projection (Réd.2D's) – Méthodes Basiques - Bag of Words



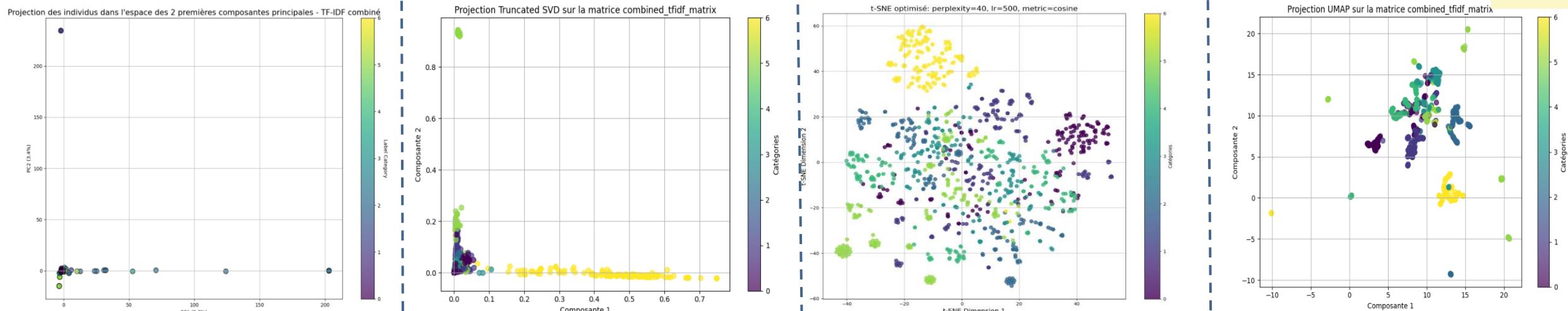
Performance K-means (Réd.2D) – Méthodes basiques



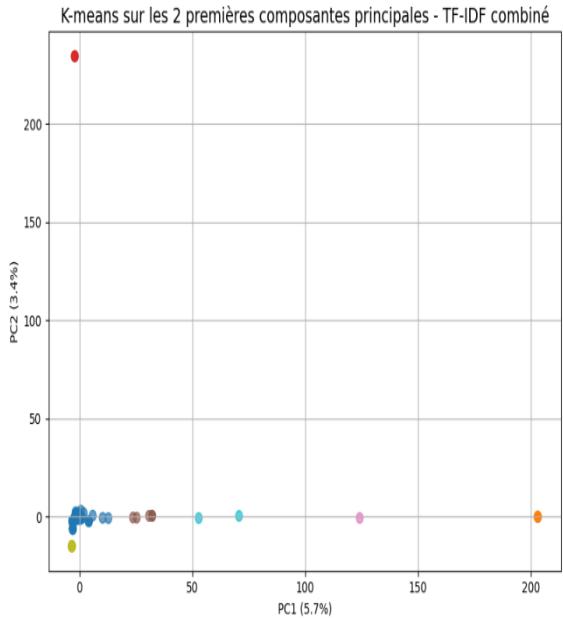
Projection (Réd.2D) – Méthodes Basiques – TF IDF



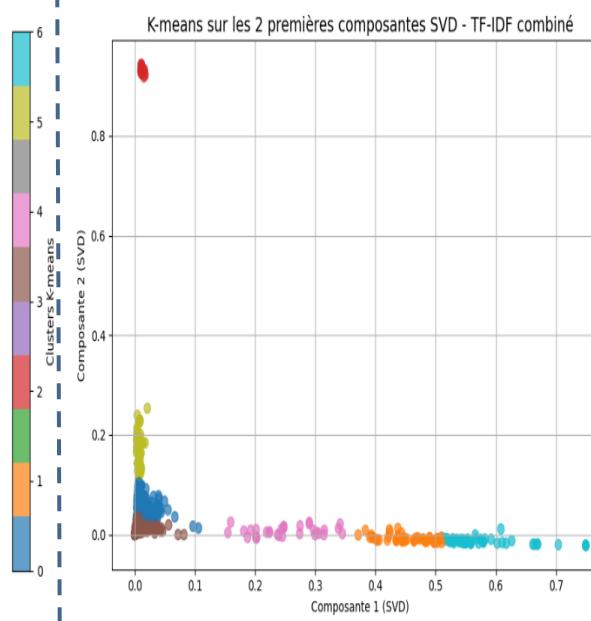
Projection (Réd.2D's)



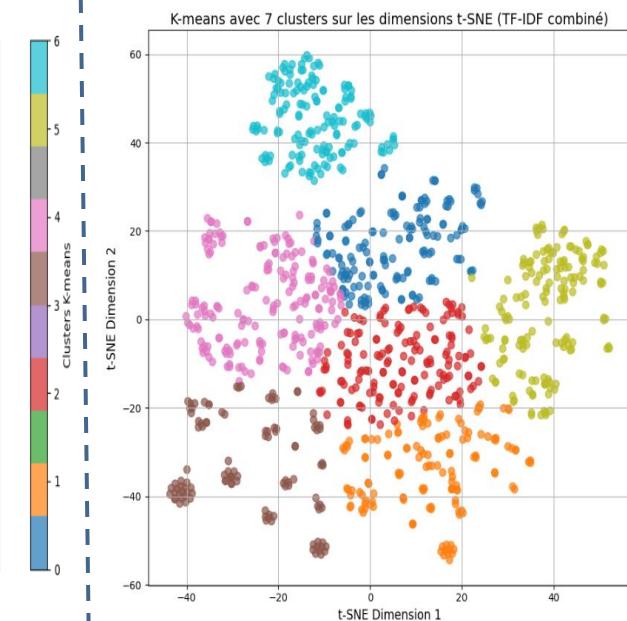
Clustering [ACP]



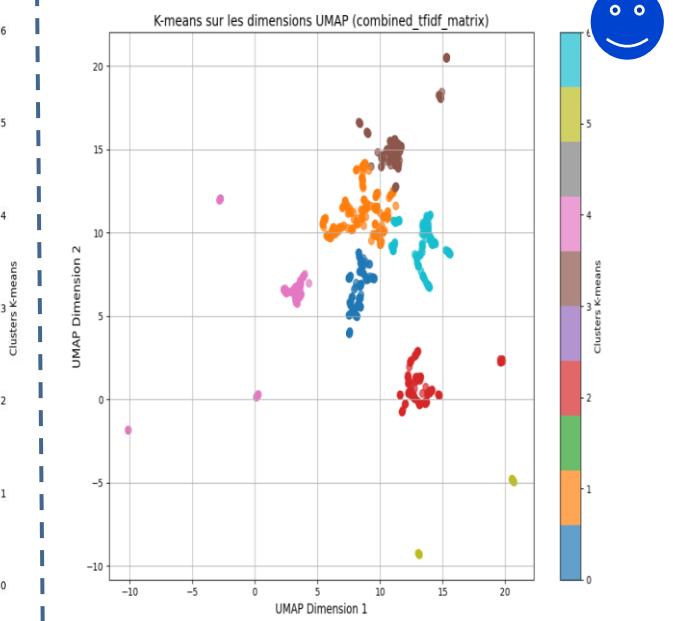
Clustering [SVD]



Clustering [t-SNE]



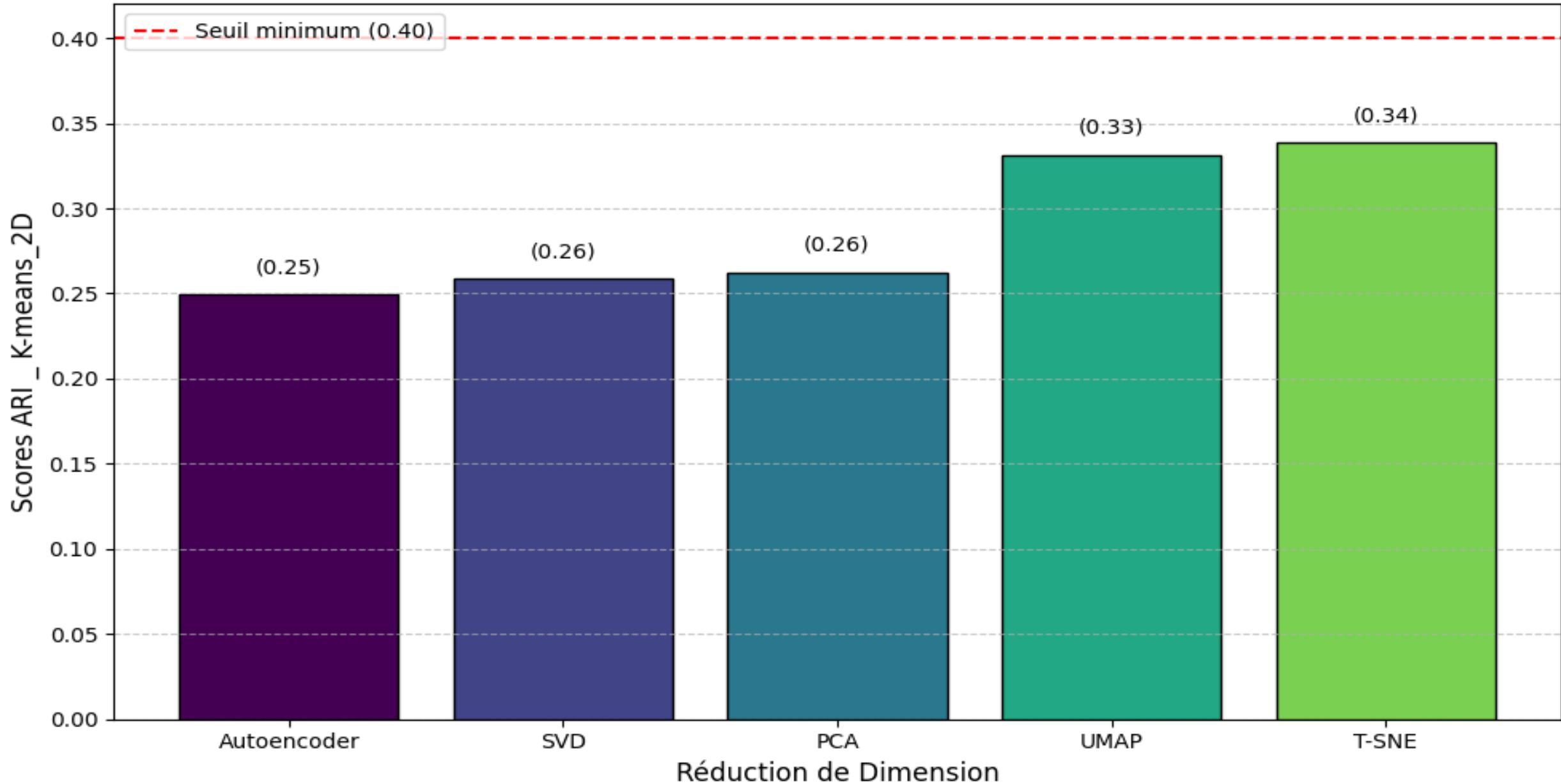
Clustering [UMAP]



Performance K-means (Réd.2D) – Méthodes avancées



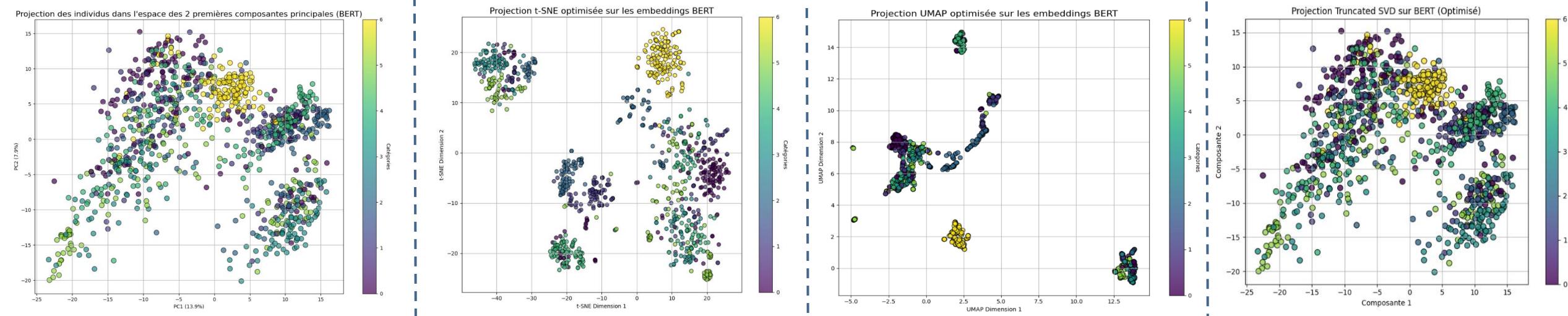
Scores ARI K-Means sur feature ****BERT**** (Réduction 2D)



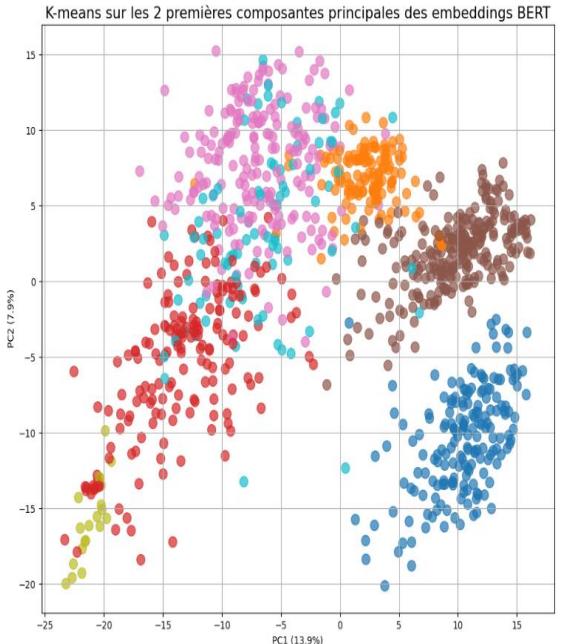
Projection (Réd.2D's) – Méthodes Avancées – BERT



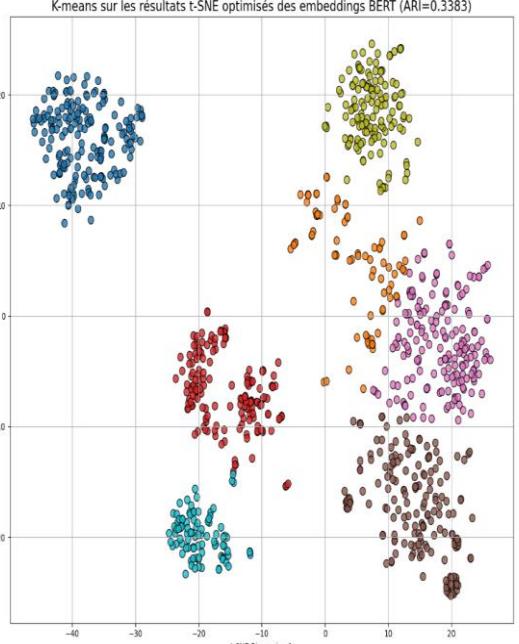
Projection (Réd.2D's)



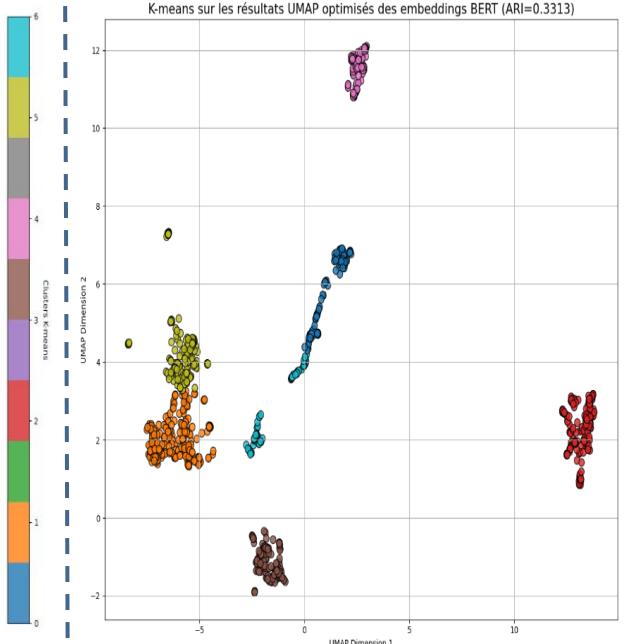
Clustering [ACP]



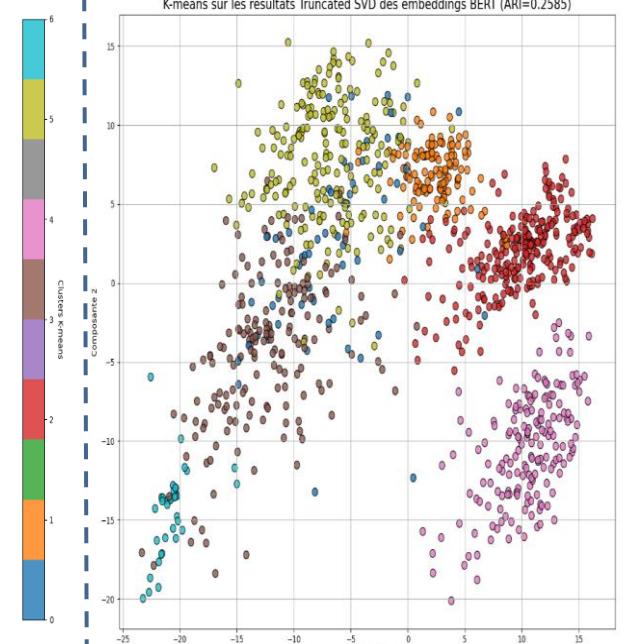
Clustering [t-SNE]



Clustering [UMAP]



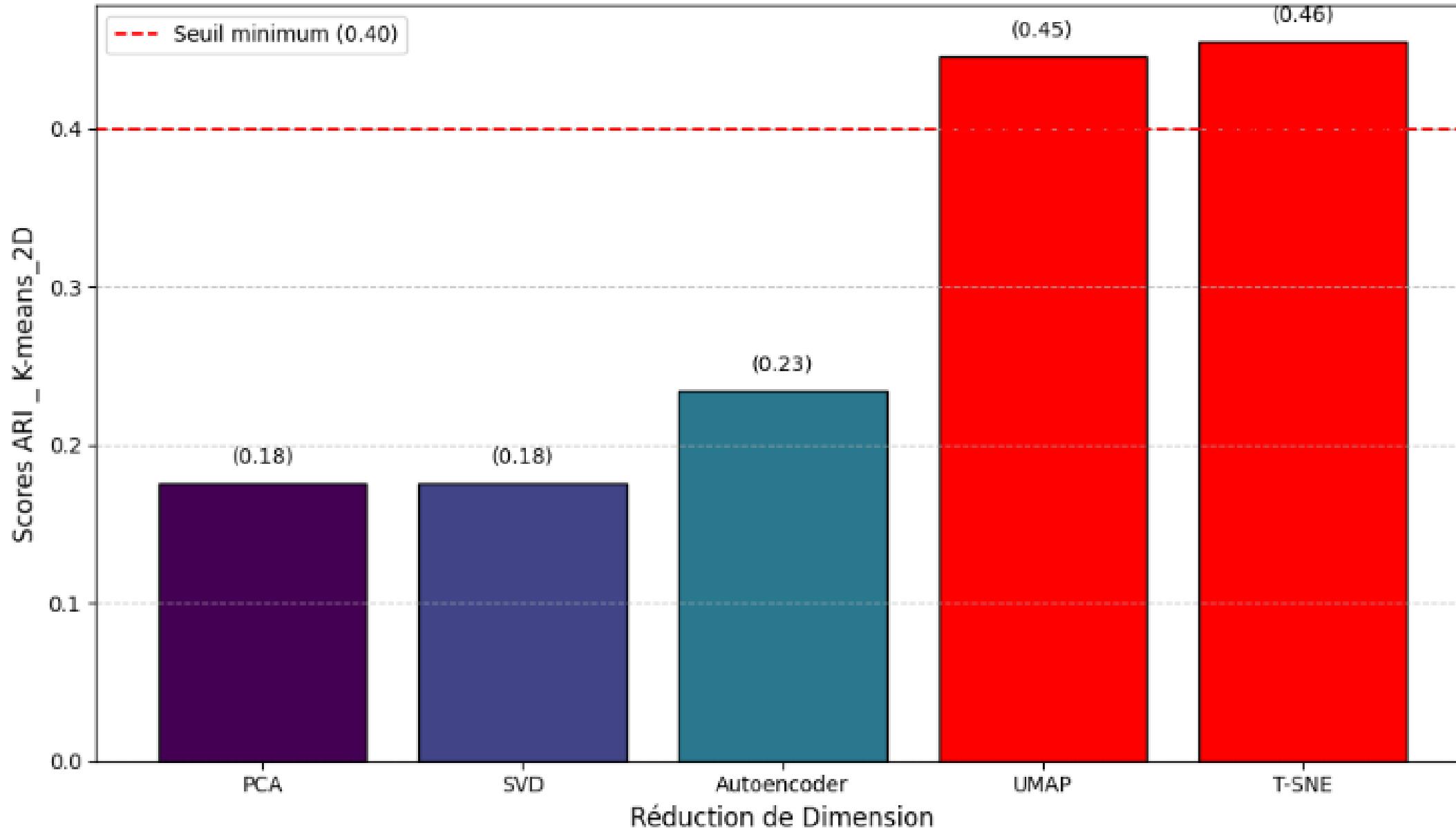
Clustering [SVD]



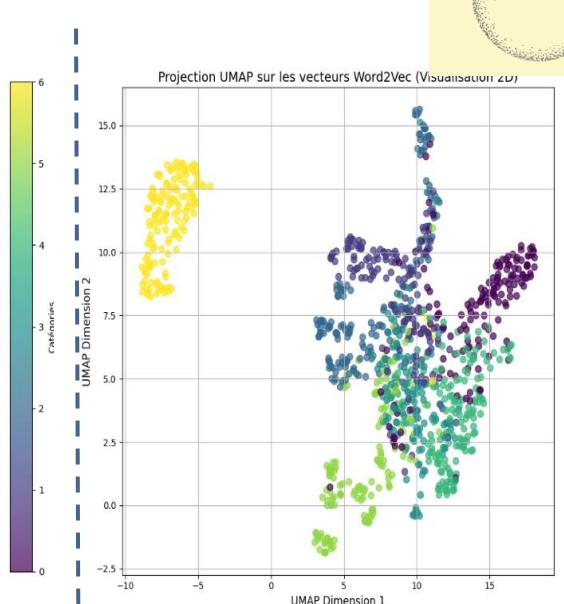
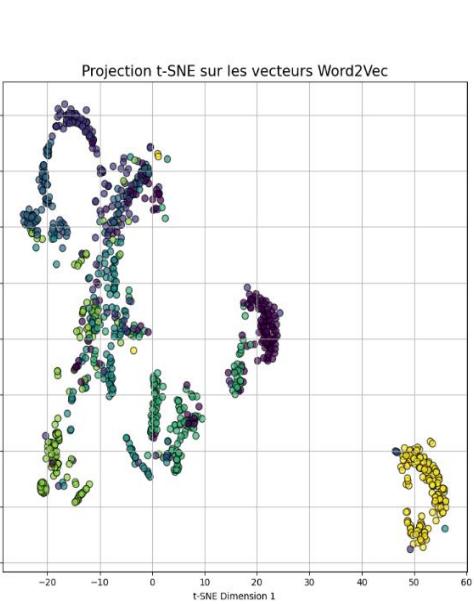
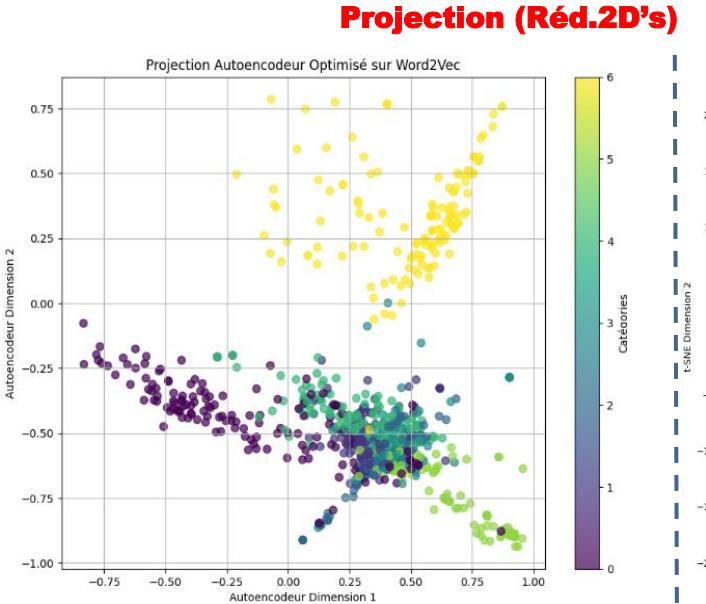
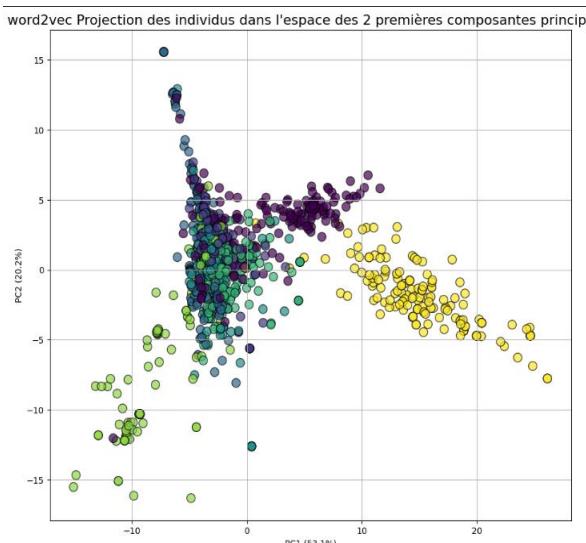
Performance K-means (Réd.2D) – Méthodes avancées



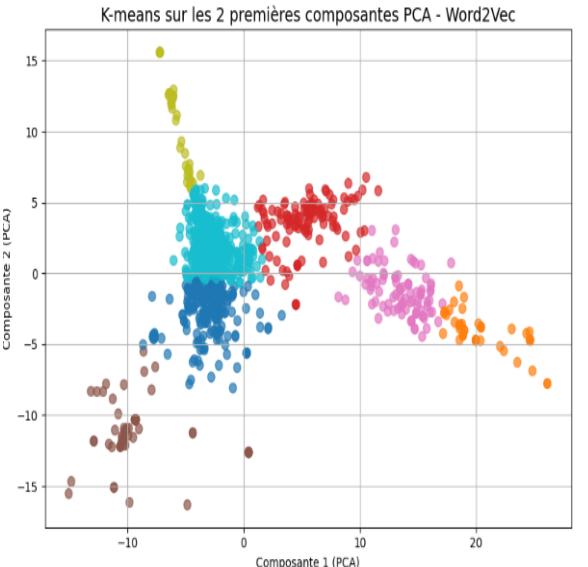
Scores ARI K-Means sur feature ****Word2Vec**** (Réduction 2D)



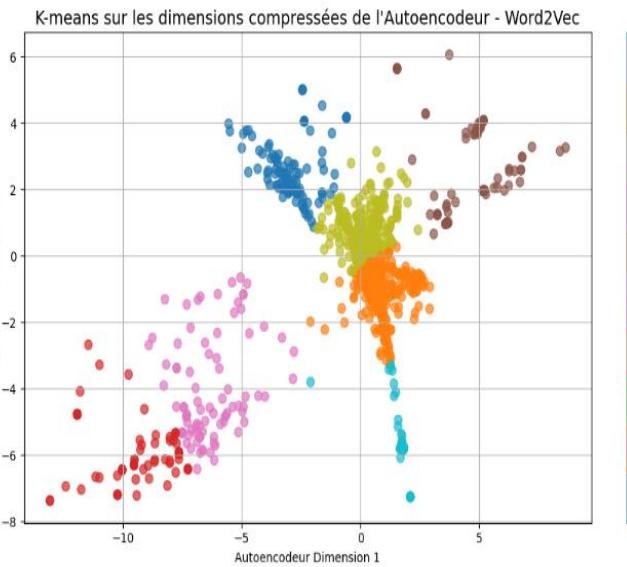
Projection (Réd.2D's) – Méthodes Avancées – Word2Vec



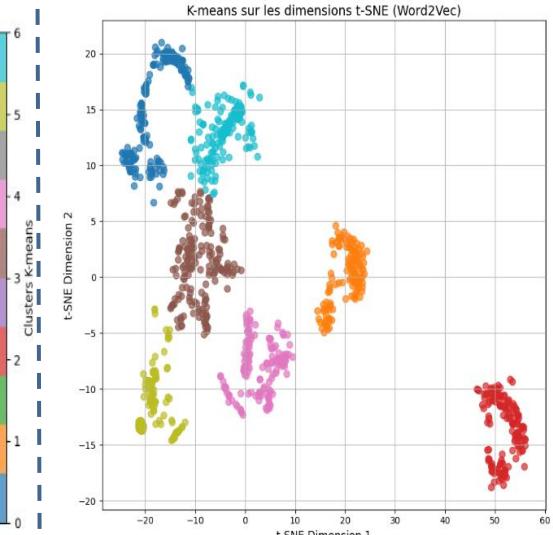
Clustering [ACP]



Clustering [AUTOENCODEUR]



Clustering [t-SNE]

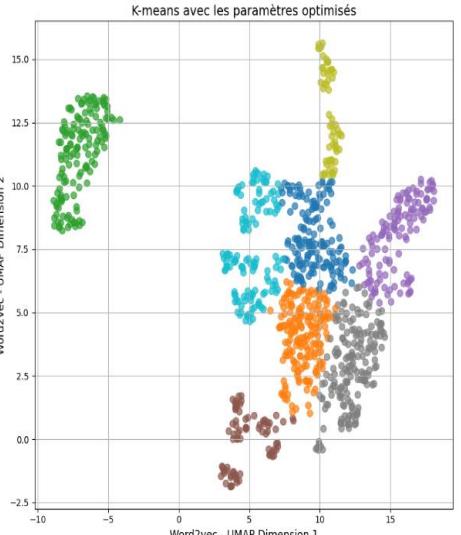


ARI 0.45

ARI 0.45



Clustering [UMAP]



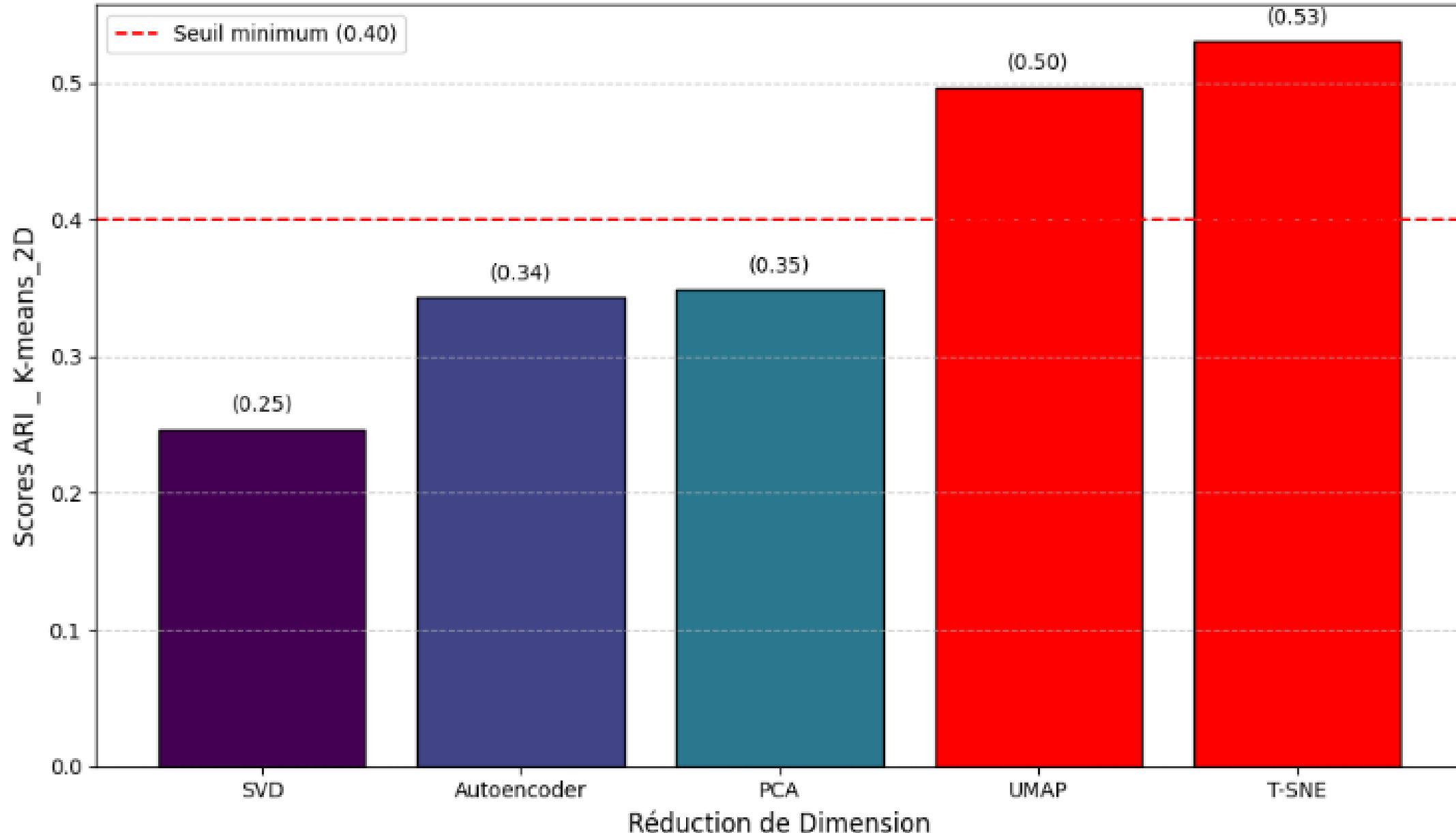
ARI 0.45



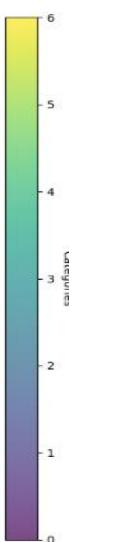
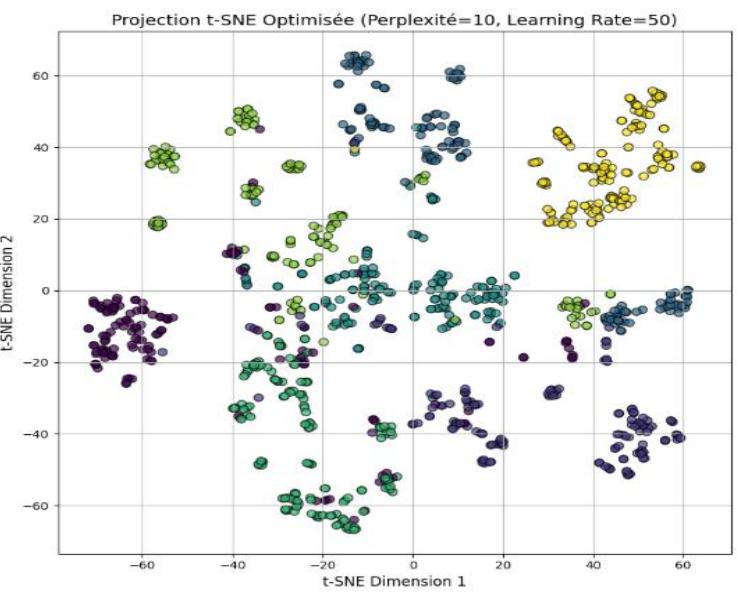
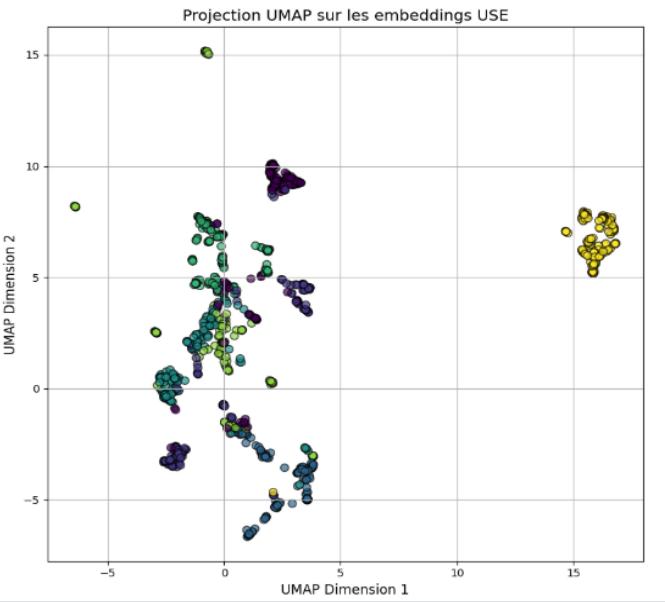
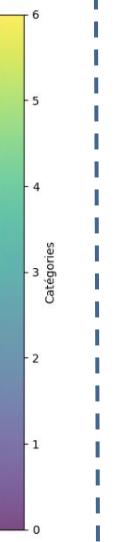
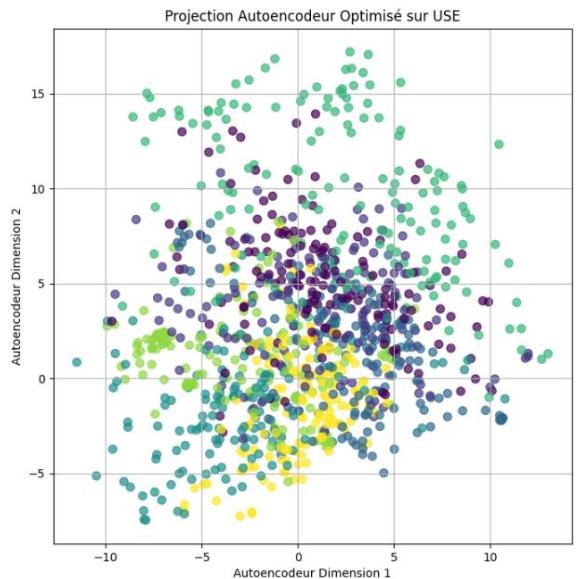
Performance K-means (Réd.2D) – Méthodes Avancées



Scores ARI K-Means sur feature ****USE**** (Réduction 2D)

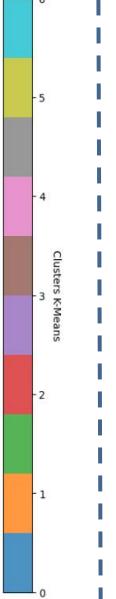
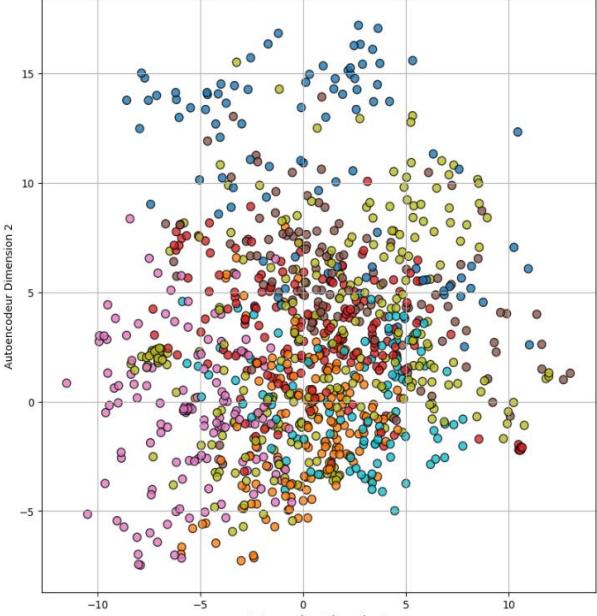


Projection (Réd.2D's) – Méthodes Avancées – USE



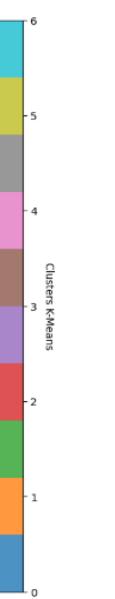
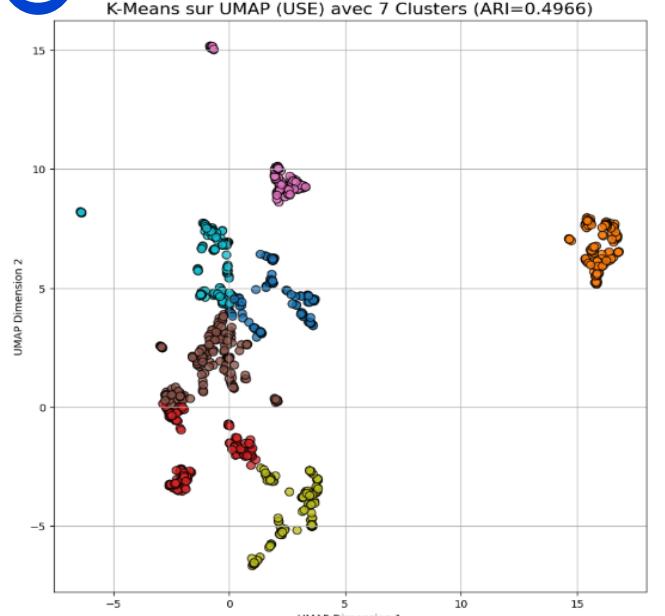
Clustering [AUTOENCODEUR]

K-Means (USE) avec 7 Clusters (ARI=0.3665)



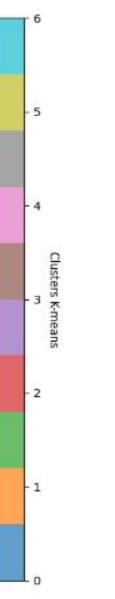
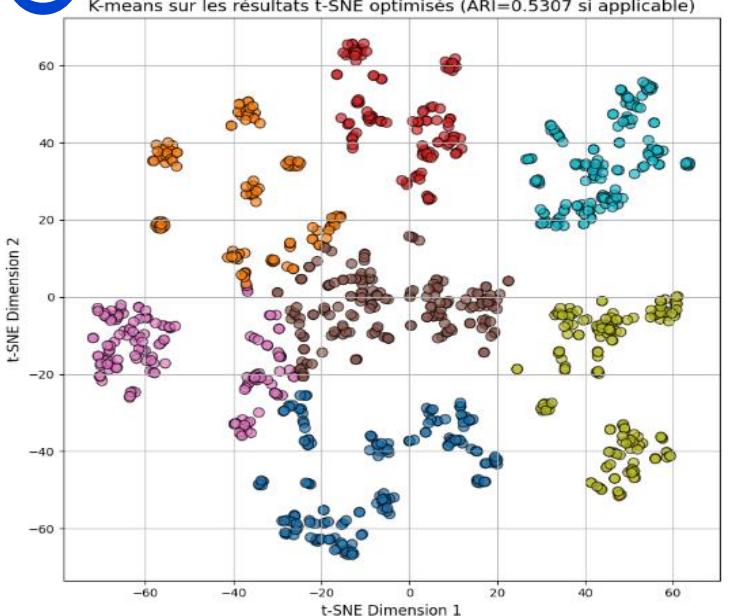
Clustering [UMAP]

K-Means sur UMAP (USE) avec 7 Clusters (ARI=0.4966)



Clustering [t-SNE]

K-means sur les résultats t-SNE optimisés (ARI=0.5307 si applicable)



Pipeline faisabilité de classification de l'image par catégorie



[Collecte des images] → [Prétraitement des images] → [Extraction des Features]

1. Identification
2. Chargement

1. Conversion aux niveaux de gris
2. Filtrage du bruit avec filtre gaussien
3. Augmentation du contraste

Bibliothèque OpenCV

Open Source Computer Vision Library

1. **SIFT** : Extraction de descripteurs (128 dimensions)
2. **ORB**: Points clés détectés et décrits
3. **CNN**: réseau de neurones artificiels

[Réduction de dimension 2d] → [Clustering K-means] → [Matrice de Confusion]

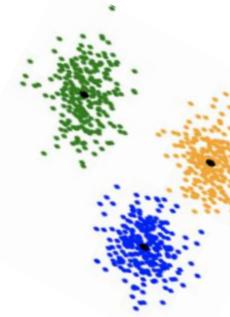
1. ACP : Analyse en Composante principale #2

1. T-SNE : t-distributed Stochastic Neighbor Embedding)

1. UMAP : (Uniform Manifold Approximation and Projection)

+

ARI



[Matrice de Confusion]

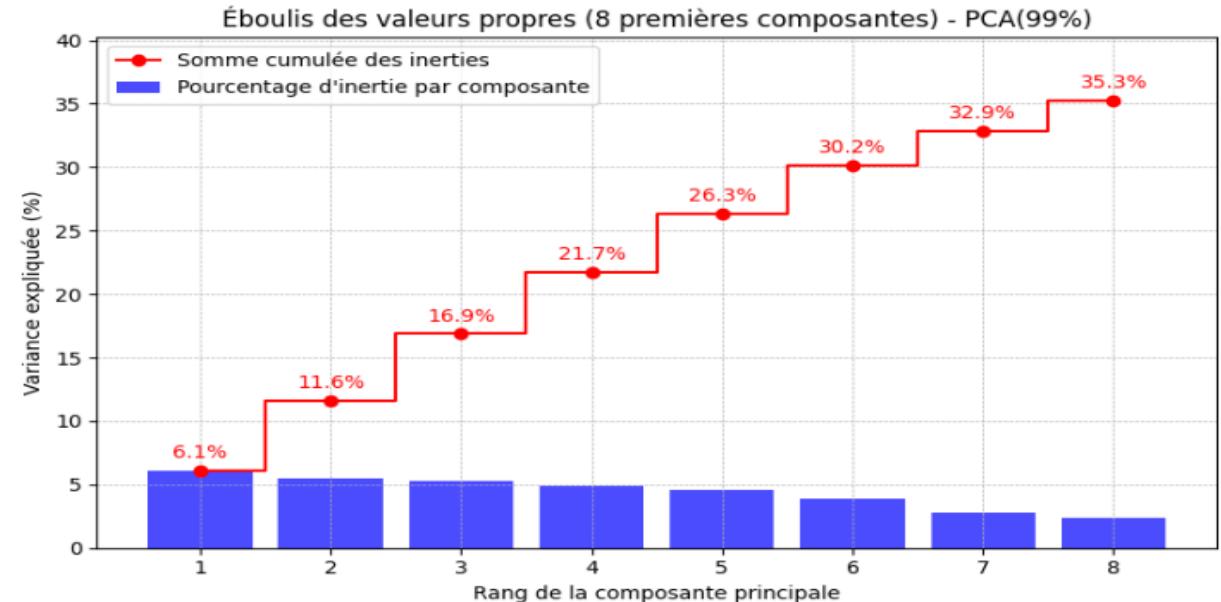
Catégories Réelles		Clusters Prédits Clustering K-Means						
		Home Furnishing	Baby Care	Watches	Home Decor & Festive Needs	Kitchen & Dining	Beauty and Personal Care	Computers
Home Furnishing	93	52	0	2	1	0	2	
Baby Care	0	61	78	8	1	1	1	
Watches	0	0	33	46	27	2	42	
Home Decor & Festive Needs	0	0	0	137	3	0	10	
Kitchen & Dining	0	0	0	28	95	26	1	
Beauty and Personal Care	0	0	0	3	2	125	20	
Computers	8	11	5	1	13	14	98	

Ce pipeline permet de classifier les images par catégories grâce à une combinaison d'extraction de features, réduction de dimension et de clustering

Projection (Réd.2D's) – Méthodes Basiques – SIFT

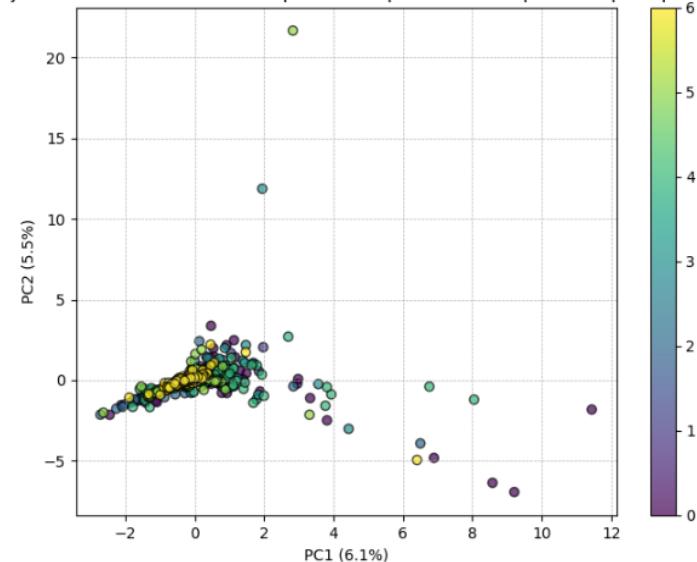


Projection (Réd.2D [ACP])



Dimensions dataset avant réduction PCA : (1050, 719)
Dimensions dataset après réduction PCA : (1050, 502)

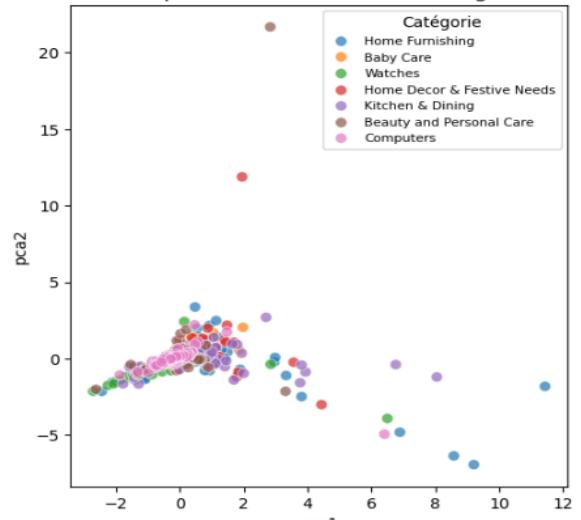
Projection des individus dans l'espace des 2 premières composantes principales



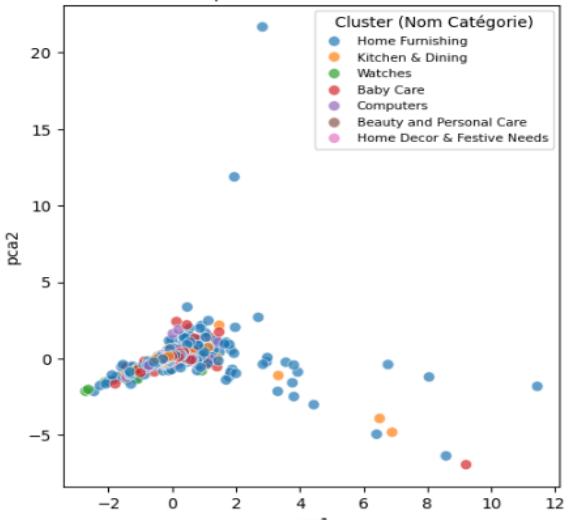
Clustering [ACP]

ARI : 0.0034

PCA-SIFT - Représentation selon les catégories réelles



PCA-SIFT - Représentation selon les clusters



Matrice de Confusion

Segmentation & Clustering | Algorithme: SIFT + Réduction PCA 2D

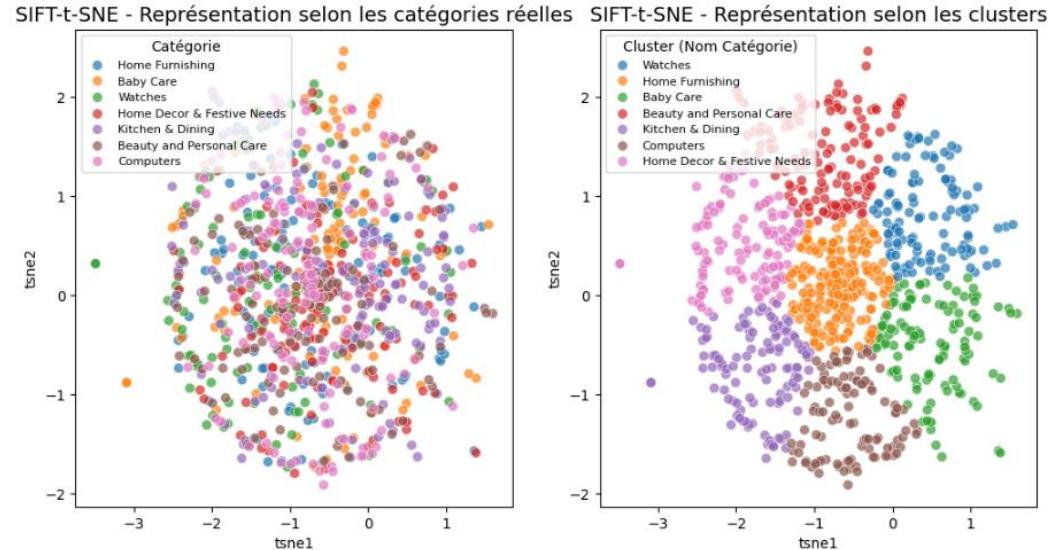
Catégories Réelles	Clusters Prédits Clustering K-Means						
	Home Furnishing	Baby Care	Watches	Home Decor & Festive Needs	Kitchen & Dining	Beauty and Personal Care	Computers
Home Furnishing	10	0	0	12	89	11	28
Baby Care	5	2	0	14	96	6	27
Watches	4	0	0	16	93	13	24
Home Decor & Festive Needs	1	0	0	18	94	13	24
Kitchen & Dining	2	0	0	8	110	6	24
Beauty and Personal Care	7	0	0	12	100	15	16
Computers	2	0	2	9	76	16	45

Projection (Réd.2D's) – Méthodes Basiques – SIFT



Clustering [t-SNE]

ARI : 0.0165



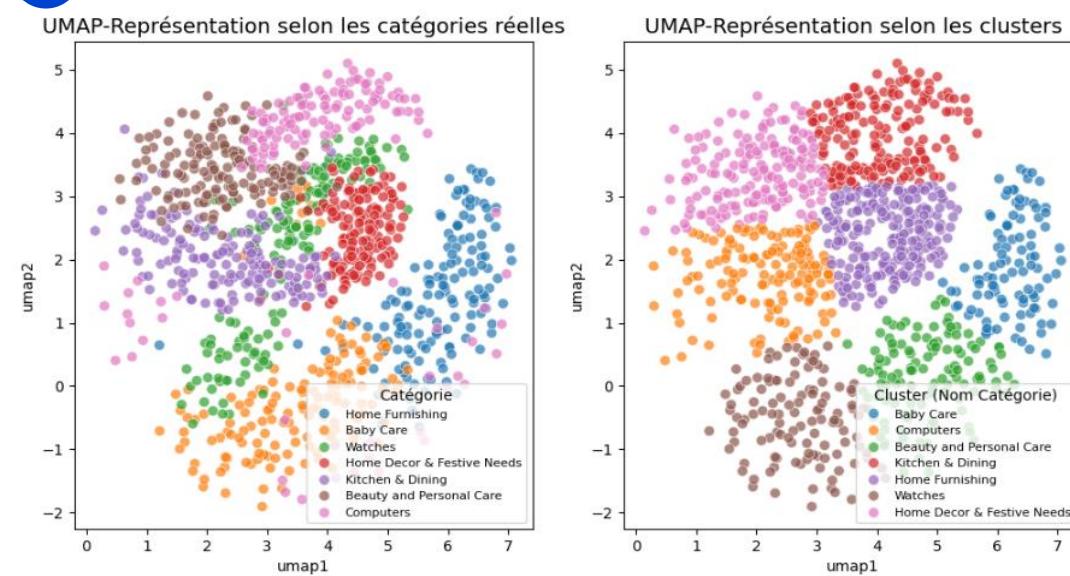
Matrice de confusion

Segmentation & Clustering | Algorithme: SIFT + Réduction t-SNE 2D

Catégories Réelles	Clusters Prédits Clustering K-Means						
	Home Furnishing	Baby Care	Watches	Home Decor & Festive Needs	Kitchen & Dining	Beauty and Personal Care	Computers
Home Furnishing	22	19	18	16	21	37	17
Baby Care	19	38	12	12	12	44	13
Watches	11	16	30	22	8	51	12
Home Decor & Festive Needs	18	10	13	16	27	54	12
Kitchen & Dining	28	19	17	12	34	27	13
Beauty and Personal Care	7	9	18	20	13	67	16
Computers	17	31	21	14	7	34	26

Clustering [UMAP]

ARI : 0.4179



Matrice de confusion

Segmentation & Clustering | Algorithme: SIFT + Réduction UMAP 2D

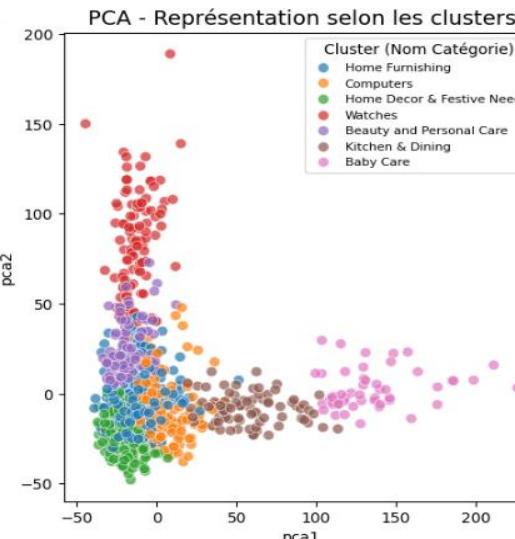
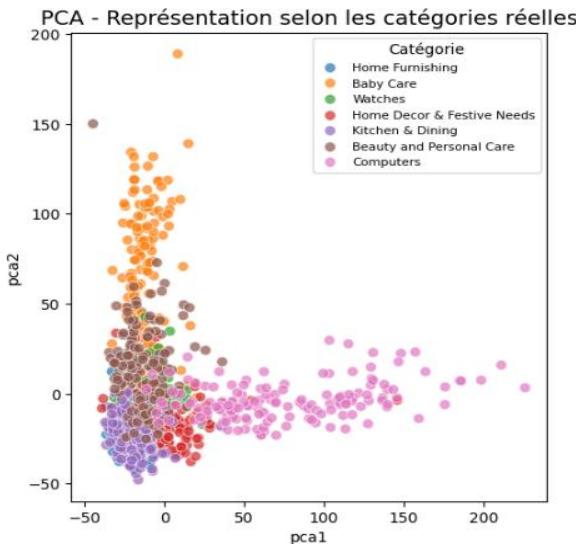
Catégories Réelles	Clusters Prédits Clustering K-Means						
	Home Furnishing	Baby Care	Watches	Home Decor & Festive Needs	Kitchen & Dining	Beauty and Personal Care	Computers
Home Furnishing	93	52	0	2	1	0	2
Baby Care	0	61	78	8	1	1	1
Watches	0	0	33	46	27	2	42
Home Decor & Festive Needs	0	0	0	137	3	0	10
Kitchen & Dining	0	0	0	28	95	26	1
Beauty and Personal Care	0	0	0	3	2	125	20
Computers	8	11	5	1	13	14	98

Projection (Réd.2D's) – Méthode Avancée – CNN



Clustering [ACP]

ARI : 0.3552



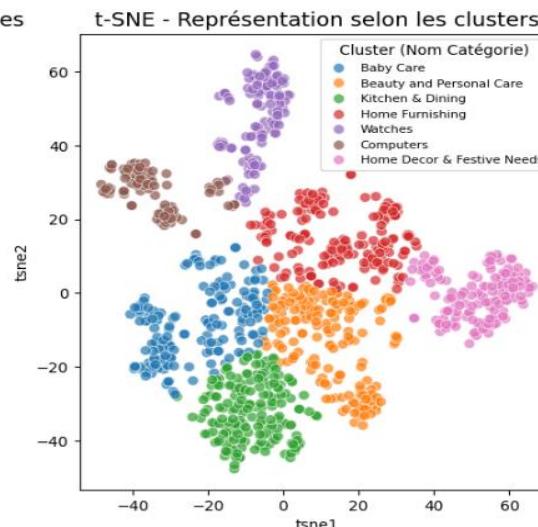
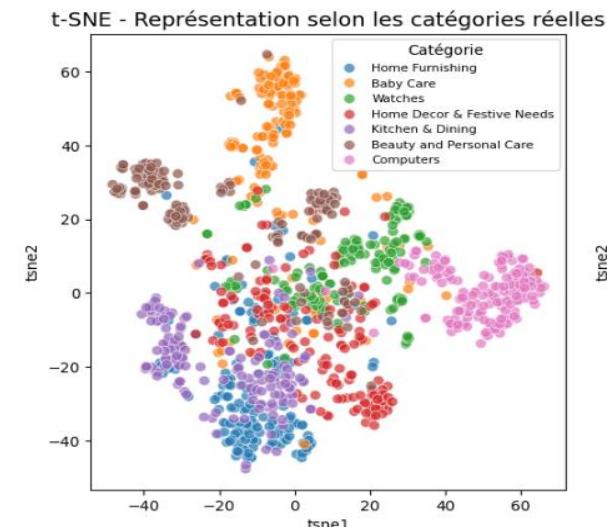
Matrice de confusion

Segmentation & Clustering | Algorithme: CNN + Réduction PCA 2D

Catégories Réelles	Clusters Prédits Clustering K-Means						
	Home Furnishing	Baby Care	Watches	Home Decor & Festive Needs	Kitchen & Dining	Beauty and Personal Care	Computers
Home Furnishing	0	2	29	8	110	1	0
Baby Care	0	98	40	5	6	0	1
Watches	0	0	138	7	4	0	1
Home Decor & Festive Needs	1	0	49	93	6	0	1
Kitchen & Dining	0	0	26	4	120	0	0
Beauty and Personal Care	0	2	29	36	0	83	0
Computers	50	0	17	4	0	0	79

Clustering [t-SNE]

ARI : 0.4133



Matrice de confusion

Segmentation & Clustering | Algorithme: CNN + Réduction t-SNE 2D

Catégories Réelles	Clusters Prédits Clustering K-Means						
	Home Furnishing	Baby Care	Watches	Home Decor & Festive Needs	Kitchen & Dining	Beauty and Personal Care	Computers
Home Furnishing	107	2	5	9	26	1	0
Baby Care	1	109	16	11	10	2	1
Watches	2	2	83	47	9	6	1
Home Decor & Festive Needs	7	1	13	82	46	0	1
Kitchen & Dining	65	0	2	10	73	0	0
Beauty and Personal Care	0	2	43	19	4	82	0
Computers	0	0	8	2	1	0	139

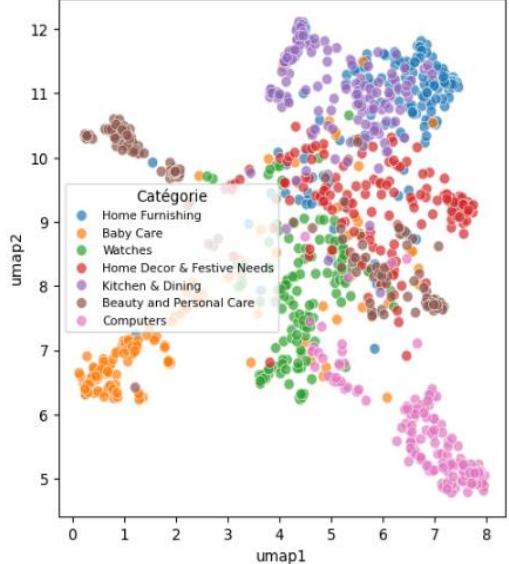
Projection (Réd.2D's) – Méthode Avancée – CNN



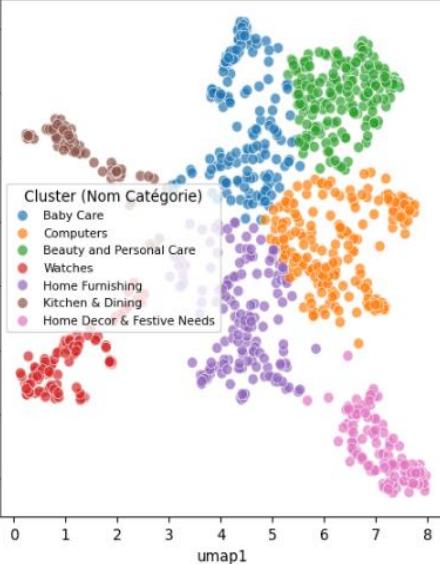
Clustering [UMAP]

ARI : 0.4384

UMAP-Représentation selon les catégories réelles



UMAP-Représentation selon les clusters

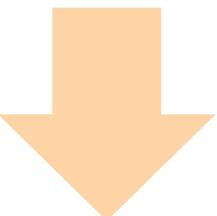


Matrice de confusion

Segmentation & Clustering | Algorithme: CNN + Réduction UMAP 2D

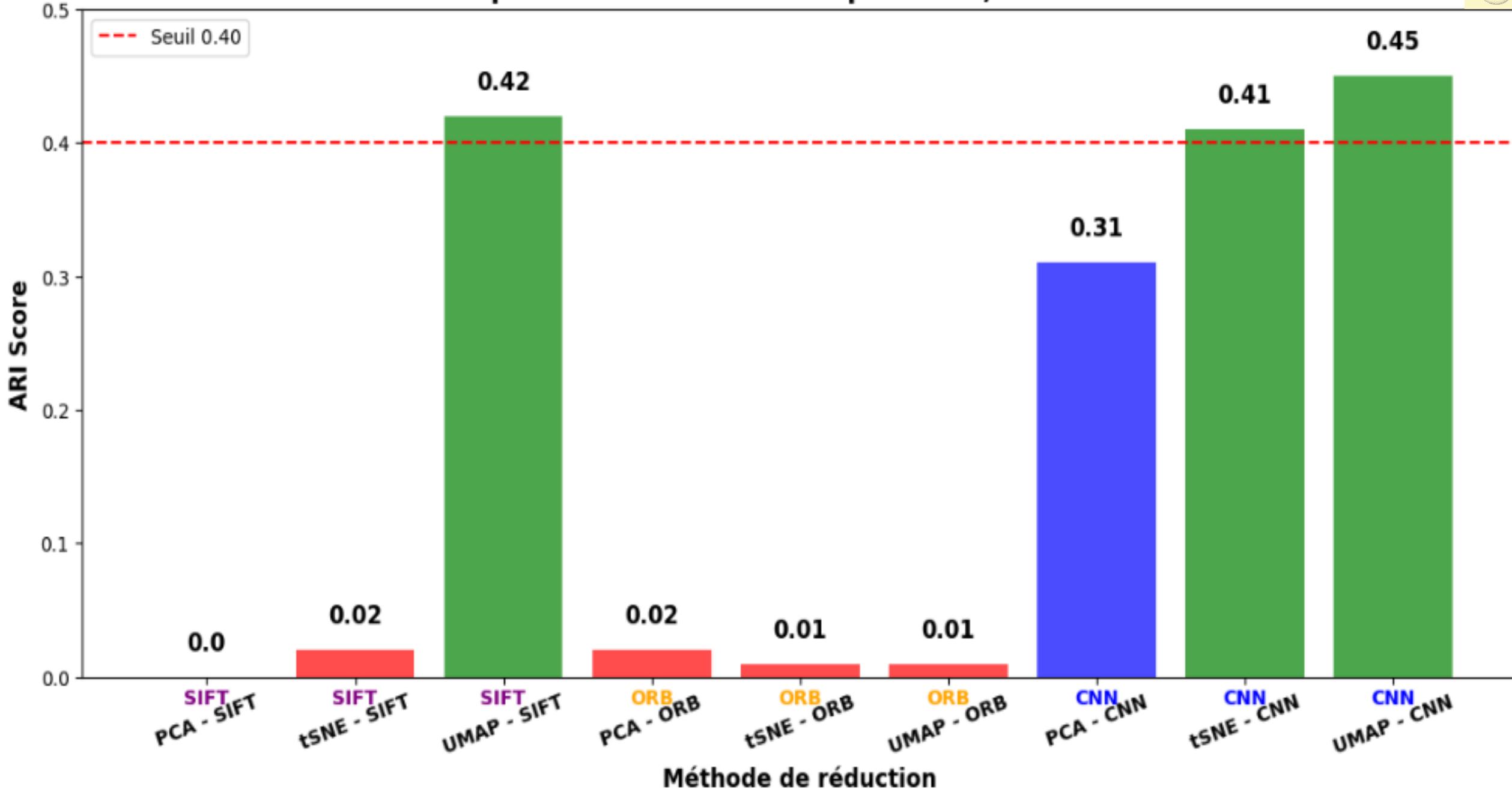
Catégories Réelles	Clusters Prédits Clustering K-Means						
	Home Furnishing	Baby Care	Watches	Home Decor & Festive Needs	Kitchen & Dining	Beauty and Personal Care	Computers
Home Furnishing	108	2	4	10	25	1	0
Baby Care	3	108	15	14	8	1	1
Watches	0	0	117	20	11	2	0
Home Decor & Festive Needs	11	0	6	91	39	1	2
Kitchen & Dining	67	0	2	2	79	0	0
Beauty and Personal Care	1	1	3	61	2	82	0
Computers	0	0	24	4	0	0	122

RESULTAT



Comparaison des scores ARI pour SIFT, ORB et CNN

place de monté



Répartition du data set CNN Train / Validation / Test



Dataset	Contenu	Description
x_train	Images d'entraînement	Utilisé pour entraîner le modèle (75%)
y_train	Labels d'entraînement	Labels associés aux images d'entraînement
x_val	Images de validation	Utilisé pour évaluer le modèle (15%)
y_val	Labels de validation	Labels associés aux images de validation
x_test	Images de test	Utilisé pour tester le modèle (10%) avec les meilleurs paramètres
y_test	Labels de test	Labels associés aux images pour effectuer le test

Répartition homogène des catégories (TRAIN) - TRAIN contient 787 images :

Catégorie	label_name	Nombre d'images	Nombre total de labels
1	Beauty and Personal Care	113	113
5	Kitchen & Dining	113	113
0	Baby Care	113	113
4	Home Furnishing	112	112
6	Watches	112	112
2	Computers	112	112
3	Home Decor & Festive Needs	112	112

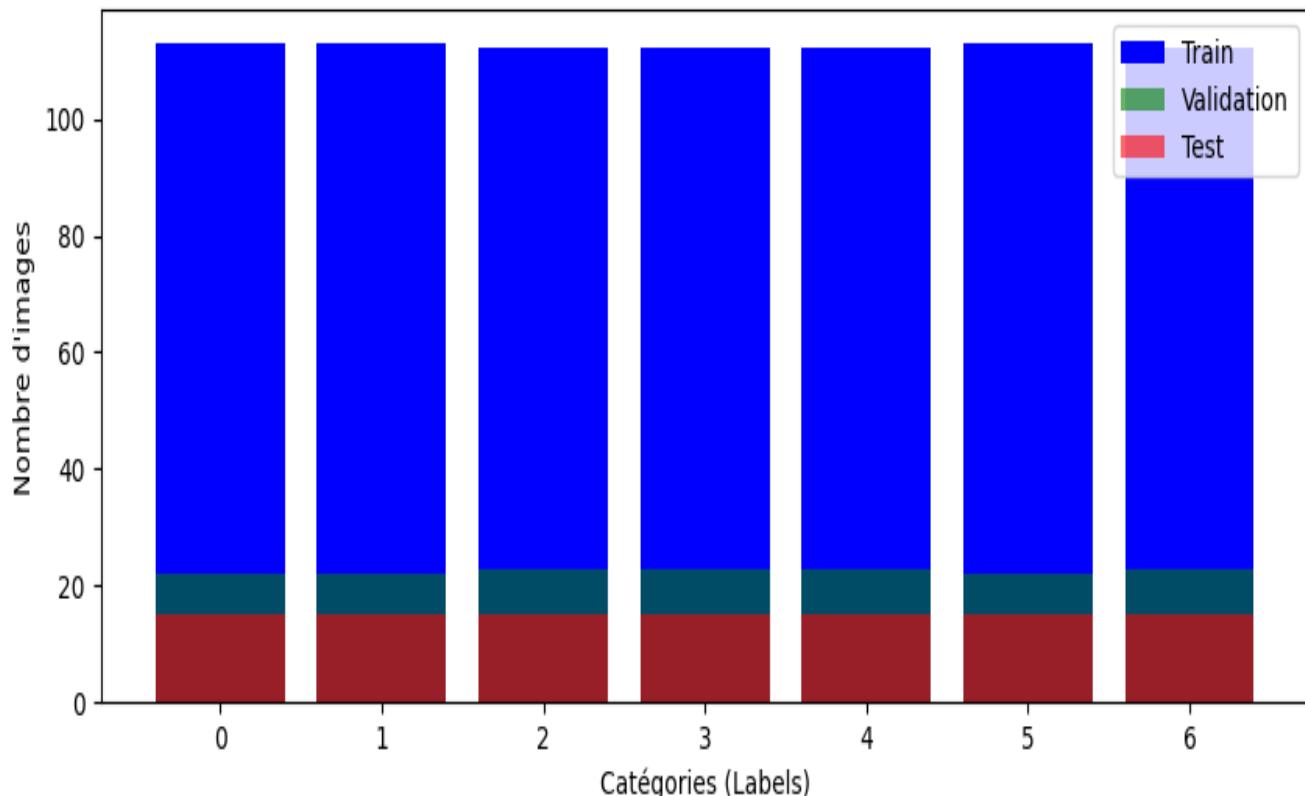
Répartition homogène des catégories (VAL) - VAL contient 158 images :

Catégorie	label_name	Nombre d'images	Nombre total de labels
3	Home Decor & Festive Needs	23	23
6	Watches	23	23
2	Computers	23	23
4	Home Furnishing	23	23
5	Kitchen & Dining	22	22
1	Beauty and Personal Care	22	22
0	Baby Care	22	22

Répartition homogène des catégories (TEST) - TEST contient 105 images :

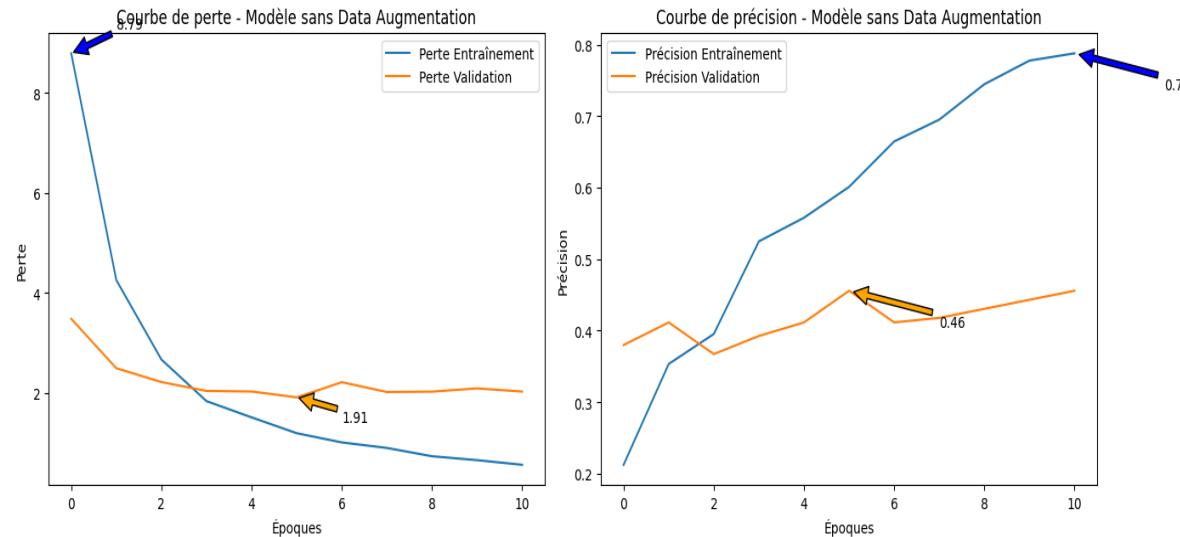
Catégorie	label_name	Nombre d'images	Nombre total de labels
1	Beauty and Personal Care	15	15
0	Baby Care	15	15
6	Watches	15	15
5	Kitchen & Dining	15	15
3	Home Decor & Festive Needs	15	15
4	Home Furnishing	15	15
2	Computers	15	15

Distribution des labels dans Train 75% / Validation 15% / Test 10%



CNN sans data augmentation

Entraînement sur **train** 81% – Evaluation sur la **validation** 45%



ENTRAÎNEMENT : **Précision 81%** | **Validation 45%**

La courbe de perte :

- **Entraînement** : baisse progressive, bon apprentissage.
- **Validation** : plus linéaire, un surajustement est visible.*

La courbe de précision :

- **Entraînement**: augmentation progressive pour l'entraînement et la validation.
- Corrélation visible entre l'entraînement et de validation epoch (0-2).

Entraînement **train** avec le meilleur modèle – Evaluation sur le **test** 40%



TEST : **Précision 40%**

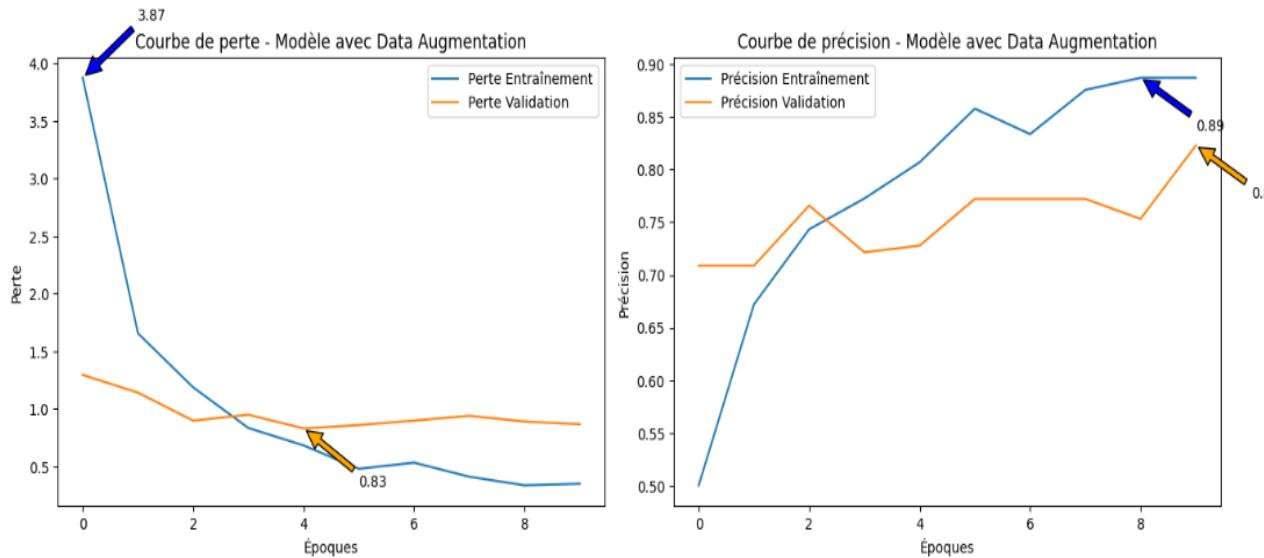
Entraînement **train** avec le meilleur modèle sur le test :

- Précision limitée en **test**
- Précision cohérente entre **train** et **validation** mais un écart est observé
- Bon apprentissage car la perte finale est estimée à 1.95% (bonne généralisation)
- Précision estimée à 40% ce qui est faible

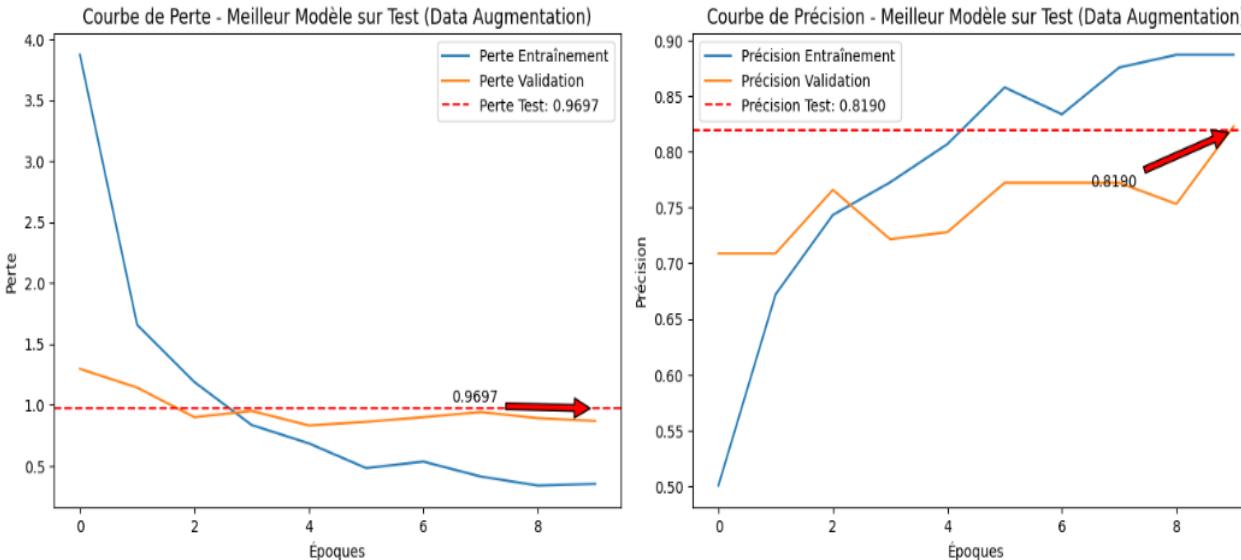
*Le modèle montre un **surajustement** et la précision observée est faible sur le **test** (40%)

CNN avec data augmentation

Entraînement sur **train** 96% – Evaluation sur la **validation** 77%



Entraînement **train** avec le **meilleur modèle** – Evaluation sur le **test** 81%



Entraînement : **Précision 96% | Validation 82%**
La courbe de perte :

- Entraînement : Diminution rapide (valeur minimale epoch 8 = 0.83%)
- Validation : Perte stable après l'epoch 8 (meilleure généralisation par rapport au modèle sans data augmentation).

La courbe de précision :

- Entraînement: La précision augmente progressivement (96 % à l'époque finale)
- Validation : Précision atteint à 82% (amélioration significative par rapport au modèle sans data augmentation).

TEST : **Précision 81% vs 40% sans data augmentée**

Entraînement **train** avec le **meilleur modèle sur le test** :

- Perte : 0.96% (inférieure à 1.95 sans data augmentation) - Stable
- Précision : 0.81% - Supérieure au modèle sans data augmentation.

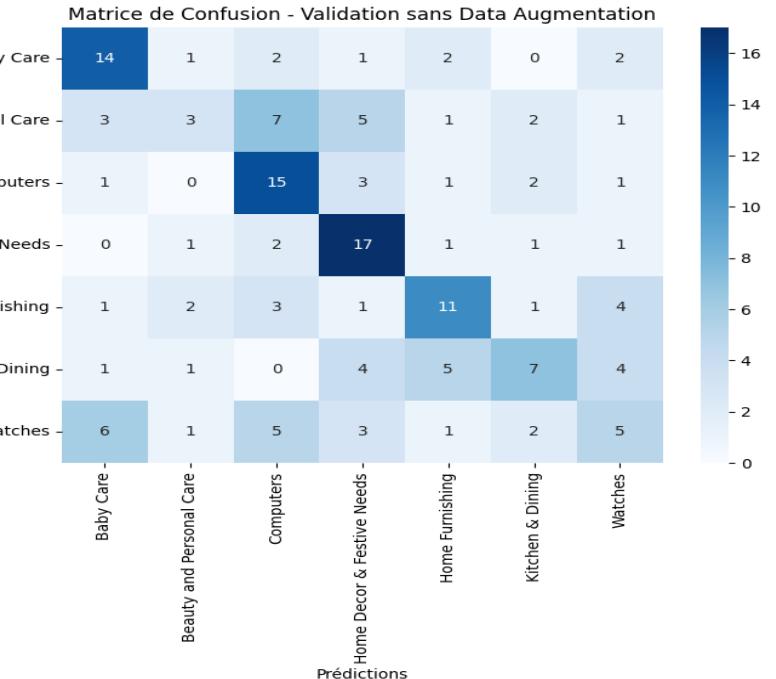
La data augmentation améliore la généralisation du modèle :

- Précision test augmentée **40% → 81%** grâce à la data augmentation.

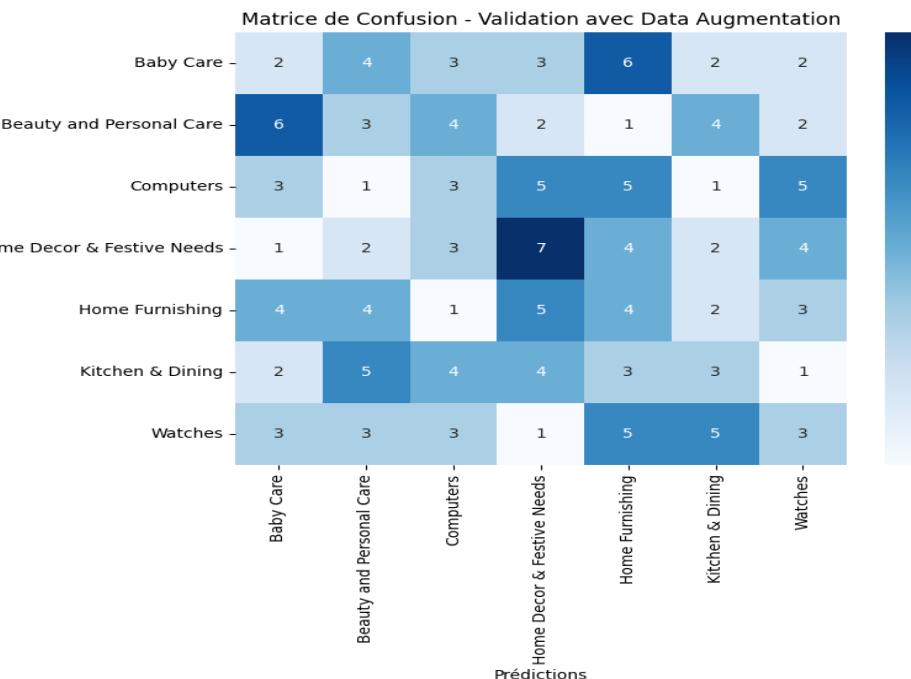
Les matrices de confusions



Vrais labels



Vrais labels



- Moins de dispersion mais la généralisation des classes est limitée

Rapport de classification - Validation :

	precision	recall	f1-score	support
Baby Care	0.54	0.64	0.58	22
Beauty and Personal Care	0.33	0.14	0.19	22
Computers	0.44	0.65	0.53	23
Home Decor & Festive Needs	0.50	0.74	0.60	23
Home Furnishing	0.50	0.48	0.49	23
Kitchen & Dining	0.47	0.32	0.38	22
Watches	0.28	0.22	0.24	23
accuracy			0.46	158
macro avg	0.44	0.45	0.43	158
weighted avg	0.44	0.46	0.43	158

- Prédiction plus variée, meilleures généralisations attendues
- La data augmentation augmente la généralisation mais les données variées sont aussi plus sont plus difficiles à mémoriser)

Rapport de classification - Validation :

	precision	recall	f1-score	support
Baby Care	0.10	0.09	0.09	22
Beauty and Personal Care	0.14	0.14	0.14	22
Computers	0.14	0.13	0.14	23
Home Decor & Festive Needs	0.26	0.30	0.28	23
Home Furnishing	0.14	0.17	0.16	23
Kitchen & Dining	0.16	0.14	0.15	22
Watches	0.15	0.13	0.14	23
accuracy				158
macro avg	0.15	0.16	0.16	158
weighted avg	0.16	0.16	0.16	158

Élargir la gamme de produit de l'épicerie fine

1

[Définition et requête de l' API]

2

[Verification de l'extraction]

3

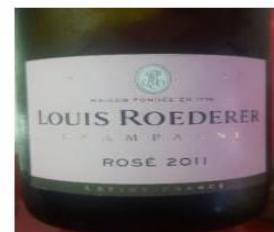
[Visualisation de l'extraction]

1. **URL**
2. **Champs** d'extraction
1. **Paramétrage** des filtres sur les champs d'extraction produit (#10)

1. **Produits** provenant de l'épicerie fine « **category** »
2. **Produit** contenant « Champagne » dans les ingrédients: («**foodContentsLabel**»)

1. **Fonction** : Récupération des url de la requête API
2. **Visualisation** : Produits de l'épicerie fine à base de Champagne.

Ci-dessous les produits extraits à base de Champagne:



Conclusion



1

Classification par le **texte** : faisabilité confirmée (seuil > 0.40)



2

Classification par **l'image** : faisabilité confirmée (seuil > 0.40)



■ L'extraction avancée des caractéristiques **textuelles** et **visuelles** (TFIDF, Word2Vec, USE, CNN) permet la création d'un moteur de classification automatique.

■ La marge d'erreur du moteur de classification automatique peut être réduite en choisissant la technique de réduction de dimension optimale tout en optimisant ses hyperparamètres UMAP dans cette étude.

3

La mise en place de l'algorithme de classification supervisé CNN, le réseau de neurone convolutif nous permet:

- ✓ D'utiliser un modèle avec l'attribution automatique des catégories via CNN. *Approche CNN et VGG16 Visual Geometry Group.*
- ✓ De proposer une classification automatique des produits mis en ligne par les vendeurs.

4

L'impact positif de la mise en place d'un moteur de classification automatique est tout autant bénéfique pour **l'acheteur** que pour le **vendeur** car nous obtenons :

- Une amélioration de l'expérience utilisateur des **vendeurs**
- Une recherche simplifiée pour les **acheteurs**

Le moteur de classification améliore la généralisation et facilite la gestion des produits du site « place de marché ».



Merci

QUESTION(S) – RÉPONSE(S)