CLASSIFICATION AUTOMATIQUE DE PRODUIT

Antonine LAROZE-CERVETTI

OBJECTIFS

- Effectuer une étude de faisabilité pour une classification automatique à partir des données textuelles et visuelles
- Classification supervisée des données visuelles
- Requête sur l'API d'OpenFoodFacts pour des produits contenant du champagne

PRÉPARATION DES DONNÉES

Base de données

- 1050 produits
- Variables intéressantes pour notre étude:
« description », « product_name », « image»
et « product_category_tree »

Valeurs manquantes

- 0 pour les variables quantitatives
- « Not Available » pour les variables qualitatives

Variable catégorie

La variable *product_category_tree* représente la catégorie de chaque produit (7 niveaux) et est décomposée en 7 sous-variables. Seul le premier niveau (*category_1*) est conservé pour l'analyse.

```
Colonne: category_1
Nombre de valeurs uniques: 7

category_1
home furnishing              150
baby care                    150
watches                      150
home decor & festive needs   150
kitchen & dining             150
beauty and personal care     150
computers                    150
Name: count, dtype: int64
```

ANALYSE DES DONNÉES TEXTUELLES

Variables
« product_name », texte court
« description », texte plus long pouvant parfois porter à confusion

NETTOYAGE INITIAL

1. Suppression des caractères spéciaux, des chiffres et des espaces en trop.
2. Conversion de tous les mots en minuscules
3. Correction des contractions s'il y en a

```
# Test des fonctions
print(remove_special("Ceci est le 1er test !"))
print(convert_to_lower("C'est LE 2ème TEST !"))
print(normalize_spaces("C'est le    3ème   TEST  !"))
print(expand_contractions("Y'all can't expand contractions I'd think."))
```

```
Ceci est le er test
c'est le 2ème test !
C'est le 3ème TEST !
You all cannot expand contractions I would think.
```

NETTOYAGE AVANCÉE

(BIBLIOTHÈQUE NLTK)

1. Tokenisation : processus de découpage d'un texte en unités élémentaires (mots, phrases etc.)
2. Suppression des stopwords: mots courants (ex. : "le", "et", "de") généralement exclus de l'analyse car peu informatifs.
3. **Stemmisation** : réduction des mots à leur racine, souvent sans garantie de sens (ex. : *manger*, *mangeons* → *mang*).
4. **Lemmatisation** : réduction des mots à leur forme canonique (lemme) en tenant compte du contexte grammatical (ex. : *mangeons*, *mangé* → *manger*).

FEATURES EXTRACTION

Méthode	Définition	Avantages	Inconvénients
BoW	Compte la fréquence de chaque mot dans un document (sac de mots sans ordre)	Simple et rapide	Ignore le contexte, l'ordre des mots et la sémantique.
Tf-Idf	Pondère les mots selon leur fréquence et leur rareté. TF = fréquence ; IDF = rareté	Réduit l'impact des mots trop fréquents, plus informatif que BoW.	Toujours basé sur la fréquence, ne capture pas le contexte.
Word2Vec	Apprend des vecteurs de mots selon leur contexte. (Réseaux de neurones CBOW = prédit le **mot cible** à partir du **contexte** ou Skip-Gram = prédit **le contexte** à partir du **mot cible**).	Capture le contexte et la sémantique	Nécessite beaucoup de données
Bert	Modèle de langage pré-entraîné prenant en entrée des séquences de mots en tenant compte du **contexte des mots voisins**.	Puissant et comprend le contexte	Modèle lourd coûteux en ressources
Use	Modèle pré-entrainé basé sur des réseaux de neurones ; Encode toute la phrase comme un tout (pas mot par mot comme BERT)	Très performant pour les comparaisons de phrases, rapide à intégrer.	Plus considéré comme une « boîte noire »

VISUALISATIONS - MÉTHODOLOGIE

💡 Si **les catégories se distinguent visuellement**, cela indique que les représentations capturent des différences suffisamment marquées et cohérentes, rendant **une classification automatique réalisable.**

🔍 **Méthodologie**:

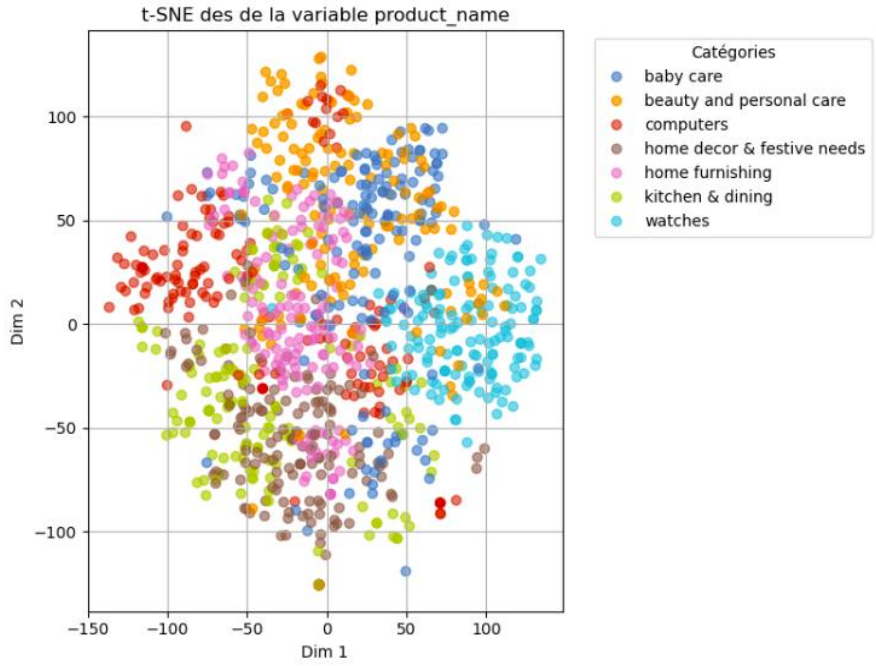
- Réduction de dimensionnalité via PCA puis t-SNE pour une visualisation 2D et une exploration de la répartition des observations,
- Application du clustering K-Means (7 clusters, correspondant aux catégories) et évaluation de la qualité de la séparation à l'aide des métriques ARI (Adjusted Rand Index) et NMI (Normalized Mutual Information,

⚠️ La réduction de dimension entraîne une perte d'information ; une projection 2D ne reflète pas toujours fidèlement la structure réelle des données.

⚠️ Un score de clustering faible peut survenir même si les catégories semblent bien séparées visuellement, car t-SNE peut créer des effets de séparation artificiels, et le clustering peut mal attribuer les labels.

VISUALISATIONS - GRAPHIQUES

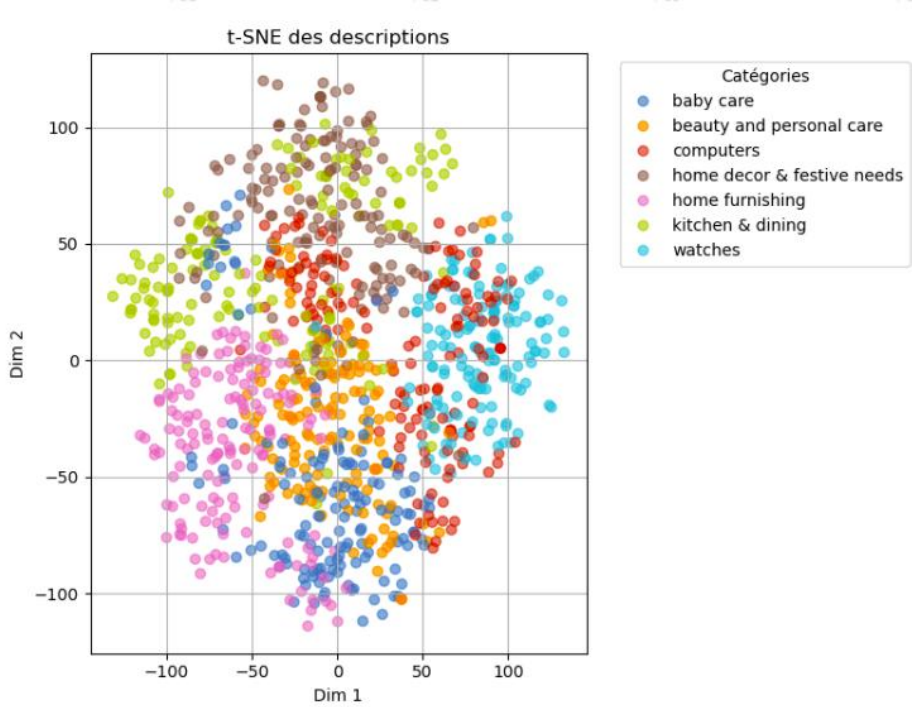
Variable product_name avec TF-IDF



ARI = 0,24 , NMI= 0,55

- Séparation modérée entre les catégories ; les représentations capturent des distinctions basiques mais exploitables.
- Adapté aux textes courts comme les noms produits, où les mots exacts comptent plus que le sens général.

Variable description avec TF-IDF

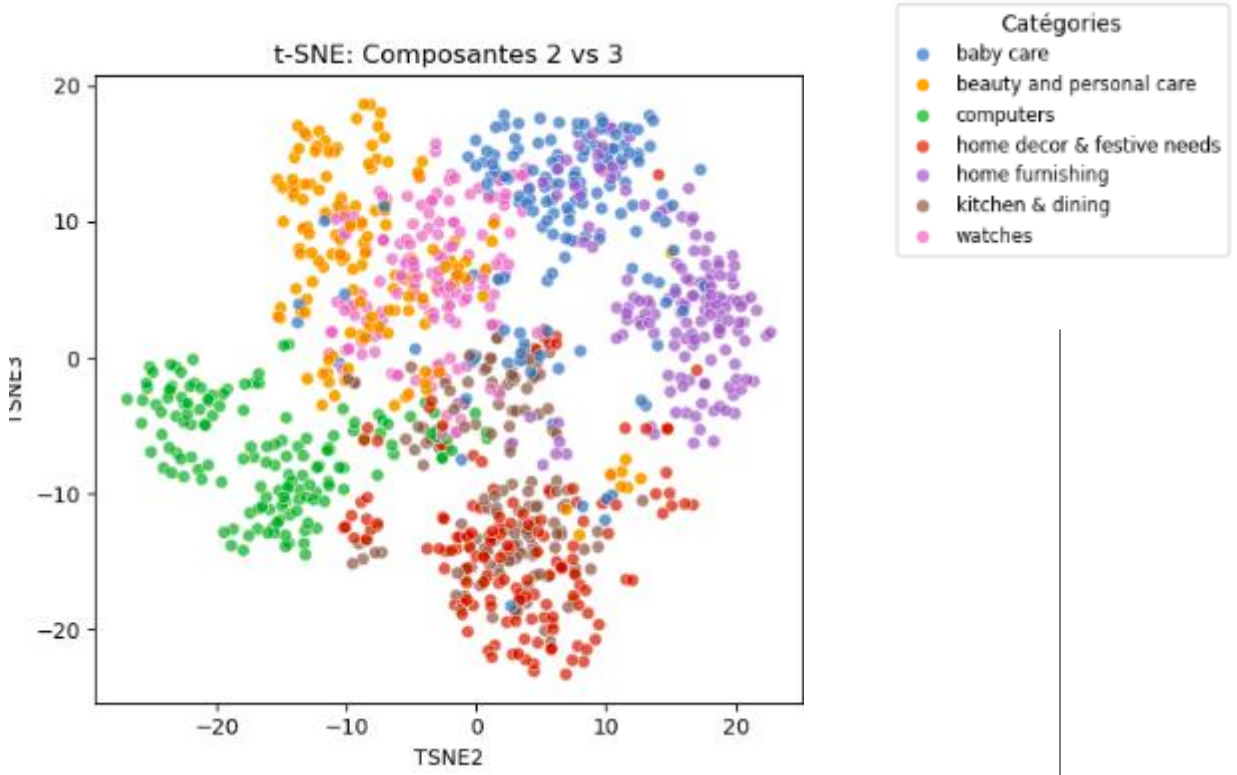


ARI = 0,26, NMI= 0,50

- Séparation modérée entre les catégories ; quantité d'informations modérées partagée entre clusters et vraies classes.
- Les descriptions étant plus longues, TF-IDF capture efficacement les termes distinctifs propres à chaque catégorie.

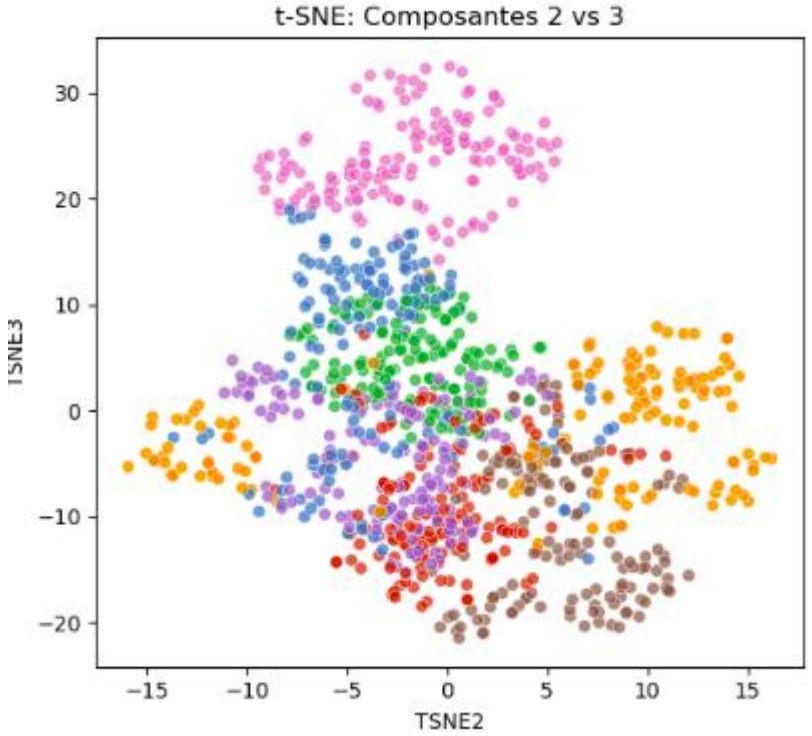
VISUALISATIONS - GRAPHIQUES

Variable product_name avec USE



- ARI = 0,51, les clusters se rapprochent plus des vraies catégories ; cela suggère que USE capture bien les différences entre les noms de produits.
- NMI= 0,68, une bonne part d'information catégorielle est préservée dans les clusters

Variable description avec USE



- ARI = 0,23, séparation modérée
- NMI= 0,40, la quantité d'informations partagée reste modérée
- Bien que USE encode le sens général des descriptions, il ne met pas assez en valeur les mots-clés discriminants.

MODÈLES DE CLASSIFICATION – RÉSULTATS

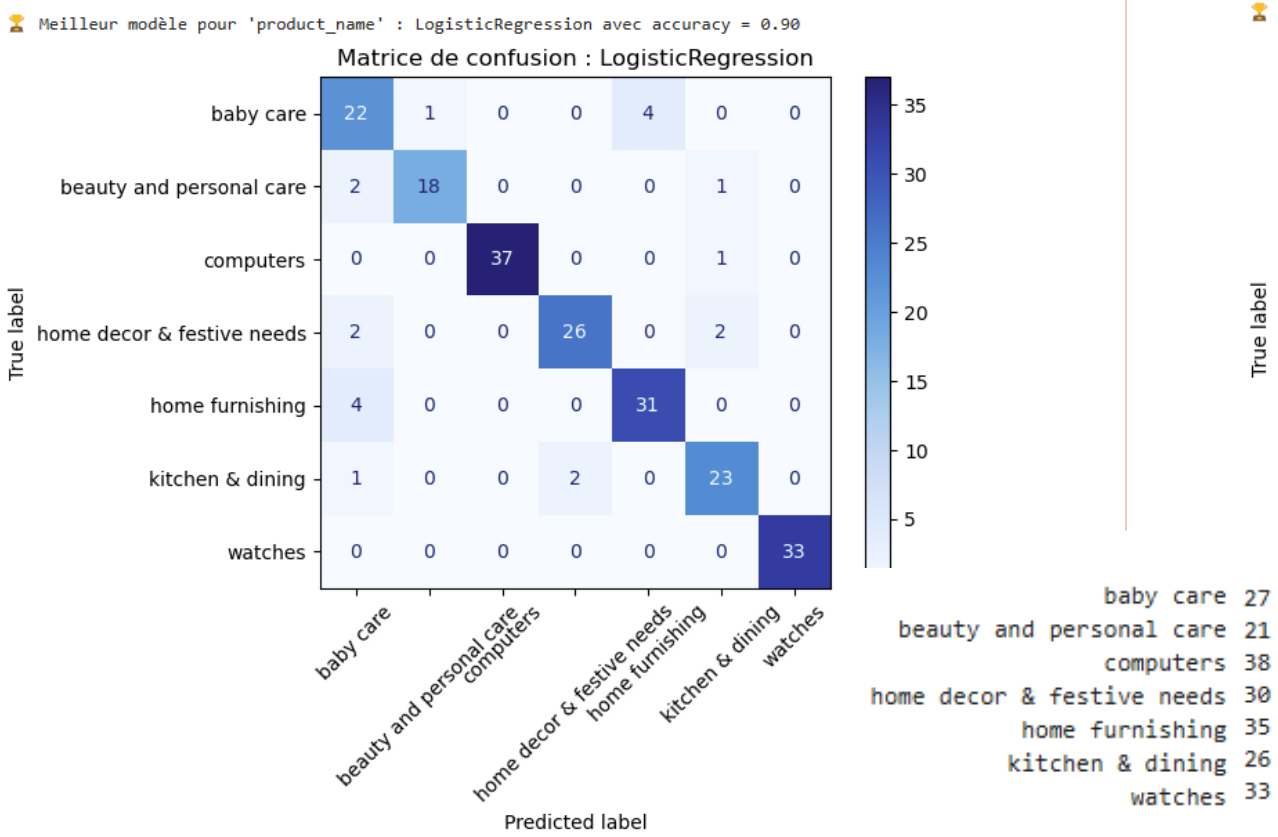
- **LogisticRegression:** modèle linéaire simple qui prédit la probabilité d'appartenance à une classe
- **Xgboost:** modèle de de boosting basé sur des arbres de décision
- **MLP:** un réseau de neurones artificiels avec plusieurs couches qui apprend des représentations non linéaires pour la classification

Variables	Modèles	Accuracy	F1-score	Timer
Product_name (USE)	Logistic Regres.	0,90	0,91	5
	Xgboost	0,90	0,90	137
	MLP	0,88	0,87	4
Description (BERT)	Logistic Regres.	0,89	0,89	5
	Xgboost	0,84	0,84	254
	MLP	0,90	0,90	8
Product_name (Tf-IDF)	Logistic Regres.	0,93	0,93	0,6
	Xgboost	0,88	0,88	5
	MLP	0,94	0,94	24
Description (Tf-IDF)	Logistic Regres.	0,96	0,96	1
	Xgboost	0,95	0,95	26
	MLP	0,95	0,95	50

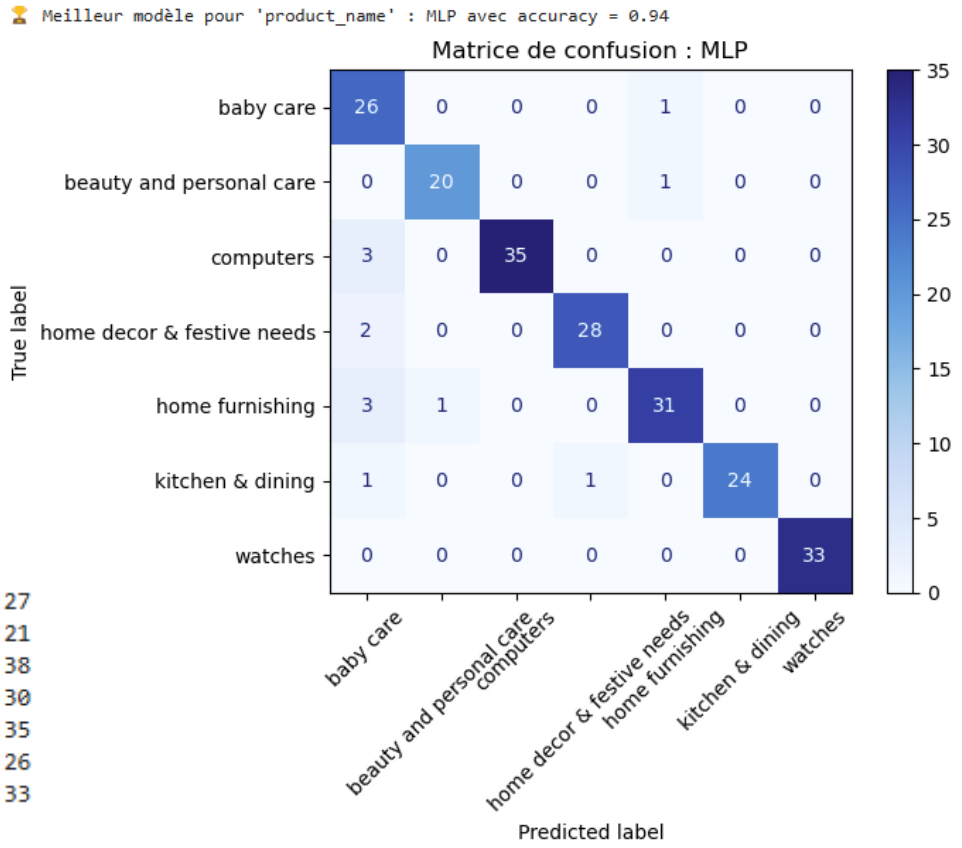
MODÈLES DE CLASSIFICATION-MATRICE DE CONFUSION

Quelques erreurs notables sur les classes baby care et home furnishing (quelques produits peuvent être similaires: chaise haute etc.). L'extraction **USE** permet une représentation sémantique mais peut être trop générale pour différencier certaines catégories proches.

Les erreurs sont moins prononcées pour les catégories baby care et home furnishing. L'approche TF-IDF + MLP semble mieux capter les mots discriminants, peut-être car ils sont courts et spécifiques.

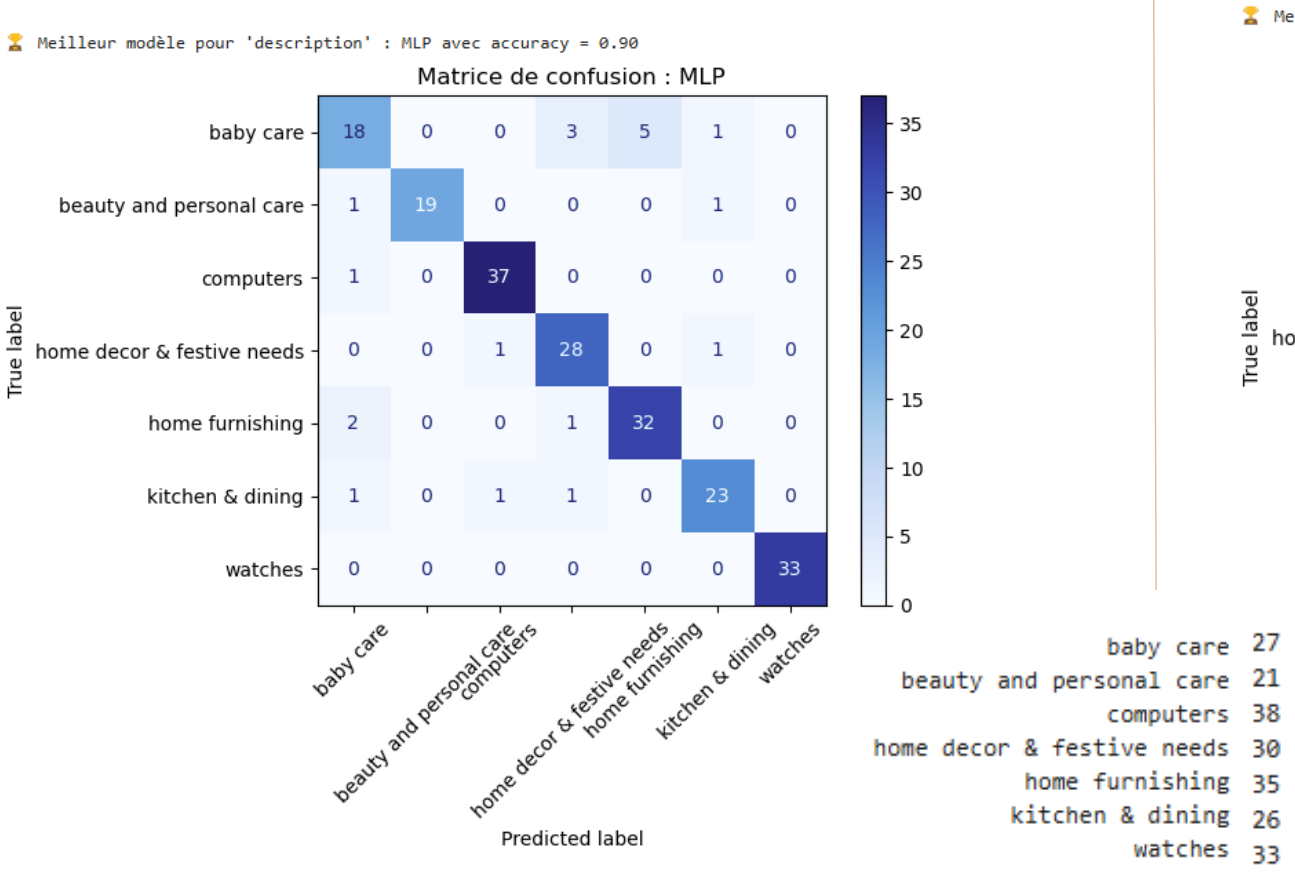



```
baby care                    27
beauty and personal care     21
computers                    38
home decor & festive needs   30
home furnishing              35
kitchen & dining             26
watches                      33
```



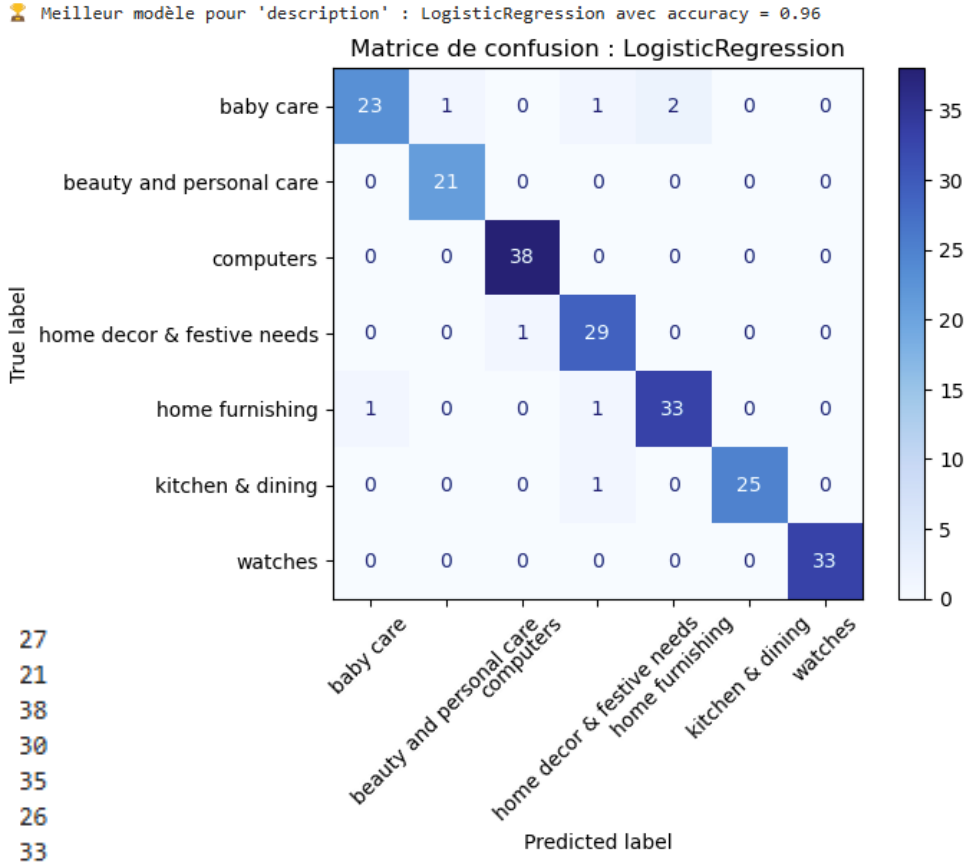
MODÈLES DE CLASSIFICATION-MATRICE DE CONFUSION

Erreurs encore plus marquées entre baby care et home furnishing car classes qui peuvent porter à confusion. Sinon l'approche Bert généralise plutôt bien et est puissante.




```
baby care                  27
beauty and personal care   21
computers                  38
home decor & festive needs 30
home furnishing            35
kitchen & dining           26
watches                    33
```

Les erreurs sont moins prononcées pour les catégories baby care et home furnishing. L'approche avec TF-IDF généralise mieux.



ANALYSE DES DONNÉES VISUELLES

1050 images

PRÉ-TRAITEMENT

Niveau de gris: simplifier l’image en supprimant les informations de couleur tout en conservant la structure de base

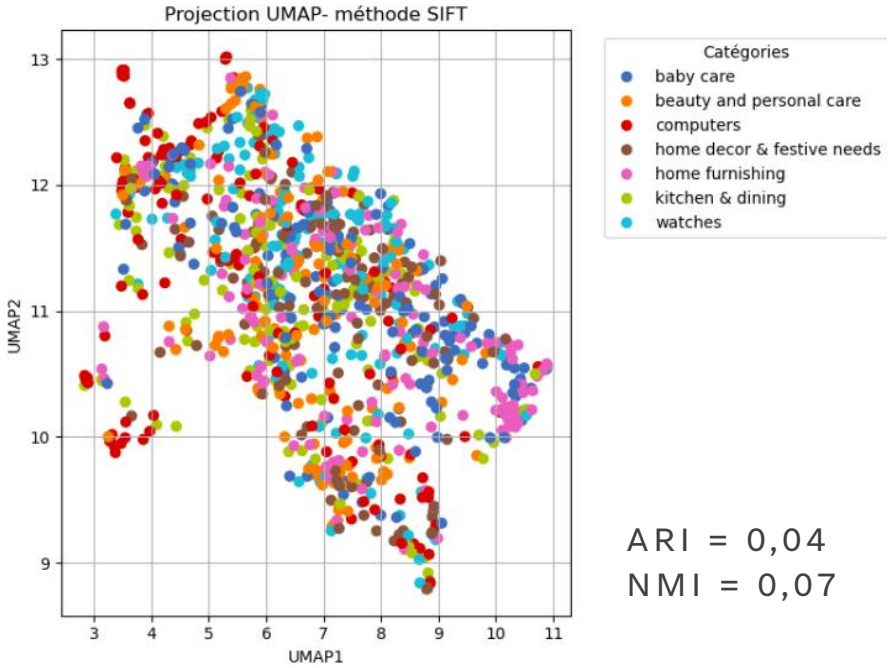
Filtre gaussien : réduire les petits bruits visuels, floute légèrement les détails trop fins

Amélioration du contraste: améliorer le contraste d'une image en redistribuant les niveaux de gris de manière uniforme

EXTRACTION DE FEATURES: SIFT

SIFT (scale-invariant feature transform) est un algorithme qui détecte et décrit des points d'intérêt (keypoints) d'une image de manière robuste aux changements d'échelle, de rotation et partiellement à l'illumination. La même méthode utilisée avec les données textuelles est appliquée

Pour cette image, SIFT a détecté 1259 keypoints



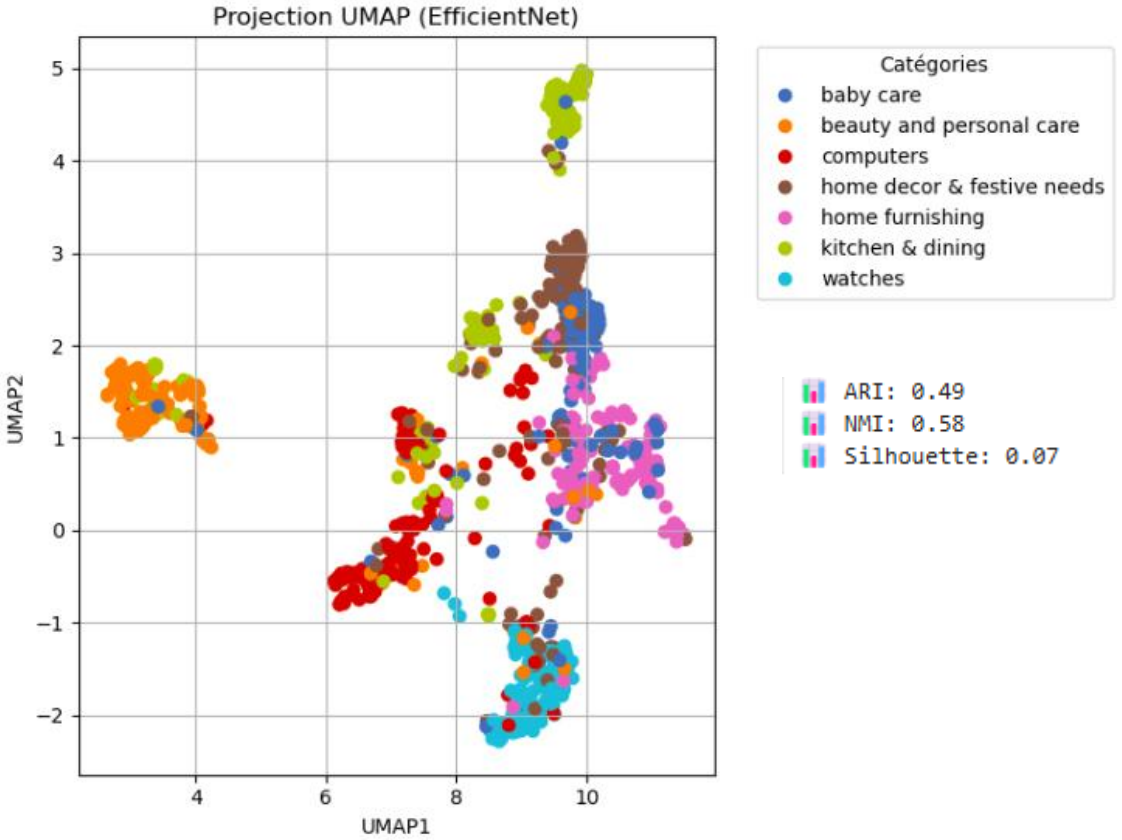
ARI = 0,04
NMI = 0,07

Le graphique, tout comme les métriques indiquent une séparation très faible entre les catégories. Cela suggère que la représentation actuelle ne parvient pas à capturer des différences pertinentes entre les classes

EXTRACTION DE FEATURES: TRANSFERT LEARNING

Test de VGG16, ResNet EfficientNet qui sont des architectures de réseaux de neurones convolutifs (CNN) avancées, conçues pour l'analyse d'images, Chacune proposant une approche différente pour améliorer la performance : simplicité pour VGG16, profondeur pour ResNet, et optimisation de la rapidité et de la performance pour EfficientNet.
EfficientNet sort comme le modèle le plus optimal.

Le graphique ainsi que les métriques montrent une bonne séparation entre les catégories.



MODÉLISATION DE CLASSIFICATION AUTOMATIQUE- MÉTHODE

Approche 1: Modèle CNN EffficientNet sans data augmentation

Approche 2: Modèle CNN EfficientNet avec data augmentation **offline** (Augmentations appliquées une fois sur les images avant entraînement. Les images transformées sont fixes et stockées)

Approche 3: Modèle CNN EficientNet avec data augmentation **inline** (Augmentations appliquées en temps réel à chaque époque)

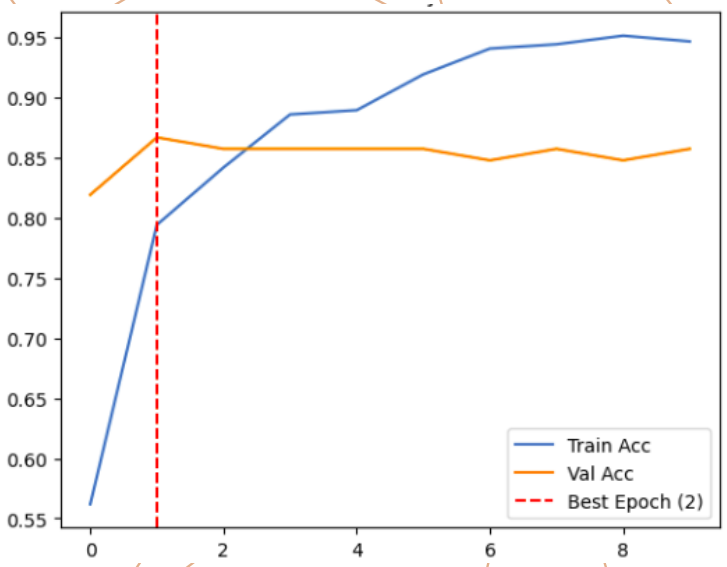
MODÉLISATION – RÉSULTATS

Approches	Accuracy (test)	Best accuracy (val)	Overfitting	F1 score	Test loss
Base	0,80	0,87 (Epoch2)	Fort	0,80	0,18
Data Augmentation Offline	0,87	0,88 (Epoch 4)	Léger	0,87	0,15
Data Augmentation Inline	0,81	0,87 (Epoch 12)	Faible	0,81	0,45

Sans data augmentation, le modèle montre des signes précoces de surapprentissage : dès l'epoch 3, la précision sur le jeu train continue d'augmenter, tandis que la précision sur le jeu de validation stagne, indiquant une mauvaise généralisation.
Avec data augmentation offline, l'overfitting est mieux contrôlé : il apparaît plus tard et de façon plus modérée. Le modèle apprend plus progressivement, ce qui se traduit par de meilleures performances globales
Enfin, **la data augmentation inline** permet de stabiliser les courbes d'apprentissage et de limiter fortement l'overfitting. Cependant l'erreur reste élevée, sûrement à cause de la variabilité qu'apporte la data augmentation inline.

MODÉLISATION-COURBES ACCURACY

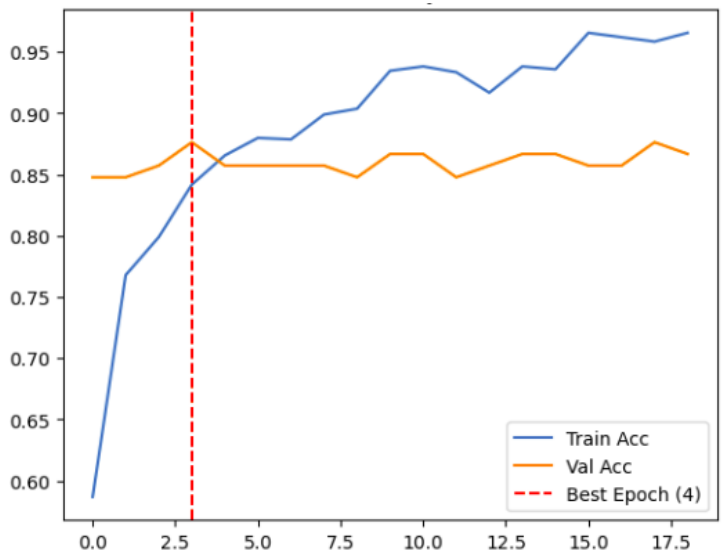
Approche 1



- Train Accuracy monte vite et dépasse la Val Accuracy dès l'epoch 3.
- Val Accuracy stagne rapidement (dès epoch 2-3), avec très peu d'évolution.
- Gap entre train et val reste élevé

=> Le modèle apprend très vite sur train (mémorise) mais ne généralise pas bien = **Overfitting fort.**

Approche 2

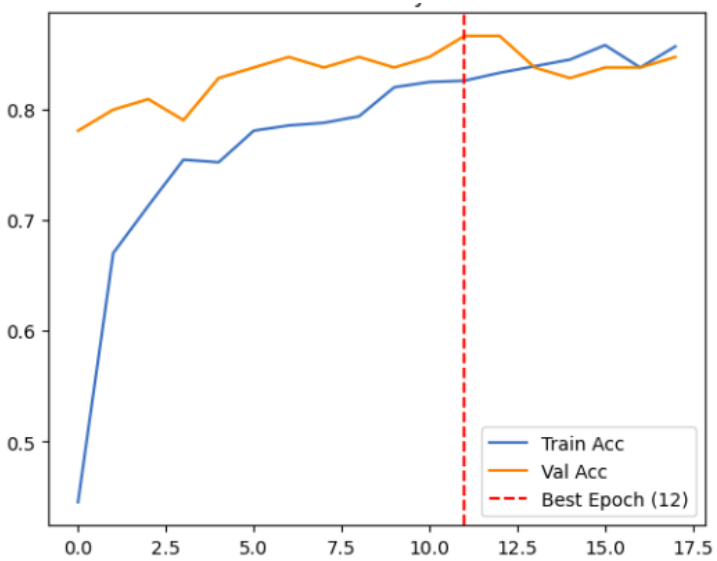


- Train Accuracy progresse plus lentement
- Val Accuracy monte puis oscille, avec une forme de "dents de scie« autour de 0,87/0,88.

⇒ Le modèle apprend plus progressivement, meilleure généralisation.

⇒ Les oscillations arrivent à l'epoch 5/6 donc l'overfitting est retardé.

Approche 3

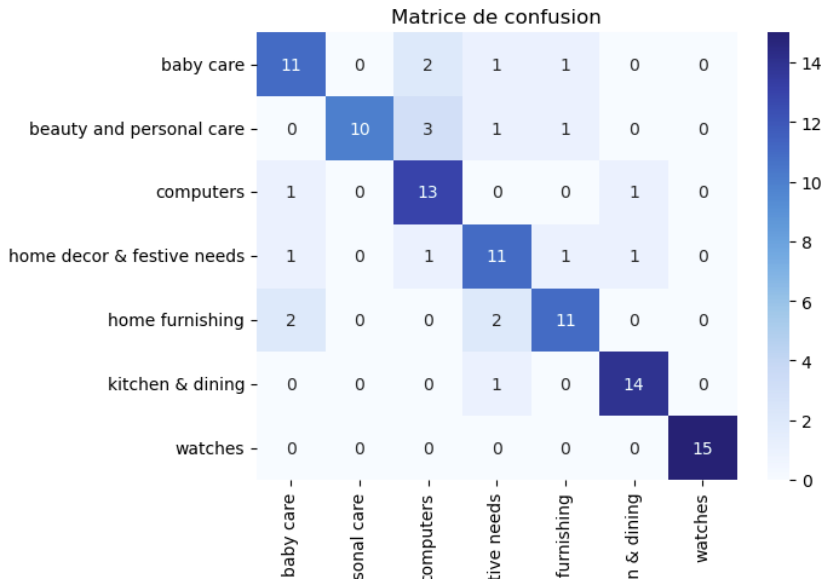


- Les courbes Train et Val Accuracy se suivent de près : excellente stabilité.
- L'écart entre les deux est faible, ce qui montre peu d'overfitting.
- Mais les deux plafonnent vers 0.85

=> Très peu d'overfitting mais potentiel d'apprentissage plafonné.

MODÉLISATION – MATRICE DE CONFUSION

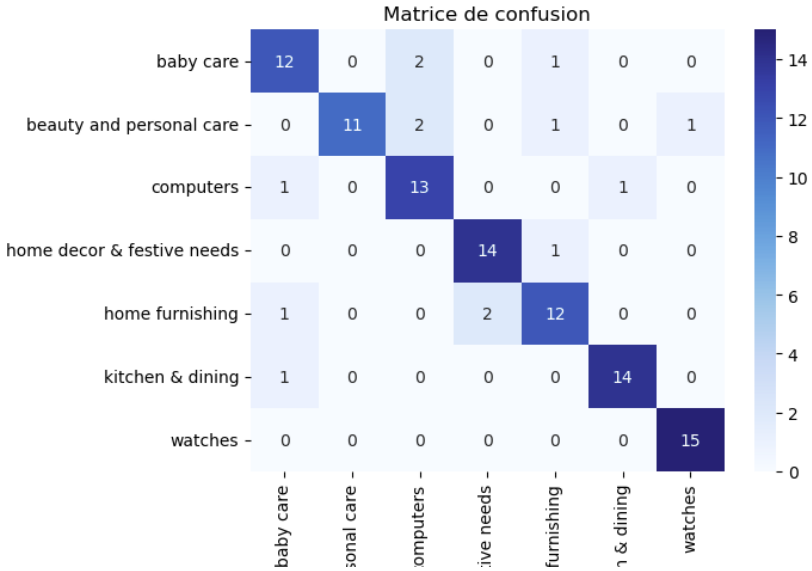
APPROCHE 1



- Plusieurs confusions notables : par exemple, *"beauty and personal care"* avec *"baby care"*.
- Moins de robustesse inter-classe.

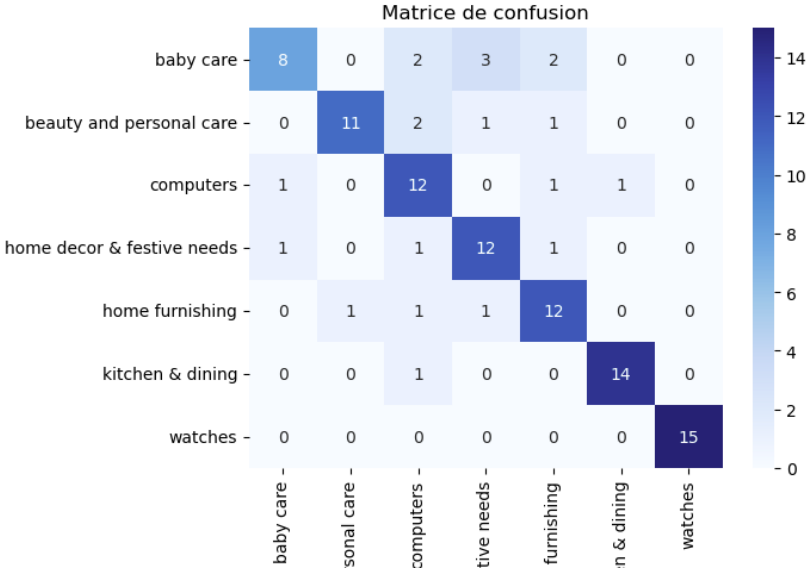
15 observations par classe

APPROCHE 2



- Meilleure répartition des prédictions : moins de confusion,
- Très bonne reconnaissance des classes comme *"home decor"*, *"kitchen & dining"*, ou *"watches"*.
- La **meilleure cohérence inter-classes**, surtout pour des classes proches.

APPROCHE 3



- Correcte mais moins nette que la 2 : certaines classes montrent une **légère dégradation** (ex. *baby care* et *home furnishing*).
- Malgré peu d'overfitting, la capacité discriminante reste inférieure à l'approche 2.

ARI : REQUÊTE SUR LE CHAMPAGNE

Objectif: tester la collecte de données de produits alimentaires contenant du champagne à l'aide de l'API OpenFoodFacts (API libre collaborative et gratuite)

Produit : Wouah gras vegan
Catégorie : Meat alternatives,fr:Faux Gras,fr:Plat-festif,Foie gras substitutes

Produit : Champagne Ruinart
Catégorie : Boissons, Boissons alcoolisées, Vins, Vins effervescents, Champagnes

	foodId	label	category	foodContentsLabel	image
0	8711812380571	Faux Gras	Produits à tartiner, Produits à tartiner salés...	Eau, levure alimentaire, huile de coco, amidon...	https://images.openfoodfacts.org/images/produc...
1	4005514008807	Wouah gras vegan	Meat alternatives,fr:Faux Gras,fr:Plat-festif,...	Eau, levure alimentaire*, huile de coco*, amid...	https://images.openfoodfacts.org/images/produc...
2	4056489843696	Rillettes de homard au cognac	Seafood, Fishes and their products, Fish prepa...	Chair de homard américain 49%, huile de colza,...	https://images.openfoodfacts.org/images/produc...
3	3258431220000		Boissons, Boissons alcoolisées, Vins, Vins eff...	Champagne	https://images.openfoodfacts.org/images/produc...
4	3049610004104	Veuve Clicquot Champagne Ponsardin Brut	Boissons et préparations de boissons, Boissons...	Champagne	https://images.openfoodfacts.org/images/produc...
5	3282946015837	Nicolas Feuillatte	Boissons, Boissons alcoolisées, Vins, Vins fra...	Champagne, Contient des _sulfites_	https://images.openfoodfacts.org/images/produc...
6	3416181017169	Champagne AOP, brut	Boissons, Boissons alcoolisées, Vins, Vins eff...	Champagne	https://images.openfoodfacts.org/images/produc...
7	3185370283905	Champagne Ruinart	Boissons, Boissons alcoolisées, Vins, Vins eff...	champagne	https://images.openfoodfacts.org/images/produc...
8	3245391237858	Champagne CHARLES VINCENT BRUT	Boissons, Boissons alcoolisées, Vins, Vins fra...	Champagne brut.	https://images.openfoodfacts.org/images/produc...
9	3256930103817	Champagne Blue Top Brut	Boissons et préparations de boissons, Boissons...	Champagne (_sulfites_)	https://images.openfoodfacts.org/images/produc...

CONCLUSION

- **Faisabilité confirmée :** L'étude démontre la viabilité d'une classification automatique des données visuelles et textuelles, avec des résultats performants, même sur un jeu de données réduit.
- **Données textuelles:** Les méthodes avancées d'extraction de features comme BERT et USE ont montré de très bonnes performances, mais c'est l'approche plus classique TF-IDF qui s'est révélée la plus efficace dans ce contexte.
- **Données visuelles :** Le modèle **EfficientNet** s'est imposé comme la méthode la plus performante pour l'extraction de features sur les images.
- **Modèle le plus robuste :** Le CNN combiné à une data augmentation offline a permis d'obtenir le meilleur compromis entre performance, réduction de l'erreur et contrôle de l'overfitting,
- **Respect du RGPD :** L'ensemble des données utilisées a été traité conformément aux principes du RGPD. Aucun élément personnel ou sensible n'a été collecté, et toutes les données exploitées sont anonymisées, non traçables et utilisées uniquement à des fins expérimentales.

MERCI