

Parcours AI Engineer

SOUTENANCE PROJET 6

“Classification de biens de consommation par le texte et l’image automatisée”

Stéphanie Duhem - Décembre 2024

00. CONTEXTE & DÉROULÉ DU PROJET

LE CONTEXTE

“Place de Marché” est une marketplace de e-commerce.

Elle souhaite mettre en place une classification automatique des produits mis en ligne par les vendeurs grâce aux photos ou aux descriptions des produits. Elle demande donc une étude de faisabilité d'un moteur de classification de produits automatique, basé sur des textes ou sur des images.

LES 4 GRANDES ÉTAPES DU PROJET

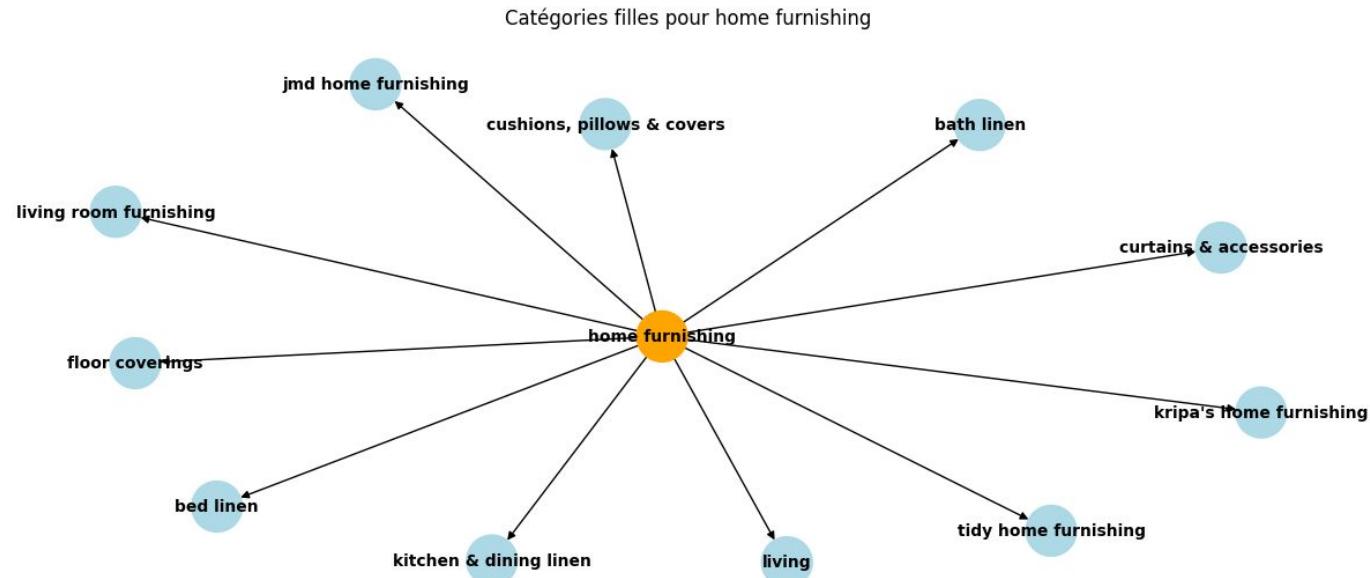
- Analyses et étude de faisabilité basée sur les descriptions
- Analyses et étude de faisabilité basée sur les photos
- Tests de modèle pour une classification supervisée basée sur les images
- Étude pour mise en place d'une collecte de données via API pour ajouter une nouvelle catégorie de produits



01-A. Analyse exploratoire globale

LES DONNÉES

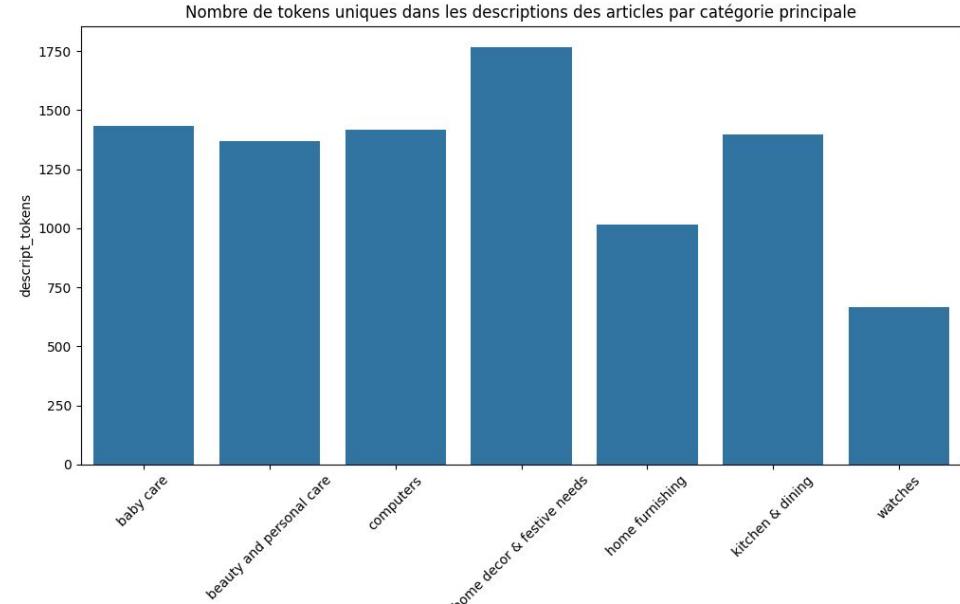
- 1050 articles classés par catégories avec leurs descriptions et leurs images
- Il y a 7 catégories principales (150 articles par catégories) qui ont chacune beaucoup de sous-catégories de produits
 - certaines catégories mères ont jusqu'à 5 niveaux de catégories filles
 - pour la classification, utilisation des 7 catégories principales



01-B. Analyse exploratoire des TEXTES

LES DESCRIPTIONS

- certaines catégories ont des descriptions plus variées avec un champ lexical bien plus importants que d'autres
- beaucoup de mots en commun entre certaines catégories



Nuage de mots des descriptions des articles pour la catégorie principale: home furnishing



Nuage de mots des descriptions des articles pour la catégorie principale: baby care

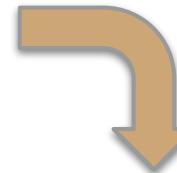


01-C. Prétraitements des TEXTES- Approche BoW

- Suppression des stops words en anglais ('of', 'they', 'did', 'or' etc.)
- Suppression des mots présents dans + de 3 catégories
- Tokenisation (résultat sans ponctuation et chiffres seuls)

```
key features of mom and kid baby girl's printed blue, grey top & pyjama set fabric: cotton brand color: blue, grey,  
mom and kid baby girl's printed blue, grey top & pyjama set price: rs. 309 girls pyjamaset,specifications of mom an  
d kid baby girl's printed blue, grey top & pyjama set general details pattern printed ideal for baby girl's night s  
uit details number of contents in sales package pack of 1 fabric cotton type top & pyjama set neck round nack in th  
e box 1 top & pyjama set
```

484



```
['mom', 'kid', 'girl', 'pyjama', 'cotton', 'mom', 'kid', 'girl', 'pyjama', 'girls', 'pyjamaset', 'mom', 'kid', 'gi  
l', 'pyjama', 'girl', 'cotton', 'pyjama', 'neck', 'round', 'nack', 'pyjama']
```

22

01-C. Prétraitements des TEXTES- Approche BoW

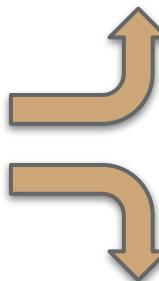
2 méthodes de réduction de dimensions des mots :

- Lemmatisation (forme canonique)
- Stemming (racine des mots)

Lemmatisation

```
{'antique', 'small', 'wonderful', 'surface180', 'indeed', 'showpiece', 'limitation', 'carving', 'panel', 'drawer',
'fascinate', 'front', 'feature', 'h', 'vibrant', 'astonish', 'handcraft', 'look', 'allure', 'golden', 'photography'}
```

21



```
surface180 wonderful wooden antique drawer box showpiece - 20 cm (wooden, green)
price:
rs. 1,112

the vibrant antique drawer box looks astonishing with alluring carvings and bright colors. fea-
turing body dimensions of 13 (l) x 10 (w) x 18 (h) cms, it is indeed an amazing home decorative item. the small and
compact wooden antique style box has different golden carvings on each front panel of the drawer and around the dr-
awer box , which makes it fascinating.

note: this is handcraft item so each item will be different than other du
e to limitation of photography.

the vibrant antique drawer box looks astonishing with alluring carvings and bright
colors. featuring body dimensions of 13 (l) x 10 (w) x 18 (h) cms, it is indeed an amazing home decorative item. t
he small and compact wooden antique style box has different golden carvings on each front panel of the drawer and a
round the drawer box , which makes it fascinating.

note: this is handcraft item so each item will be different t
han other due to limitation of photography.

1078
```

Stemming

```
{'carv', 'allur', 'featur', 'small', 'surface180', 'photographi', 'panel', 'fascin', 'drawer', 'inde', 'front', 'sh
owpiec', 'h', 'vibrant', 'wonder', 'astonish', 'handcraft', 'antiqu', 'look', 'limit', 'golden'}
```

21

01-D. Extraction des features des TEXTES

Tests de 5 approches différentes :

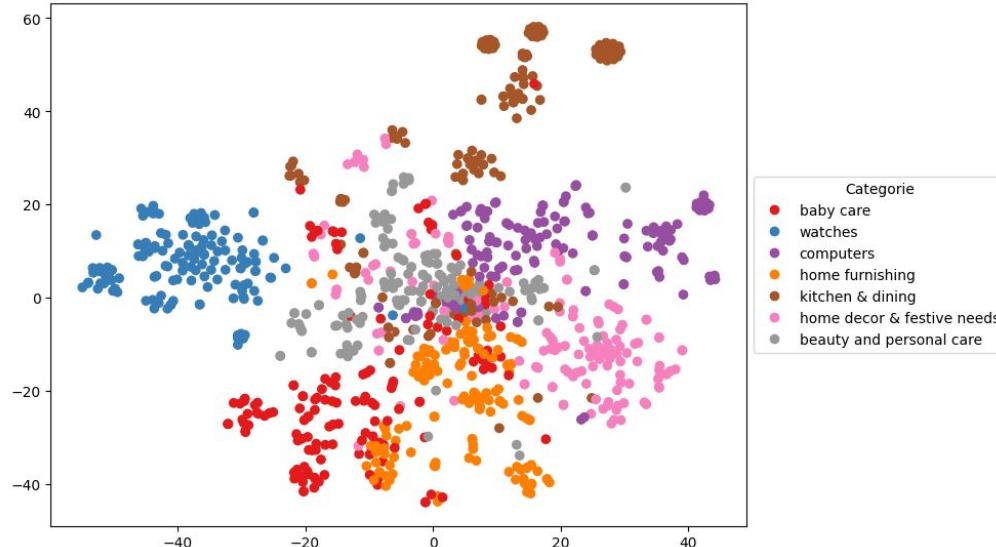
- 2 approches de type bag-of-words (*après prétraitements et tokenisation*, représenter chaque document par un vecteur de la taille du vocabulaire $|V|$) :
 - CountVectorizer (comptage simple de mots) ,
 - Tf-idf (Term Frequency - Inverse Document Frequency, pondérer l'importance d'un mot par rapport à l'ensemble du corpus) ;
- 3 approches de type word/sentence embedding (*après tokenisation*, représenter en vecteurs de taille inférieure, similarité de sens) :
 - Word2Vec (word embedding classique) ,
 - BERT (Bidirectional Encoder Representations from Transformers, pour comprendre le contexte des mots dans une phrase) ,
 - USE (Universal Sentence Encoder, spécifiquement pour générer des embeddings de phrase) .

01-E. Etude de faisabilité sur les TEXTES- Résultats

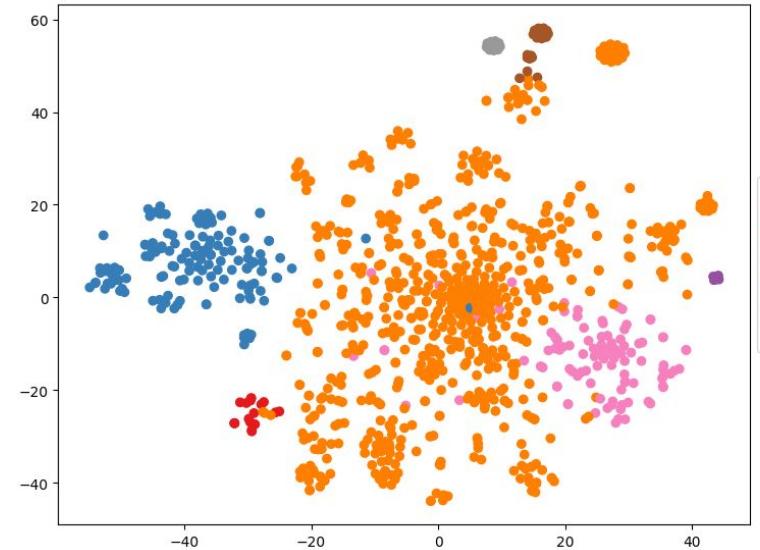
Le plus mauvais résultat est basé sur une approche BoW avec CountVectorizer :

- peu de 'noyaux' de catégories bien définis,
- superpositions, éparpillements des ensembles,
- ARI Score très faible du Kmeans (0,16) => très faible correspondance avec les catégories réelles.

Représentation des features textuelles issues de CountVectorizer - Par catégories réelles des produits
Silhouette Score: -0.0587

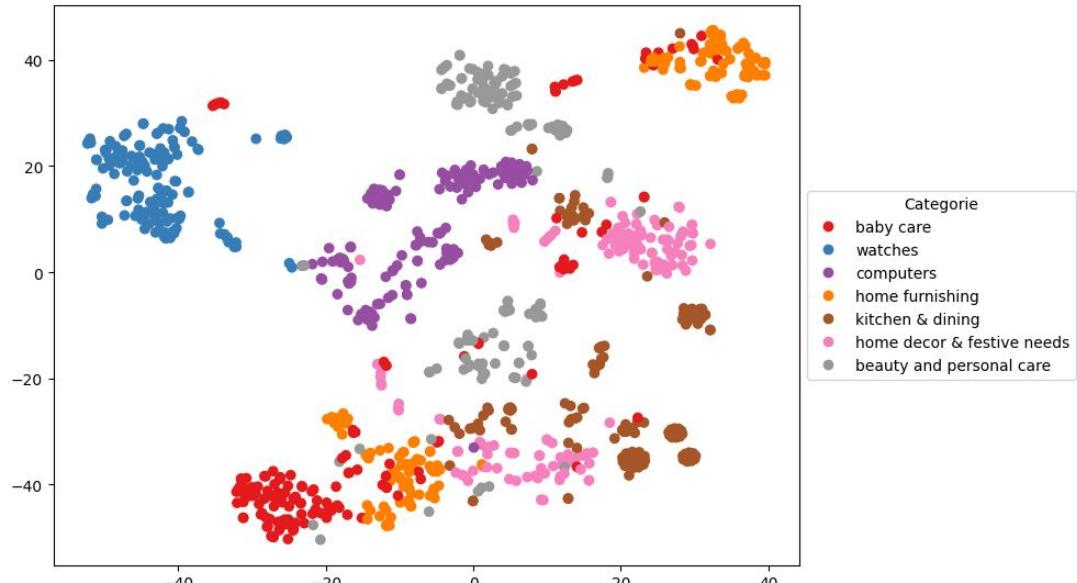


Représentation des features textuelles issues de CountVectorizer - Par clusters KMeans
Silhouette Score: -0.0144

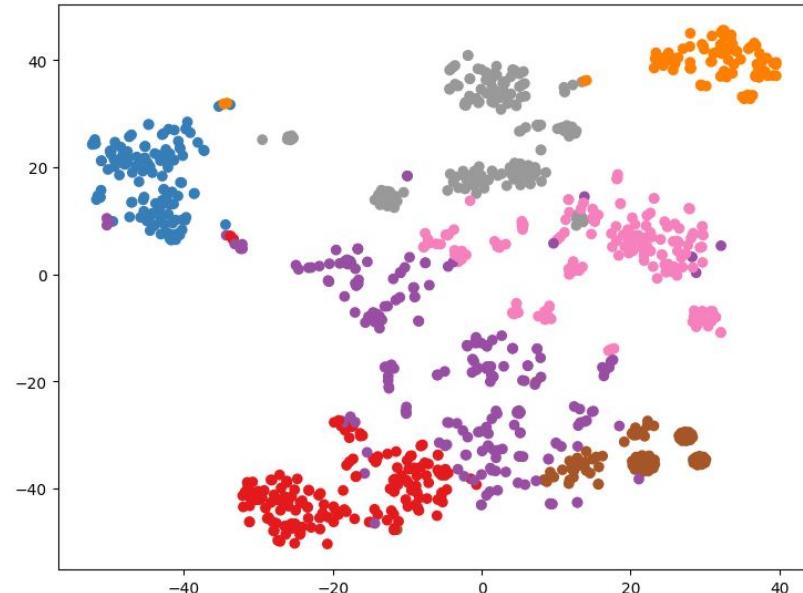


01-E. Etude de faisabilité sur les TEXTES- Résultats

Représentation des features textuelles issues de USE - Par catégories réelles des produits
Silhouette Score: 0.0613



Représentation des features textuelles issues de USE - Par clusters KMeans
Silhouette Score: 0.1018



Le meilleur résultat est basé sur une approche Word2Vec avec USE :

- clusters des catégories mieux définis, plus ramassés,
- meilleur ARI Score sur le KMeans obtenus de toute l'étude (0,33) qui reste faible, peu de correspondance avec les catégories réelles.

01-E. Etude de faisabilité sur les TEXTES - Résultats

	Méthode	Silhouette_Score	Silhouette_Score_KMeans	ARI_Score_Réel_Kmeans	Durée_extraction_features
0	CountVectorizer	-0.058690	-0.014398	0.169308	0 min 0 sec
1	TfidfVectorizer	0.029729	0.046718	0.210076	0 min 0 sec
2	Word2Vec	0.085745	0.333172	0.218812	0 min 0 sec
3	BERT bert-base-uncased	0.036839	0.133158	0.268401	2 min 34 sec
4	roBERTa - roberta-base	0.006083	0.115572	0.201345	2 min 28 sec
5	USE	0.061317	0.101817	0.327074	0 min 1 sec

CONCLUSIONS

- La meilleure modélisation semble être la USE (meilleur ARI Score notamment)
- Scores de silhouette proches de zéro (classes peu distinctes)
- ARI scores faibles (mauvaise correspondance avec le réel)

=> La complexité des données n'est jamais correctement récupérée par les différentes approches.

=> Les différentes classes semblent avoir des similarités sémantiques importantes et des subtilités de contextes difficiles à capter par les différentes approches.

PERSPECTIVES

- Refaire les catégories et les sous-catégories ?

Car le nombre important de catégories et de sous-catégories mériterait une simplification, en prenant garde à ne pas impacter l'expérience client.

- Augmenter le nombre d'échantillon dans nos données ?

Les informations captées permettraient une meilleure représentation des données, et donc une meilleure classification.

02-A. Analyse exploratoire des IMAGES

LES DONNÉES

- 1050 articles classés par catégories avec leurs images
- Dans l'ensemble plutôt bonne qualité d'images (majorité d'articles détournés sur fond blanc, image nette)
- Dimensions d'images très variables même au sein d'une même catégories.

PRÉ-TRAITEMENT

=> Rajout de bande de pixel sur les côtés pour obtenir des images carrées (pour les approches avec les CNN)



Home furnishing

Watches

Baby cares

Home decor &
festive needs

Kitchen &
dining

Beauty and
personal care

Computers

02-B. Extraction des features des IMAGES

Tests de 8 approches différentes :

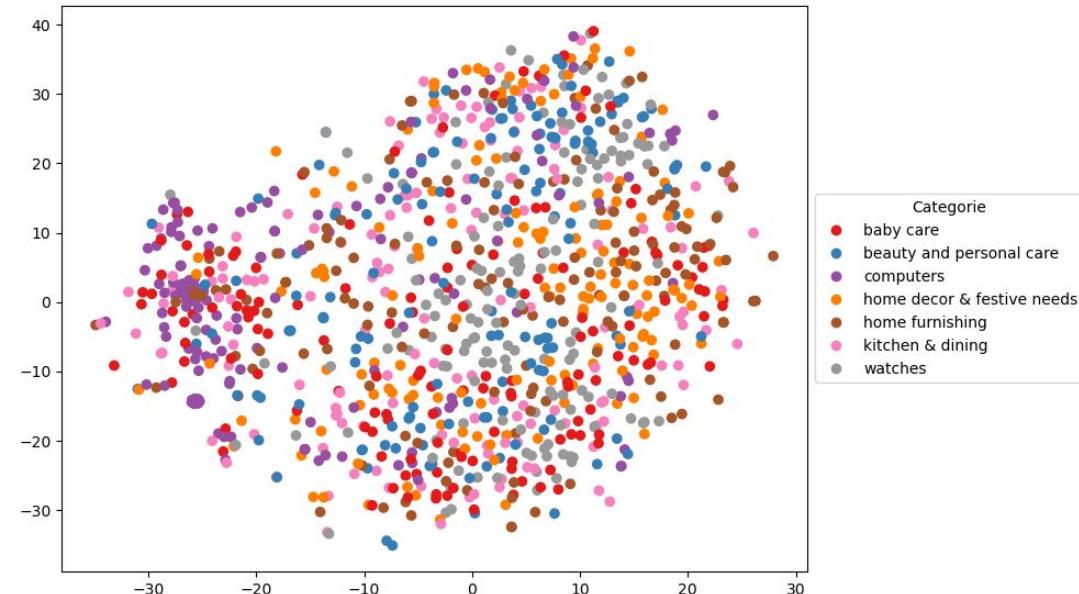
- 2 approches de type **bag-of-images** (*après extraction des caractéristiques locales puis les regrouper en 'mots visuel'*, représenter chaque image par un histogramme de mots visuels, puis utiliser en vecteur de taille fixe) :
 - SIFT (détection des points d'intérêt),
 - ORB (détection des points d'intérêt, avec invariance à la rotation) ;
- 6 approches de type **CNN par Transfert Learning** (*après redimensionnement des images, utiliser des CNN déjà entraînés sur le dataset Imagenet*) :
 - VGG16 (16 couches de convolution et de pooling),
 - ResNet50 (connexions résiduelles pour faciliter l'apprentissage),
 - InceptionV3 (chaque module Inception applique plusieurs convolutions avec des tailles de filtres différente),
 - EfficientNetB0 (mise à l'échelle uniforme pour équilibrer la profondeur, la largeur et la résolution du réseau),
 - MobileNetV2 (léger et efficace, conçu pour les mobiles),
 - NASNetMobile (recherche automatisée d'architecture, optimisés pour les mobiles),
 - DenseNet (connexions denses entre les couches).

02-C. Etude de faisabilité sur les IMAGES - Résultats

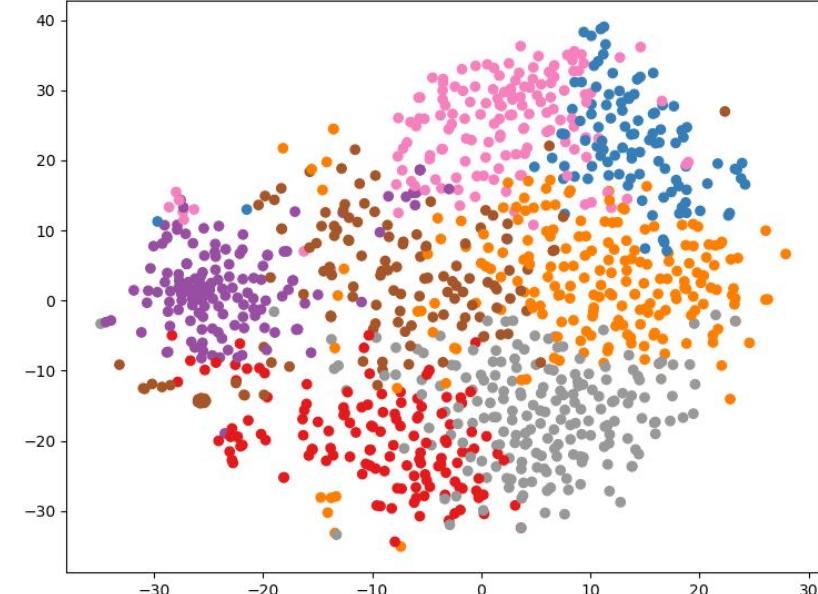
Le plus mauvais résultat est basé sur une approche Bol avec ORB :

- aucune catégorie définie,
- superpositions, éparpillements des ensembles,
- ARI Score très faible du Kmeans (0,05) => correspondance proche d'une partition obtenue par hasard.

Représentation des features des images issues de ORB - Par catégories réelles des produits
Silhouette Score: -0.0158

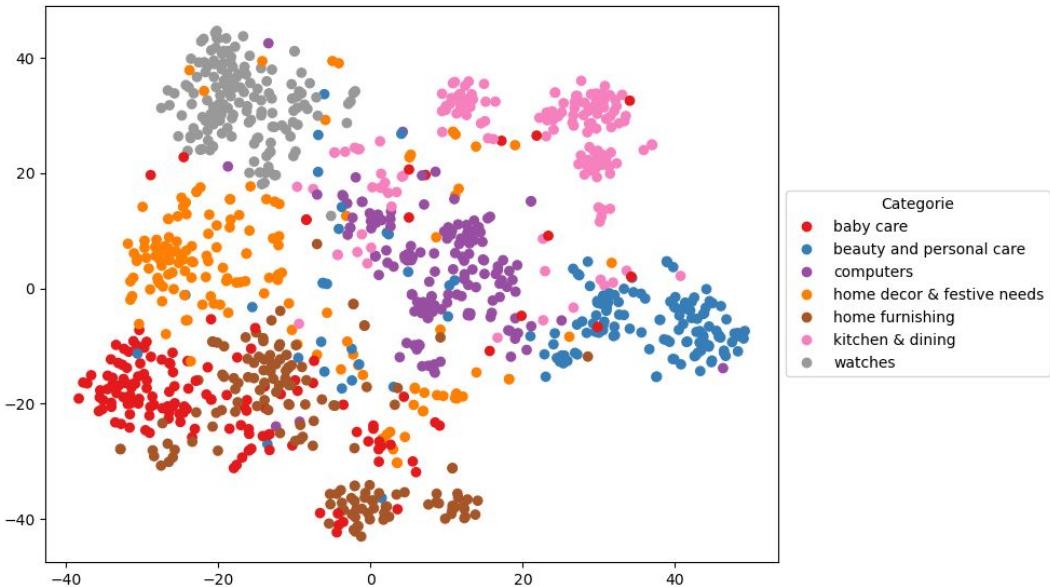


Représentation des features des images issues de ORB - Par clusters KMeans
Silhouette Score: 0.0663

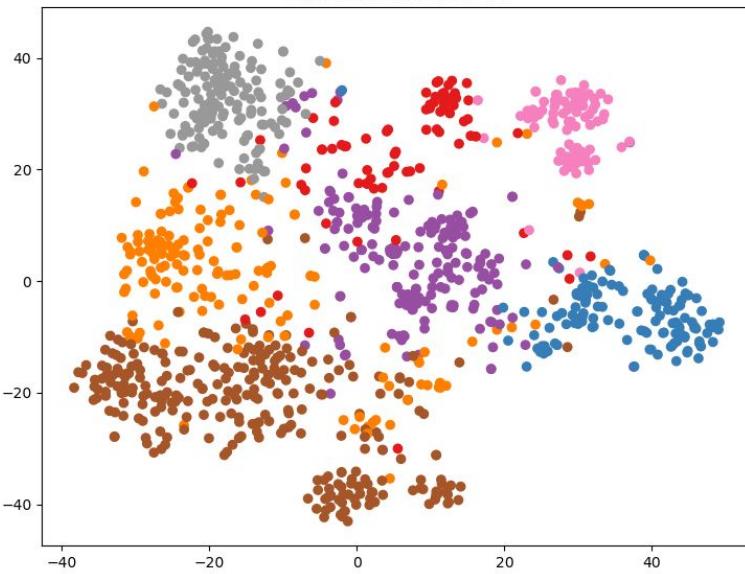


02-C. Etude de faisabilité sur les IMAGES - Résultats

Représentation des features des images issues de DenseNet - Par catégories réelles des produits
Silhouette Score: 0.0498



Représentation des features des images issues de DenseNet - Par clusters KMeans
Silhouette Score: 0.0648



Le meilleur résultat est basé sur une approche CNN par Transfert Learning avec DenseNet :

- clusters par catégories mieux définis, plus ramassés,
- meilleur ARI Score sur le KMeans obtenu de toute l'étude (0,53), correspondance moyenne avec les catégories réelles.

02-C. Etude de faisabilité sur les IMAGES - Résultats

	Méthode	Silhouette_Score	Silhouette_Score_KMeans	ARI_Score_Réel_Kmeans	Durée_extraction_features
0	SIFT	-0.111566	0.142517	0.052924	0 min 58 sec
1	ORB	-0.015801	0.066327	0.047535	0 min 31 sec
2	VGG16	-0.004886	0.035967	0.298764	1 min 17 sec
3	MobileNetV2	0.034515	0.041874	0.478463	1 min 20 sec
4	DenseNet	0.049802	0.064804	0.525897	1 min 34 sec
5	ResNet50	0.021253	0.026041	0.339214	1 min 23 sec
6	InceptionV3	0.049003	0.046550	0.341701	1 min 29 sec
7	EfficientNetB0	0.043875	0.039094	0.435899	1 min 20 sec
8	NASNetMobile	0.062614	0.083318	0.459579	1 min 46 sec

CONCLUSIONS

- Les approches Bol, SIFT et ORB, ne sont pas du tout adaptées à notre projet de classification d'images par catégorie de produits.
- Scores de silhouette proches de zéro (classes peu distinctes)
- ARI scores moyens

=> La complexité des données est relativement bien récupérée par les différents CNN.

=> Le bon compromis performances et coûts se trouve avec le DenseNet:

- 2ème au score de silhouette sur réel (0,05)
- 2ème en terme de temps de calcul (1min34)
- 1er au ARI score entre le réel et le kmeans (0,53)

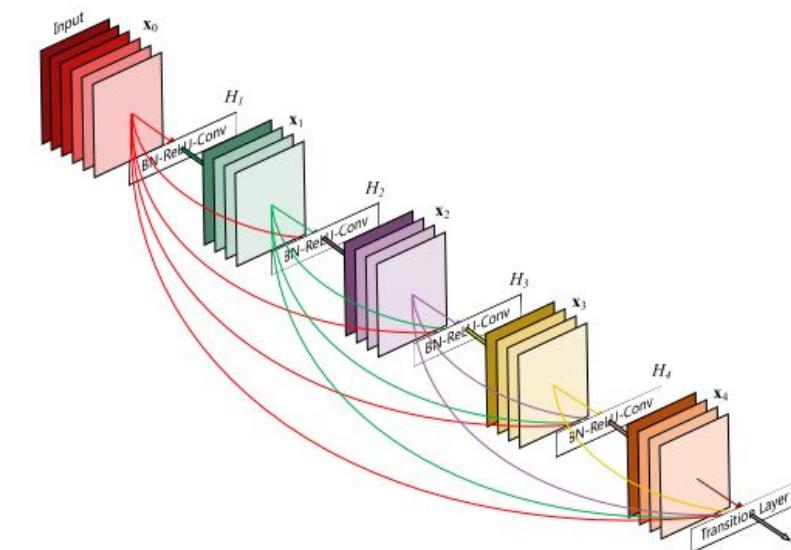
DenseNet retenu pour la classification d'images par catégories de produits.

03-A. Classification basée sur les IMAGES - DenseNet

Architecture du DenseNet121

“conv” = BatchNormalisation-ReLU-Convolution ; “Transition Layer” = Convolution et Pooling

Layers	Output Size	DenseNet-121
Convolution	112×112	
Pooling	56×56	
Dense Block (1)	56×56	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$
Transition Layer (1)	56×56	
	28×28	
Dense Block (2)	28×28	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$
Transition Layer (2)	28×28	
	14×14	
Dense Block (3)	14×14	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 24$
Transition Layer (3)	14×14	
	7×7	
Dense Block (4)	7×7	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 16$
Classification Layer	1×1	



Couche de sortie initiale retirée et remplacée par des couches adaptées à notre problème

03-B. Classification basée sur les IMAGES - Scénarii

LES 4 SCÉNARI

- Modèle 1A : Entraînement avec les poids imagenet - SANS Data Augmentation
- Modèle 2B : Entraînement avec les poids imagenet - AVEC Data Augmentation
- Modèle 2A : Entraînement SANS les poids imagenet - SANS Data Augmentation
- Modèle 2B : Entraînement SANS les poids imagenet - AVEC Data Augmentation

POINTS TECHNIQUES

- Reproductibilité des résultats
- Data Augmentation



Rotation Gauche ou Droite de 2 degrés maximum



Zoom x1,2

03-C. Classification basée sur les IMAGES - Résultats

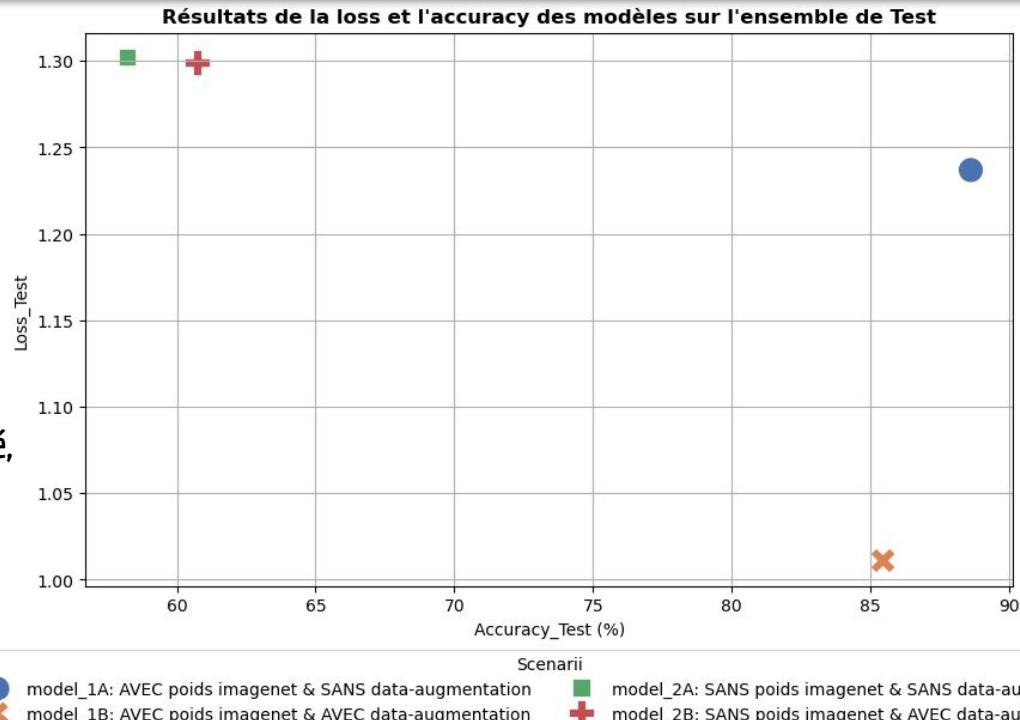
Scenario	Description	Accuracy_TrainVal	Loss_TrainVal	Accuracy_Test	Loss_Test	Durée_training
model_1A	AVEC poids Imagenet & SANS data-augmentation	97.533631	0.443065	88.607597	1.236787	3 min 4 sec
model_1B	AVEC poids Imagenet & AVEC data-augmentation	97.197312	0.381347	84.810126	1.411997	2 min 38 sec
model_2A	SANS poids Imagenet & SANS data-augmentation	93.946189	0.325691	65.189874	1.777045	17 min 57 sec
model_2B	SANS poids Imagenet & AVEC data-augmentation	80.156952	0.589508	68.987340	1.221051	14 min 10 sec

CONCLUSIONS

- En terme de coût, les modèles pré-entraînés sont bien plus économies que les modèles non pré-entraînés.
- Pour un résultat à peine mieux que le hasard le nombre d'epoch a été x4 pour les modèles non pré-entraînés
- Le meilleur modèle est le Modèle 1A

La data augmentation

Amélioration des performances sur un modèle non pré-entraîné, mais dégradation sur le modèle pré-entraîné
 => Elle rajoute de la complexité sur un modèle déjà entraîné, mais aide les modèles non pré-entraînés



03-D. Classification basée sur les IMAGES - Perspectives

AMÉLIORATION & RECOMMANDATION

- Data Augmentation : seulement pour certaines sous-catégories d'articles
- Garder le modèle pré-entraîné pour limiter le coût à l'entraînement tout en maintenant une bonne accuracy

PERSPECTIVE

Approche multimodale : certaines produits sont trompeurs

=> cumuler la classification basée sur les images et celle basée sur les textes des descriptions

=> maintenance plus conséquente

Catégorie réelle : baby care



Pred : home furnishing



Pred : home furnishing



Pred : home furnishing

04-A. Test de l'API Openfood Facts - Conformité RGPD

Règlement Général sur la Protection des Données (RGPD) de l'Union Européenne

5 grands principes :

- **Licéité, loyauté et transparence** : les données doivent être traitées de manière licite, loyale et transparente, et les individus doivent être informés de leur collecte et utilisation ;
- **Limitation des finalités** : les données doivent être collectées pour des finalités spécifiques et légitimes ;
- **Minimisation des données** : seules les données nécessaires doivent être collectées ;
- **Exactitude** : les données doivent être exactes et mises à jour ;
- **Limitation de la conservation** : les données ne doivent être conservées que le temps nécessaire pour atteindre les finalités pour lesquelles elles ont été collectées.

04-B. Test de l'API Openfood Facts - Résultats

- Les objets décrits ne sont pas ce qu'on attend (sauces, vinaigrettes, truffes au chocolat etc..)
- Manque de visuels (seulement 3 images sur 10 produits)
- Beaucoup de ressources à déployer pour utiliser ces informations.

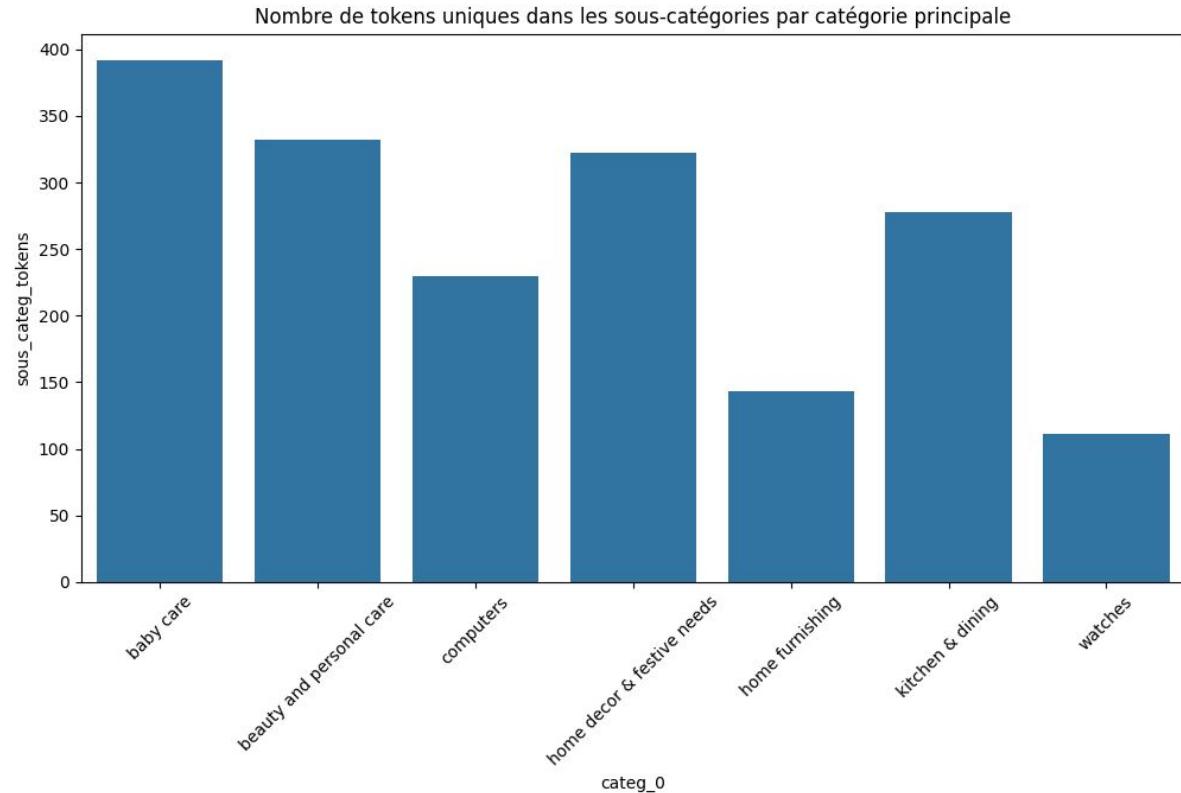
	Food ID	Label	Category	Food Contents Label	Image
0	food_a656mk2a5dmqb2adlamu6beihduu	Champagne	Generic foods	NR	https://www.edamam.com/food-img/a71/a718cf3c52add522128929ff324d2ab.jpg
1	food_b753ithamdb8psbt0w2k9aquo06c	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR; GARLIC; DIJON MUSTARD; SEA SALT.	NR
2	food_b3dyababj054xobm6r8jzbghjgqe	Champagne Vinaigrette, Champagne	Packaged foods	INGREDIENTS: WATER; CANOLA OIL; CHAMPAGNE VINEGAR; SUGAR; OLIVE OIL; SALT; DRIED GARLIC; DRIED SHALLOTS; BLACK PEPPER; XANTHAN GUM; SPICE	https://www.edamam.com/food-img/d88/d88b64d97349ed062368972113124e35.jpg
3	food_a9e0ghsamvoc45bwa2ybsa3gken9	Champagne Vinaigrette, Champagne	Packaged foods	CANOLA AND SOYBEAN OIL; WHITE WINE (CONTAINS SULFITES); WATER; VINEGARS (CHAMPAGNE AND WHITE WINE); SUGAR; SALT; MUSTARD SEED; MONOSODIUM GLUTAMATE; GARLIC*; ONION*; SPICE; XANTHAN GUM; MOLASSES; CALCIUM DISODIUM EDTA ADDED TO PROTECT FLAVOR; CHIVES*; TAMARIND; NATURAL FLAVOR.	NR
4	food_an4jjeaucpus2a3u1nl8auhe7q9	Champagne Vinaigrette, Champagne	Packaged foods	WATER; CANOLA AND SOYBEAN OIL; WHITE WINE (CONTAINS SULFITES); VINEGARS (CHAMPAGNE AND WHITE WINE); SUGAR; SALT; MUSTARD SEED; MONOSODIUM GLUTAMATE; GARLIC*; ONION*; SPICE; XANTHAN GUM; POTASSIUM SORBATE ADDED TO MAINTAIN FRESHNESS; MOLASSES; CALCIUM DISODIUM EDTA ADDED TO PROTECT FLAVOR; CHIVES*; TAMARIND.	NR
5	food_bmu5dmkazwuvpaas5prh1daa8jxs0	Champagne Dressing, Champagne	Packaged foods	SOYBEAN OIL; WHITE WINE (PRESERVED WITH SULFITES); WATER; WHITE WINE VINEGAR; SUGAR; SALT; SPICES (INCLUDING MUSTARD SEED); MONOSODIUM GLUTAMATE; GARLIC*; ONION*; XANTHAN GUM; MOLASSES; CALCIUM DISODIUM EDTA ADDED TO PROTECT FLAVOR; VINEGAR; CORN SYRUP; CARAMEL COLOR; CHIVES*; NATURAL FLAVOR; TAMARIND.	https://www.edamam.com/food-img/ab2/ab2459fc2a98cd35f68b848be2337ecb.jpg
6	food_alpl44taoyv11ra0llc1qa8xcull	Champagne Buttercream	Generic meals	sugar; butter; shortening; vanilla; champagne; milk	NR
7	food_am5egz6aq3fpjlaf8xpdkbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanilla extract; champagne; powdered sugar	NR
8	food_bcz8rlhajk1fuva0vkfmeakbouco	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; shallot; honey; Salt; pepper	NR
9	food_a79xmny6toreaeukbroa0thhh0	Champagne Chicken	Generic meals	Flour; Salt; Pepper; Boneless, Skinless Chicken Breasts; Butter; Olive Oil; Champagne; Mushrooms; Heavy Whipping Cream	NR

ANNEXES

01-A. Analyse exploratoire des TEXTES

LES SOUS CATÉGORIES

- certaines catégories filles ont des noms plus variés avec un champ lexical bien plus importants que d'autres



01-A. Analyse exploratoire des TEXTES

LES DESCRIPTIONS

- beaucoup de mots en commun entre les catégories

Nuage de mots dans les descriptions des articles pour la catégorie principale: home furnishing



Nuage de mots dans les descriptions des articles pour la catégorie principale: baby care



02-A. Analyse exploratoire des IMAGES

LES DONNÉES

- Dimensions d'images très variables même au sein d'une même catégories.

=> Rajout de bande de pixel sur les côtés pour obtenir des images carrées (pour les approches avec les CNN)

