Classifiez automatiquement des biens de consommation

Agenda

1. Contexte
2. Approche
3. Présentation des jeux de données
4. Analyse textuelle
5. Analyse des images
6. Prochaines étapes
7. Annexe : l’environnement technique

1) Contexte

Objectif	L'entreprise « **Place de marché** » souhaite lancer une market place e-commerce. Sur ce dernier, les vendeurs proposent des articles à vendre en postant une photo et une description. **Actuellement, l'affection d'une catégorie est effectuée manuellement par le vendeur donc sujette à erreur. Il devient nécessaire d'automatiser cette tâche.**
Données mises à disposition	Jeu de données d'articles avec une photo et une description associée
Mission	Réalisation d'une première étude de faisabilité d'un moteur de classification basé sur une image et une description pour l'automatisation de la catégorie de l'article.

2) Approche

Présentation des jeux de données (Taille du jeu de données, Type de données, nombre de lignes, nombre de colonnes, données manquantes…)

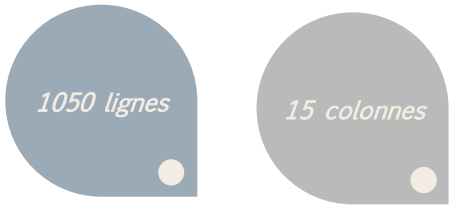
Nettoyage des données (Récupération de la catégorie principales des produits, Capwords des noms de produits, suppression de la punctuation et des stopwords…), Bag of Words, TF-IDF, BERT-Embedding, Word2Vec, Doc2Vec.

Nettoyage des images (Noir et Blanc, Suppression du bruit, Contraste, …), SIFT, CNN

3) Présentation des jeux de données (1/2)

	uniq_id	crawl_timestamp	product_url	product_name	product_category_tree	pid	retail_price
0	55b85ea15a1536d46b7190ad6fff8ce7	2016-04-30 03:22:56 +0000	http://www.flipkart.com/elegance-polyester-mul...	Elegance Polyester Multicolor Abstract Eyelet ...	["Home Furnishing >> Curtains & Accessories >>...	CRNEG7BKMFFYHQ8Z	1899.0
1	7b72c92c2f6c40268628ec5f14c6d590	2016-04-30 03:22:56 +0000	http://www.flipkart.com/sathiyas-cotton-bath-t...	Sathiyas Cotton Bath Towel	["Baby Care >> Baby Bath & Skin >> Baby Bath T...	BTWEGFZHGBXPHZUH	600.0
2	64d5d4a258243731dc7bbb1eef49ad74	2016-04-30 03:22:56 +0000	http://www.flipkart.com/eurospa-cotton-terry-f...	Eurospa Cotton Terry Face Towel Set	["Baby Care >> Baby Bath & Skin >> Baby Bath T...	BTWEG6SHXTDB2A2Y	NaN
3	d4684dcdc759dd9cdf41504698d737d8	2016-06-20 08:49:52 +0000	http://www.flipkart.com/santosh-royal-fashion-...	SANTOSH ROYAL FASHION	["Home Furnishing >> Bed Linen >>	BDSEJT9UQWHDUBH4	2699.0

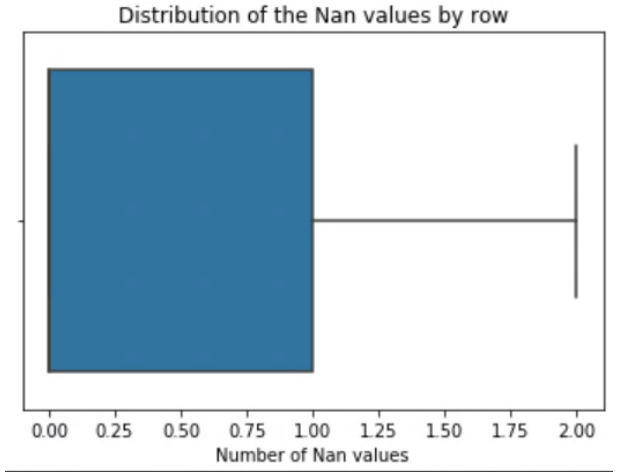
Extrait du jeu de données



Indicateurs clés du jeu de données

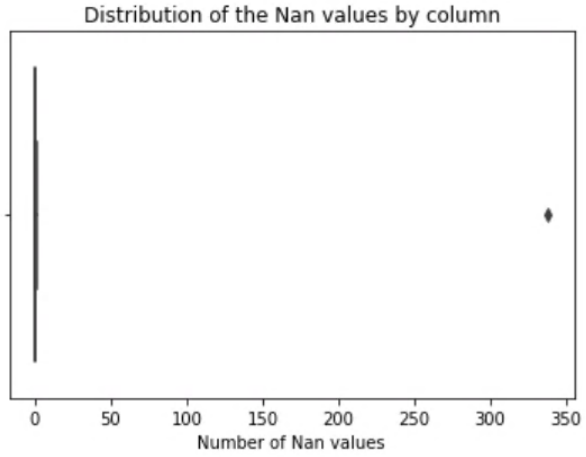
3) Présentation des jeux de données (2/2)

Analyse des Nan par ligne



Caractéristiques	Valeur
Nombre de lignes	1050
Moyenne du nombre de valeurs	0.324763
Ecart type	0.470539
Min de valeurs	0
25%	0
50%	0
75%	1
Max	2

Analyse des Nan par colonne



Caractéristiques	Valeur
Nombre de colonnes	15
Moyenne du nombre de valeurs	22.733333
Ecart type	87.216862
Min de valeurs	0
25%	0
50%	0
75%	0.5
Max	338

Le jeu de données fournit est bien complété. Aucune complétude de données n'est necessaire pour faire du Machine Learning.

4) Analyse textuelle

Nettoyage des données

1 2 3

Récupération de la catégorie principale

Les changements opérés sont :

Id	Product_category_tree
0	["Home Furnishing >> Curtains & Accessories >>..
1	["Baby Care >> Baby Bath & Skin >> Baby Bath T...
....	
1049	["Baby Care >> Baby & Kids Gifts >> Stickers >...*

Id	Product_main_category
0	Home Furnishing
1	Baby Care
...	
1049	Baby Care

Récupération de la marque dans le champs Description

Le nom de la marque est extraite du champ descrption :

Id	Description
0	Key Features of **Elegance** Polyester Multicolor ..
1	Specifications of **Sathiyas** Cotton Bath Towel (.
....	
1049	Buy **Uberlyfe** Large Vinyl Sticker for Rs.595 on...

Id	Revamped_brand
0	**Elegance**
1	**Sathiyas**
...	
1049	**Uberlyfe**

Nettoyage sur les specifications d'un produit (Suppression des clés valeurs, suppression de la ponctuation, suppression des stopwords, lemmatization)

Les changements suivants sont opérés :

Id	Product_speciifications
0	{"product_specification"=>[{"key"=>"Brand", "v..
1	{"product_specification"=>[{"key"=>"Machine Wa..
....	
1049	{"product_specification"=>[{"key"=>"Sales Pack...

Id	Clean_Product_speciifications
0	[brand, elegance, designed, door, type, eyelet...
1	[machine, washable, material, cotton, design, ...
....	
1049	[sale, package, sticker, brand, uberlyfe, type...

4) Analyse textuelle

Nettoyage des données - exemple

1) Données avant nettoyage

Id	Description	Product_specifications
0	Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 This curtain enhances the look of the interiors.This curtain is made from 100% high quality polyester fabric.It features an eyelet style stitch with Metal Ring.It makes the room environment romantic and loving.This curtain is ant-wrinkle and anti shrinkage and have elegant apparance.Give your home a bright and modernistic appeal with these designs. The surreal attention is sure to steal hearts. These contemporary eyelet and valance curtains slide smoothly so when you draw them apart first thing in the morning to welcome the bright sun rays you want to wish good morning to the whole world and when you draw them close in the evening, you create the most special moments of joyous beauty given by the soothing prints. Bring home the elegant curtain that softly filters light in your room so that you get the right amount of sunlight.,Specifications of Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) General Brand Elegance Designed For Door Type Eyelet Model Name Abstract Polyester Door Curtain Set Of 2 Model ID Duster25 Color Multicolor Dimensions Length 213 cm In the Box Number of Contents in Sales Package Pack of 2 Sales Package 2 Curtains Body & Design Material Polyester	{"product_specification"=>[{"key"=>"Brand", "value"=>"Elegance"}, {"key"=>"Designed For", "value"=>"Door"}, {"key"=>"Type", "value"=>"Eyelet"}, {"key"=>"Model Name", "value"=>"Abstract Polyester Door Curtain Set Of 2"}, {"key"=>"Model ID", "value"=>"Duster25"}, {"key"=>"Color", "value"=>"Multicolor"}, {"key"=>"Length", "value"=>"213 cm"}, {"key"=>"Number of Contents in Sales Package", "value"=>"Pack of 2"}, {"key"=>"Sales Package", "value"=>"2 Curtains"}, {"key"=>"Material", "value"=>"Polyester"}]}1 7b72c92c2f6c40268628ec5f14c6d590 2016-04-30 03:22:56 +0000 http://www.flipkart.com/sathiyas-cotton-bath-towel/p/itmegfzhxbyucwgn?pid=BTWEGFZHGBXPHZUH Sathiyas Cotton Bath Towel ["Baby Care >> Baby Bath & Skin >> Baby Bath Towels >> Sathiyas Baby Bath Towels >> Sathiyas Cotton Bath Towel (3 Bath Towel, Red, Y..."] BTWEGFZHGBXPHZUH 600.0 449.0 7b72c92c2f6c40268628ec5f14c6d590.jpg False Specifications of Sathiyas Cotton Bath Towel (3 Bath Towel, Red, Yellow, Blue) Bath Towel Features Machine Washable Yes Material Cotton Design Self Design General Brand Sathiyas Type Bath Towel GSM 500 Model Name Sathiyas cotton bath towel Ideal For Men, Women, Boys, Girls Model ID asvtwl322 Color Red, Yellow, Blue Size Mediam Dimensions Length 30 inch Width 60 inch In the Box Number of Contents in Sales Package 3 Sales Package 3 Bath Towel No rating available No rating available Sathiyas {"product_specification"=>[{"key"=>"Machine Washable", "value"=>"Yes"}, {"key"=>"Material", "value"=>"Cotton"}, {"key"=>"Design", "value"=>"Self Design"}, {"key"=>"Brand", "value"=>"Sathiyas"}, {"key"=>"Type", "value"=>"Bath Towel"}, {"key"=>"GSM", "value"=>"500"}, {"key"=>"Model Name", "value"=>"Sathiyas cotton bath towel"}, {"key"=>"Ideal For", "value"=>"Men, Women, Boys, Girls"}, {"key"=>"Model ID", "value"=>"asvtwl322"}, {"key"=>"Color", "value"=>"Red, Yellow, Blue"}, {"key"=>"Size", "value"=>"Mediam"}, {"key"=>"Length", "value"=>"30 inch"}, {"key"=>"Width", "value"=>"60 inch"}, {"key"=>"Number of Contents in Sales Package", "value"=>"3"}, {"key"=>"Sales Package", "value"=>"3 Bath Towel"}]}2 64d5d4a258243731dc7bbb1eef49ad74 2016-04-30 03:22:56 +0000 http://www.flipkart.com/eurospa-cotton-terry-face-towel-set/p/itmeg6shbrpubhca?pid=BTWEG6SHXTDB2A2Y Eurospa Cotton Terry Face Towel Set["Baby Care >> Baby Bath & Skin >> Baby Bath Towels >> Eurospa Baby Bath Towels >> Eurospa Cotton Terry Face Towel Set (20 PIECE FA..."] .....
....		

2) Données après nettoyage

Id	clean_concat
0	elegance brand elegance designed door type eyelet model name abstract polyester door curtain model duster color multicolor length number content sale package pack sale package curtain material polyester feature elegance polyester multicolor abstract eyelet door curtain floral curtainelegance polyester multicolor abstract eyelet door curtain height pack price curtain enhances look interiorsthis curtain made high quality polyester fabricit feature eyelet style stitch metal ringit make room environment romantic lovingthis curtain wrinkle anti shrinkage elegant apparancegive home bright modernistic appeal design surreal attention sure steal heart contemporary eyelet valance curtain slide smoothly draw apart first thing morning welcome bright want wish good morning whole world draw close evening create special moment joyous beauty given soothing print bring home elegant curtain softly filter light room right amount sunlightspecifications elegance polyester multicolor abstract eyelet door curtain height pack general brand elegance designed door type eyelet model name abstract polyester door curtain model duster color multicolor dimension length number content sale package pack sale package curtain body design material polyester
....	

4) Analyse textuelle

Bag of words – Résultats bruts

Principe : chaque mot est compté

	aapno	aari	aarika	abgn	abil	abklgrngrngrn	abklplplpnk	abklplpnkpnk	abl	abod	abras	abroad	absolut	absorb	abstract	abstrct	accent	acc
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1045	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1046	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1047	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1048	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1049	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

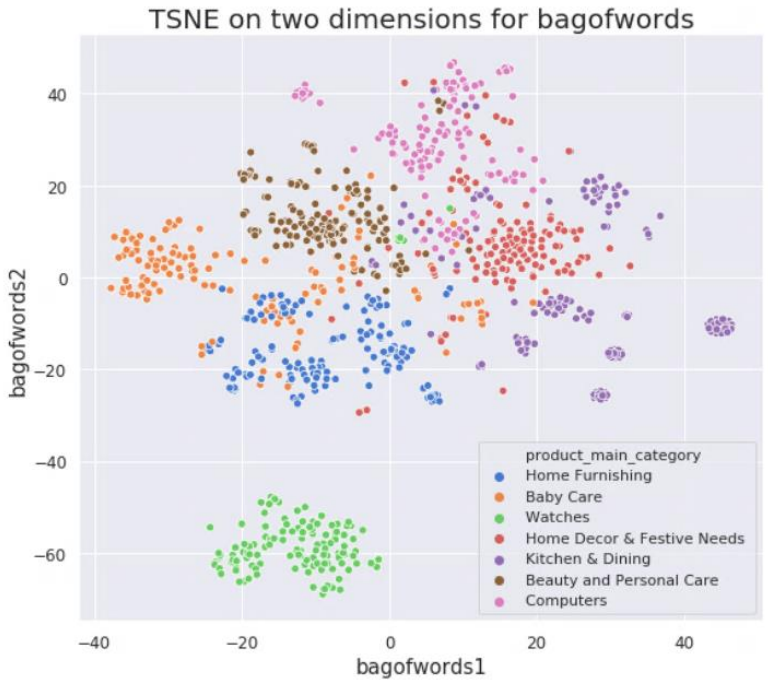
1050 rows × 4701 columns

Bag of words

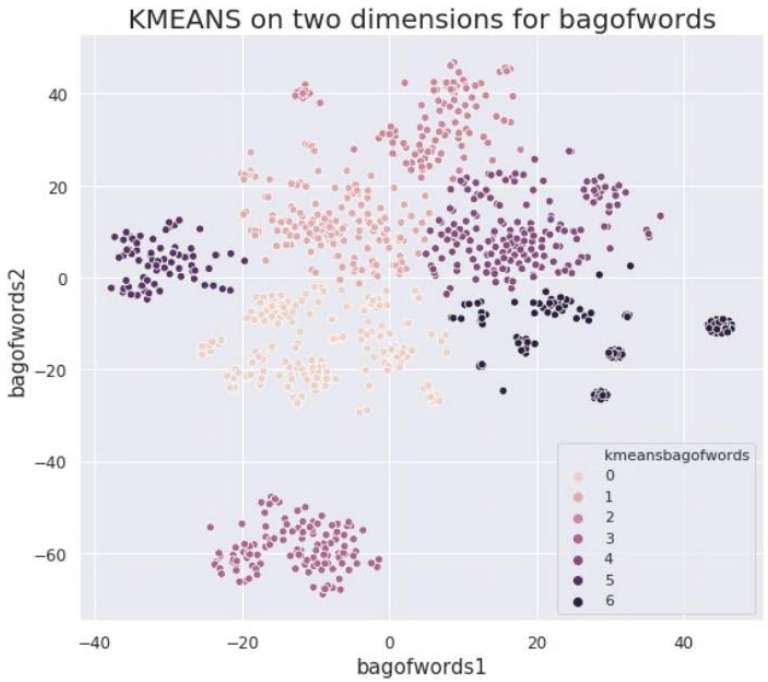
4) Analyse textuelle

Bag of words – Présentation graphique des résultats

```
silhouette_score 0.2747385
calinski_harabasz_score 671.758499071198
```



Etape 1 : Application d'une TSNE à 2 dimensions sur le résultat du bag of words afin de pouvoir afficher les résultats par catégorie



Etape 2 : Application d'un KMEANS sur le résultat de la TSNE afin de verifier visuellement si la formation des clusters est la même que pour les categories de produits

4) Analyse textuelle

TF-IDF– Résultats bruts

Principe : diminution des termes souvent employés et augmentation des termes peu utilisés

	aapno	aari	aarika	abgn	abil	abklgrngrngrn	abklplplpnk	abklplpnkpnk	abl	abod	abras	abroad	absolut	absorb	abstract	abstrct	accent	a
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.031646	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.006135	0.000000	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1045	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	
1046	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	
1047	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	
1048	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	
1049	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	

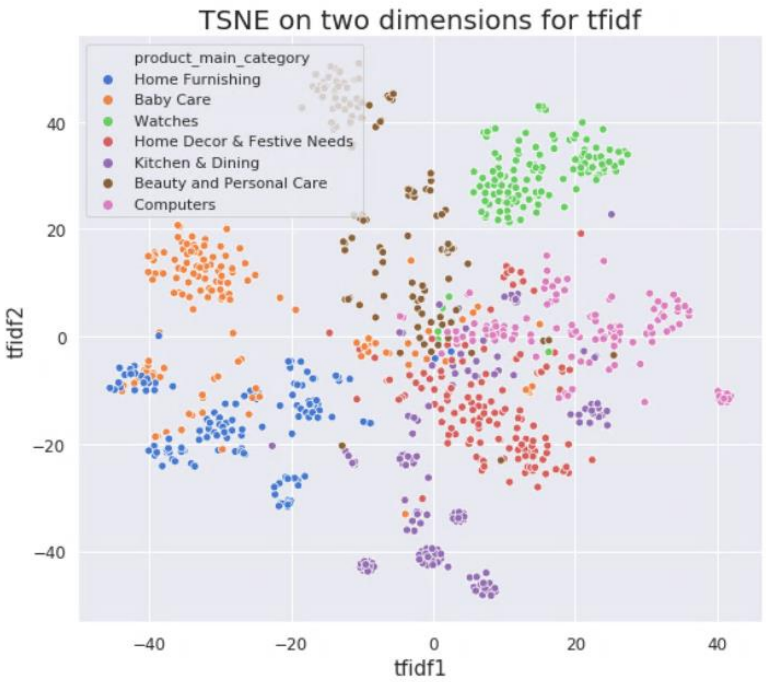
1050 rows × 4701 columns

TF-IDF

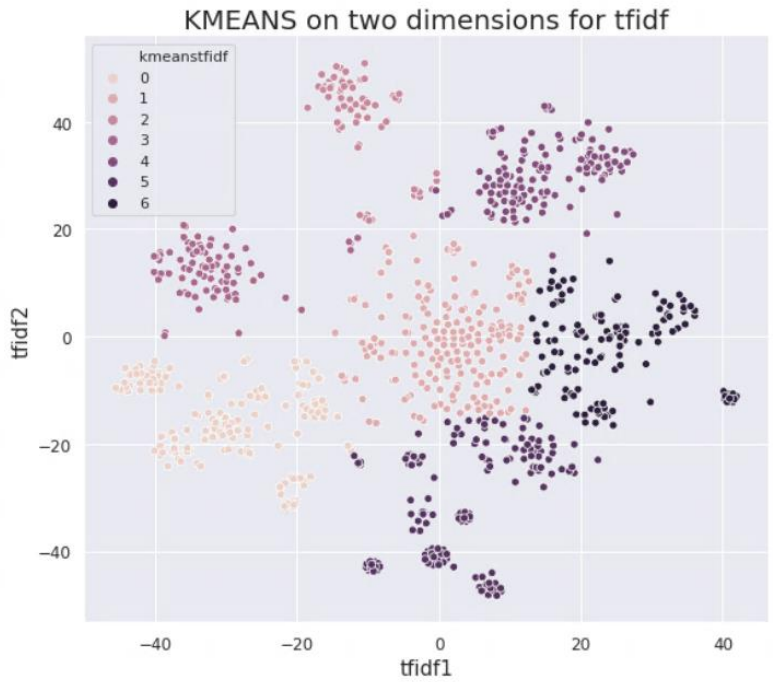
4) Analyse textuelle

TD-IDF– Présentation graphique des résultats

```
silhouette_score 0.24846697
calinski_harabasz_score 514.0654570715894
```



Etape 1 : Application d'une TSNE à 2 dimensions sur le résultat du TFIDF afin de pouvoir afficher les résultats par catégorie

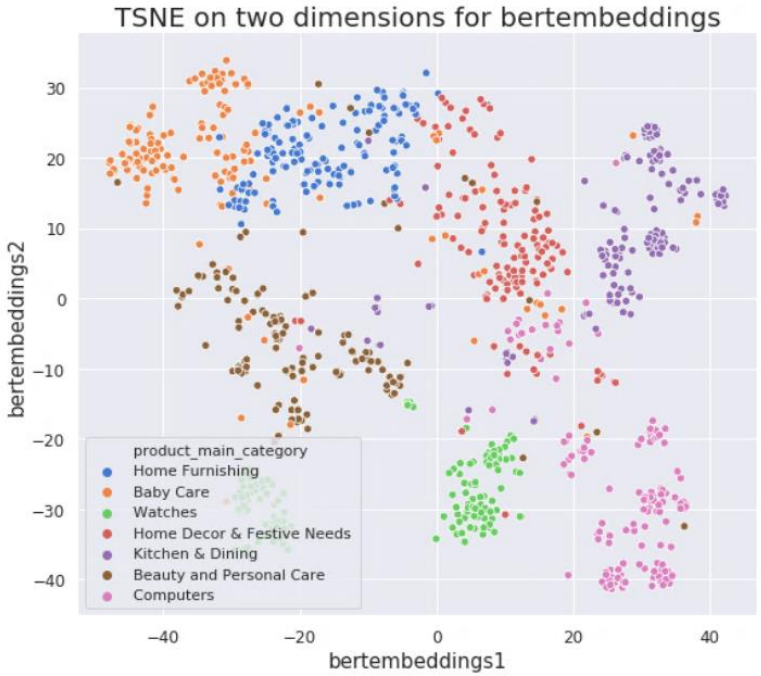


Etape 2 : Application d'un KMEANS sur le résultat de la TSNE afin de verifier visuellement si la formation des clusters est la même que pour les categories de produits

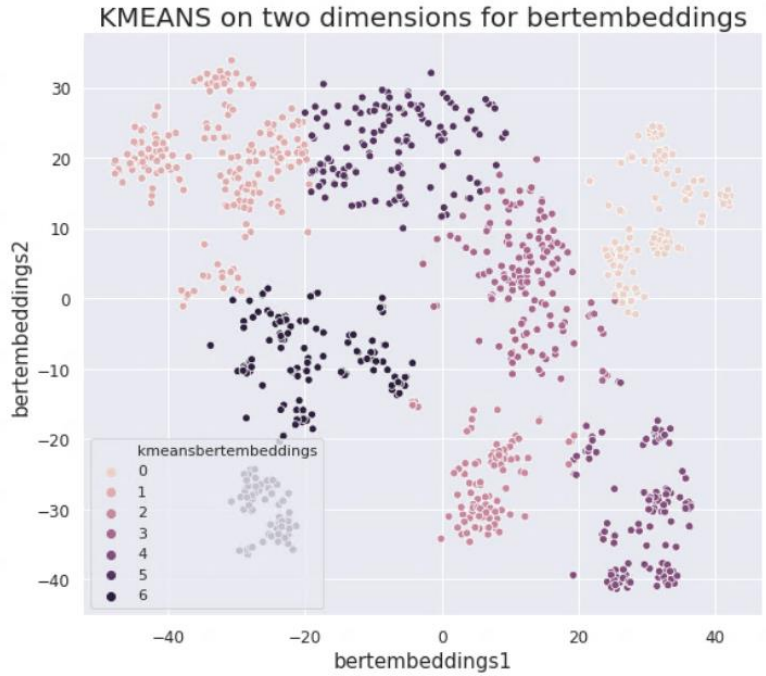
4) Analyse textuelle

BERT-EMBEDDING– Présentation graphique des résultats

```
silhouette_score 0.25909725
calinski_harabasz_score 525.3380042616367
```



Etape 1 : Application d'une TSNE à 2 dimensions sur le résultat du BERT EMBEDDING (modèle pré-entraîné bert-base-nli-mean-tokens) afin de pouvoir afficher les résultats par catégorie

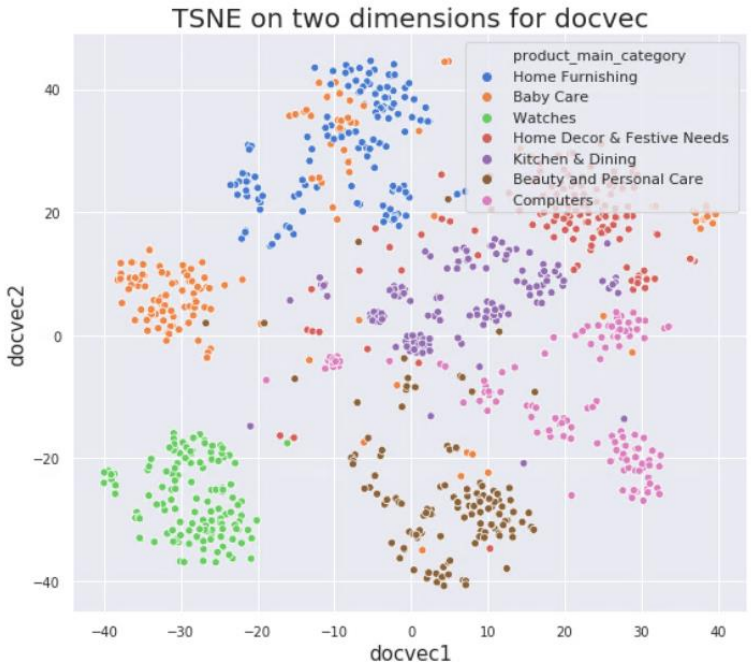


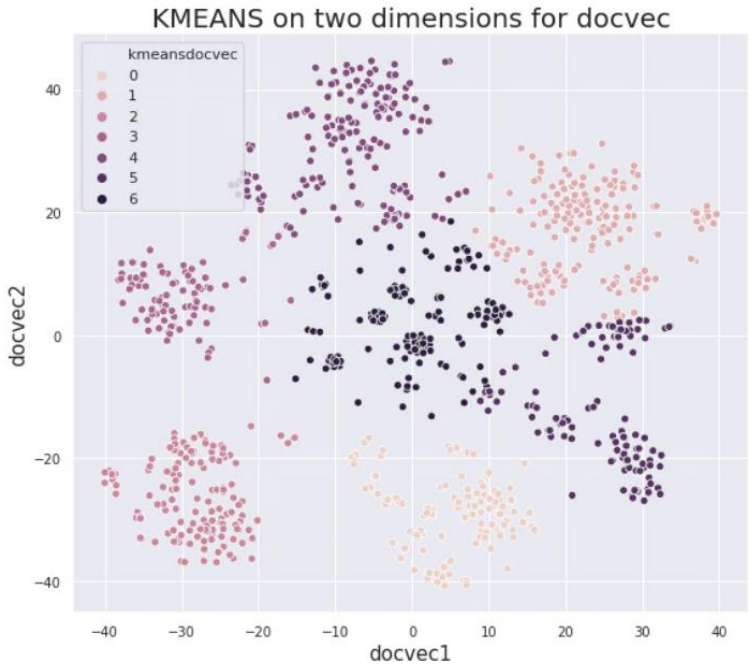
Etape 2 : Application d'un KMEANS sur le résultat de la TSNE afin de verifier visuellement si la formation des clusters est la même que pour les categories de produits

4) Analyse textuelle

DOC2VEC– Présentation graphique des résultats

```
silhouette_score 0.32153472
calinski_harabasz_score 513.9459778352266
```





Etape 1 : Application d'une TSNE à 2 dimensions sur le résultat du DOC2VEC qui a été entrainé sur le jeu de données afin de pouvoir afficher les résultats par catégorie

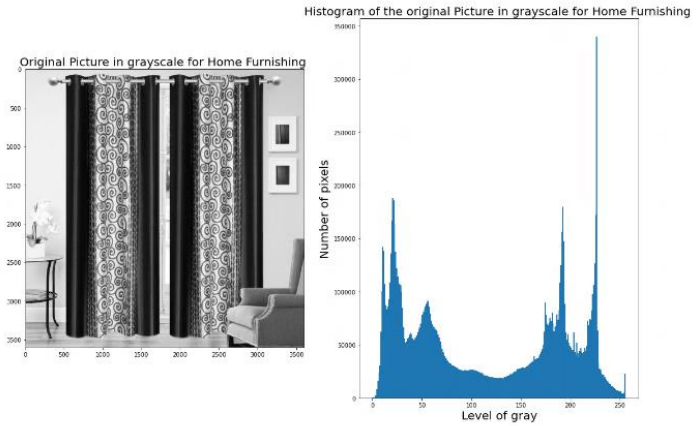
Etape 2 : Application d'un KMEANS sur le résultat de la TSNE afin de verifier visuellement si la formation des clusters est la même que pour les categories de produits

5) Analyse des images

Nettoyage des images

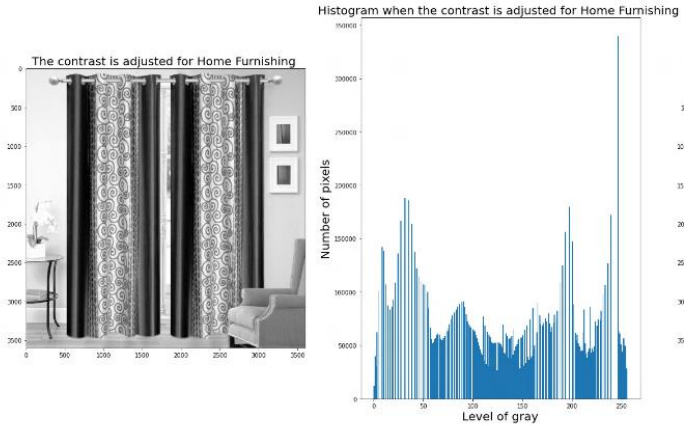
1 Passage en noir et blanc

La luminosité d'un point fournit l'essential de l'information utile. La couleur n'apporte pas grand chose.



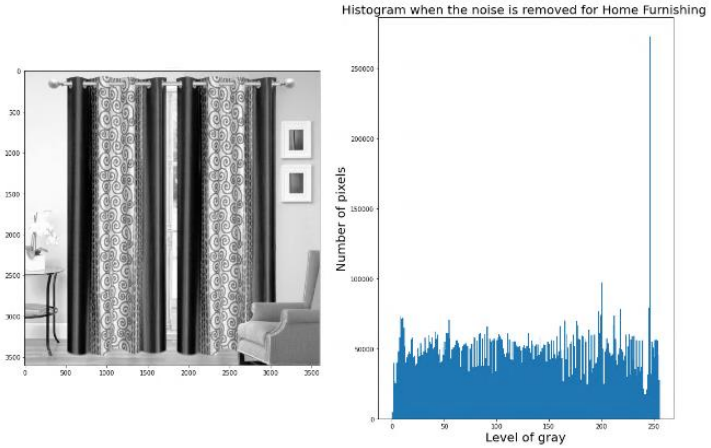
2 Ajustement du contraste

Le contraste est ajusté afin que la distribution des niveaux de gris soit plus uniforme



3 Elimination du bruit

Suppression des "grains" superflus permettant d'améliorer la qualité d'une photo



5) Analyse des images

Sélection de l'algorithme

Sur base de l'étude Image Matching Using SIFT, SURF, and ORB: Performance Comparison for Distorted Images : https://arxiv.org/ftp/arxiv/papers/1710/1710.02726.pdf

Domain	Criteria	SIFT	SURF	ORB
Varying Intensity	Time	↘	↗	↗
	Match Rate	↗	↗	→
Rotated Image	Time	↘	↗	↗
	Match Rate	↗	↗	→
Scaled Image	Time	↘	↗	↗
	Match Rate	→	→	↗
Sheared Image	Time	↘	↗	↗
	Match Rate	↗	↗	→

Légende : ↗ Good → Satisfactory ↘ Poor [ ] Choix

5) Analyse des images

Démarche SIFT

1. Préparation

1.1 Séparation Train / Test
Pour chaque catégorie :
- Train : 150 images
- Test : 50 images

1.2 Redimensionnement des images
Pour chaque catégorie :
- Conservation du rapport hauteur / largeur
- Dimension maximale 250 pixels

2. Extraction des features

2.1 Extraction
Pour chaque image extraction de toutes les features

2.2 Limitation à 150 points centraux par feature
Application d'un KMEANS sur chaque feature

3. Résultats

3.1 Réduction à deux dimensions
Application d'une TSNE

3.2 Prédiction de la classe pour les images de tests via KNN

3.3 Calcul de l'accuracy

Image originale

Image nettoyée

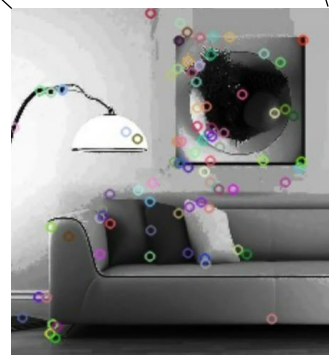
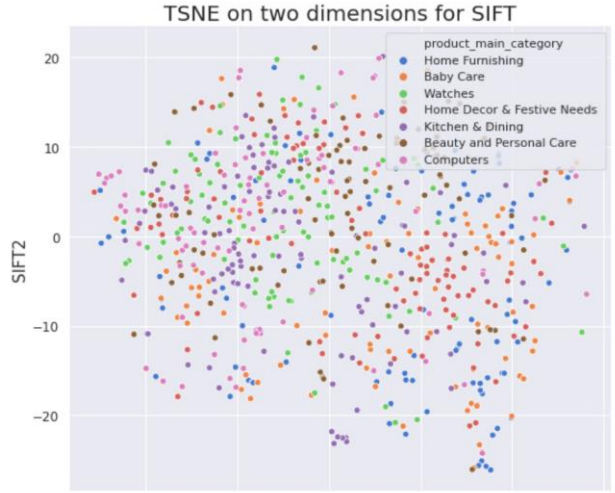


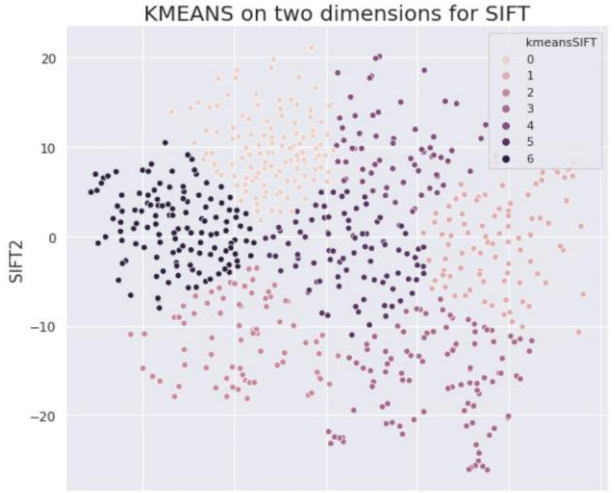
Image avec les features

5) Analyse des images

Résultats SIFT

```
silhouette_score -0.10500022
calinski_harabasz_score 15.487172863023705
```





5) Analyse des images

Démarche CNN avec Transfer Learning

1. Préparation

1.1 Séparation Train / Test / Validation
Pour chaque catégorie :
- Train : 80 images
- Test : 50 images
- Validation : 20 images

1.2 Redimensionnement des images en 224 x 224

2. Entraînement du modèle

2.1 Modèle
MobileNetV2 est choisi car il est le plus petit en taille et offre une accuracy de 70%.
L'hyperparamètre alpha est à 0.35
Pour le Transfer Learning, l'approche "fine-tuning partielle a été choisie (**Entraînement du classifieur et des couches basses)**,

2.2 Optimizer
Adam est choisi car :
- Peu gourmand en mémoire
- Tuning des hyperparamètres simplifié

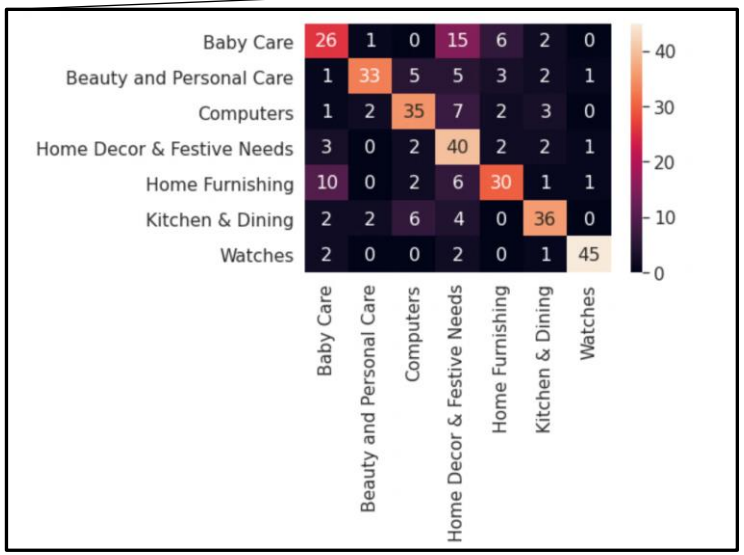
L'hyperparamètre learning_rate est à 0.0001

3. Résultats

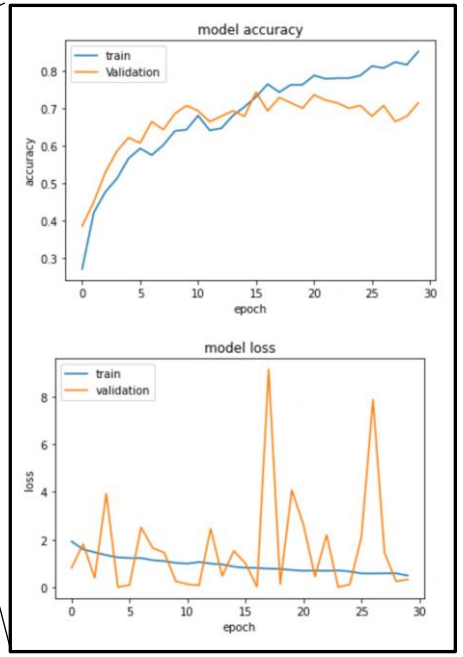
3.1 Accuracy sur le jeu de tests de train et validation

3.2 Accuracy sur le jeu de tests après entrainement : 0.894285

3.3 Matrice de confusion et classification report



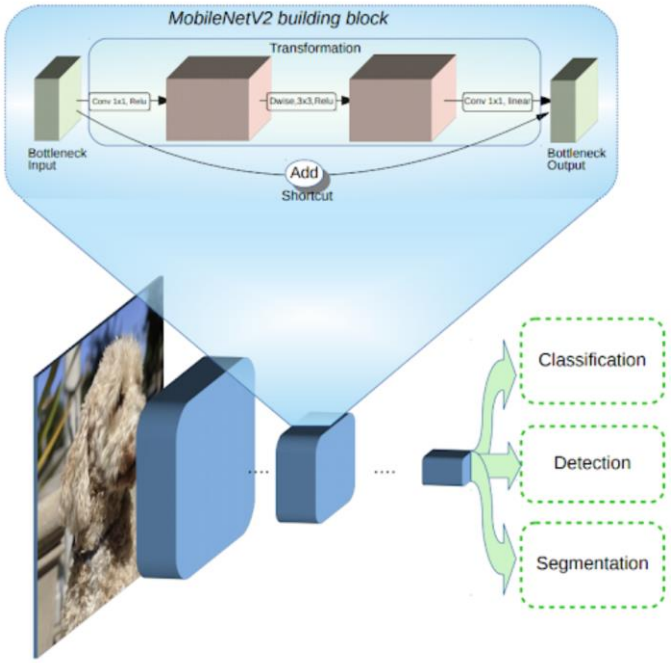
	Precision	Recall	F1-score	Support
Baby Care	0.58	0.52	0.55	50
Beauty and Personal Care	0.87	0.66	0.75	50
Computers	0.70	0.70	0.70	50
Home Décor & Festive Needs	0.51	0.80	0.62	50
Home Furnishing	0.70	0.60	0.65	50
Kitchen & Dining	0.77	0.72	0.74	50
Watches	0.94	0.90	0.92	50
Accuracy			0.70	350
Macro avg	0.72	0.70	0.70	350
Weighted avg	0.72	0.70	0.70	350

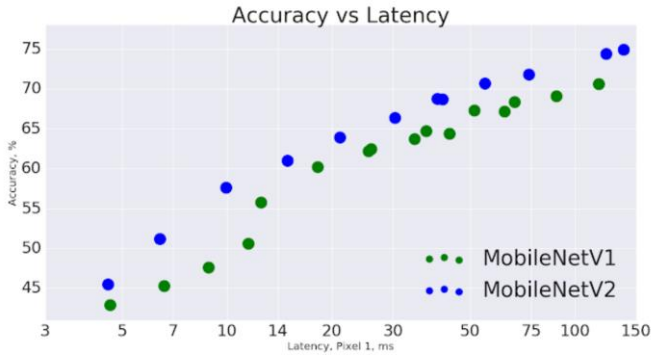


5) Analyse des images

MobileNetV2

Model	Size	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth
Xception	88 MB	0.790	0.945	22,910,480	126
VGG16	528 MB	0.713	0.901	138,357,544	23
VGG19	549 MB	0.713	0.900	143,667,240	26
ResNet50	98 MB	0.749	0.921	25,636,712	-
ResNet101	171 MB	0.764	0.928	44,707,176	-
ResNet152	232 MB	0.766	0.931	60,419,944	-
ResNet50V2	98 MB	0.760	0.930	25,613,800	-
ResNet101V2	171 MB	0.772	0.938	44,675,560	-
ResNet152V2	232 MB	0.780	0.942	60,380,648	-
InceptionV3	92 MB	0.779	0.937	23,851,784	159
InceptionResNetV2	215 MB	0.803	0.953	55,873,736	572
MobileNet	16 MB	0.704	0.895	4,253,864	88
MobileNetV2	14 MB	0.713	0.901	3,538,984	88
DenseNet121	33 MB	0.750	0.923	8,062,504	121
DenseNet169	57 MB	0.762	0.932	14,307,880	169
DenseNet201	80 MB	0.773	0.936	20,242,984	201
NASNetMobile	23 MB	0.744	0.919	5,326,716	-
NASNetLarge	343 MB	0.825	0.960	88,949,818	-
EfficientNetB0	29 MB	-	-	5,330,571	-
EfficientNetB1	31 MB	-	-	7,856,239	-
EfficientNetB2	36 MB	-	-	9,177,569	-
EfficientNetB3	48 MB	-	-	12,320,535	-
EfficientNetB4	75 MB	-	-	19,466,823	-
EfficientNetB5	118 MB	-	-	30,562,527	-
EfficientNetB6	166 MB	-	-	43,265,143	-
EfficientNetB7	256 MB	-	-	66,658,687	-





MobileNetV2 improves speed (reduced latency) and increased ImageNet Top 1 accuracy

5) Analyse des images

Exemples d'erreurs detectées

Catégorie réelle	Baby Care
Catégorie détectée	Home Decor & Festive Needs

Catégorie réelle	Beauty & Personal Care
Catégorie détectée	Computers

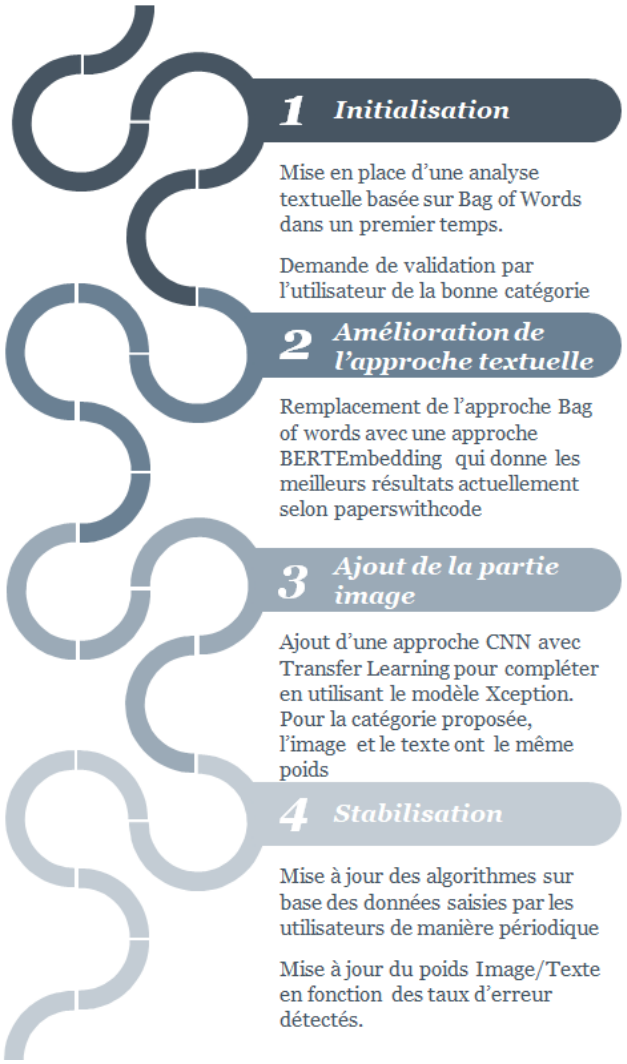
Catégorie réelle	Computers
Catégorie détectée	Home Decor & Festive Needs

Catégorie réelle	Home Furnishing
Catégorie détectée	Baby Care

Catégorie réelle	Kitchen & Dining
Catégorie détectée	Computers

Catégorie réelle	Watches
Catégorie détectée	Baby Care

6) Prochaines étapes



7) Environnement technique