Place de marché – Classification de produits

- Attribution automatique de la catégorie de produits

Préambule

- Place de marché
  - Marketplace e-commerce
  - Attribution de catégorie manuelle
  - Peu fiable
- Objectif
  - Automatiser l’attribution de catégorie
  - Amélioration de l’expérience utilisateur

Sommaire

- Présentation du jeu de données
- Exploration
- Prétraitement des données
- Extraction de features
- Classification supervisée
- API

Présentation des données

- 1050 produits
  - Description / catégorie / images sans valeurs manquantes
  - Informations annexes inutiles

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1050 entries, 0 to 1049
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   uniq_id                 1050 non-null   object
 1   crawl_timestamp         1050 non-null   object
 2   product_url             1050 non-null   object
 3   product_name            1050 non-null   object
 4   product_category_tree   1050 non-null   object
 5   pid                     1050 non-null   object
 6   retail_price            1049 non-null   float64
 7   discounted_price        1049 non-null   float64
 8   image                   1050 non-null   object
 9   is_FK_Advantage_product 1050 non-null   bool
 10  description             1050 non-null   object
 11  product_rating          1050 non-null   object
 12  overall_rating          1050 non-null   object
 13  brand                   712 non-null    object
 14  product_specifications  1049 non-null   object
dtypes: bool(1), float64(2), object(12)
```

```
# Print the number of unique values by column
df.nunique()
```

```
uniq_id                   1050
crawl_timestamp            149
product_url               1050
product_name              1050
product_category_tree      642
pid                       1050
retail_price               354
discounted_price           424
image                     1050
is_FK_Advantage_product      2
description               1050
product_rating              27
overall_rating              27
brand                      490
product_specifications     984
dtype: int64
```

Présentation des données

- Catégories
  - Arborescence
  - Nombre de sous catégories disparate

	1	2	3	4	5	6	7
0	Home Furnishing	Curtains & Accessories	Curtains	Elegance Polyester Multicolor Abstract Eyelet...	None	None	None
1	Baby Care	Baby Bath & Skin	Baby Bath Towels	Sathiyas Baby Bath Towels	Sathiyas Cotton Bath Towel (3 Bath Towel, Red...	None	None
2	Baby Care	Baby Bath & Skin	Baby Bath Towels	Eurospa Baby Bath Towels	Eurospa Cotton Terry Face Towel Set (20 PIECE...	None	None
3	Home Furnishing	Bed Linen	Bedsheets	SANTOSH ROYAL FASHION Bedsheets	SANTOSH ROYAL FASHION Cotton Printed King siz...	None	None
4	Home Furnishing	Bed Linen	Bedsheets	Jaipur Print Bedsheets	Jaipur Print Cotton Floral King sized Double ...	None	None
...	...	...	...	...	...	...	...
1045	Baby Care	Baby & Kids Gifts	Stickers	Oren Empower Stickers	None	None	None
1046	Baby Care	Baby & Kids Gifts	Stickers	Wallmantra Stickers	None	None	None
1047	Baby Care	Baby & Kids Gifts	Stickers	Uberlyfe Stickers	None	None	None
1048	Baby Care	Baby & Kids Gifts	Stickers	Wallmantra Stickers	None	None	None
1049	Baby Care	Baby & Kids Gifts	Stickers	Uberlyfe Stickers	None	None	None

	1	2	3	4	5	6	7
count	1050	1050	1047	679	405	127	57
unique	7	63	246	350	297	117	57
top	Home Furnishing	Wrist Watches	Deodorants	Combos	Dresses	Wow! Dresses	Mom and Kid Baby Girl's Printed Green Top & P...
freq	150	149	65	64	21	3	1

Données textuelles

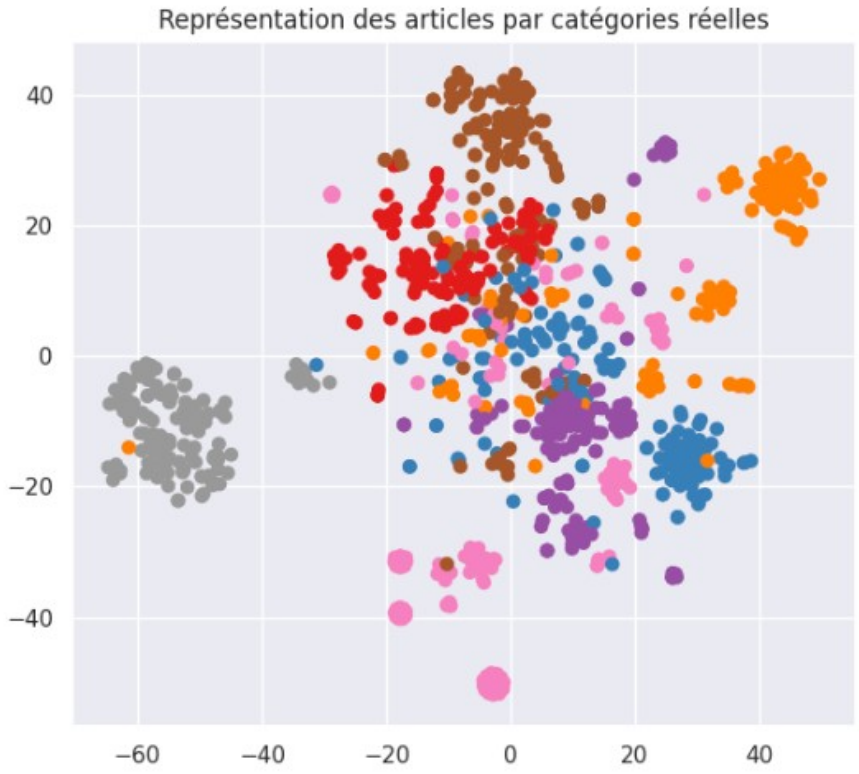
- Préprocessing
  - Lower => passage en minuscule
  - Tokenisation => isolement des mots, stopwords, mots inférieurs à 2 caractères
  - Lemmatisation => forme canonique
  - Stemmatisation => racine

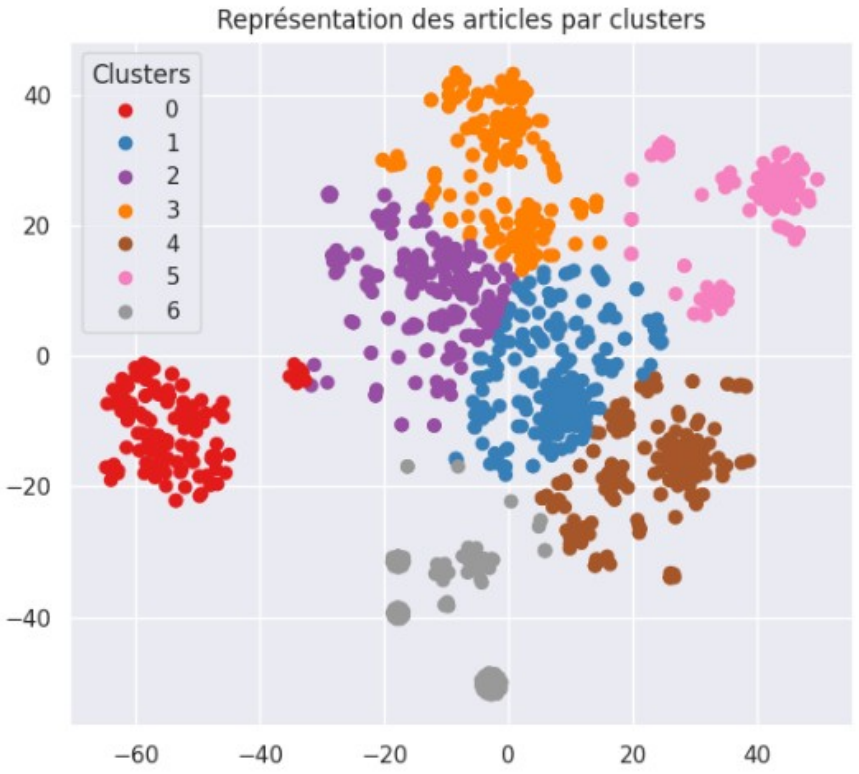
```
---------
Description originale :
nutcase sticker wrap design - teal & pink watercolors 800 ml bottle only for rs 399 . buy online @ flipkart.com. only genuine products. fre
e shipping. cash on delivery!
Description transformée :
nutcase sticker wrap design teal pink watercolor bottle buy online flipkart com genuine product free shipping cash delivery


---------
Description originale :
buy apex rolling pizza cutter for rs.69 online. apex rolling pizza cutter at best prices with free shipping & cash on delivery. only genuin
e products. 30 day replacement guarantee.
Description transformée :
buy apex rolling pizza cutter online apex rolling pizza cutter best price free shipping cash delivery genuine product day replacement guara
ntee
```

Données textuelles

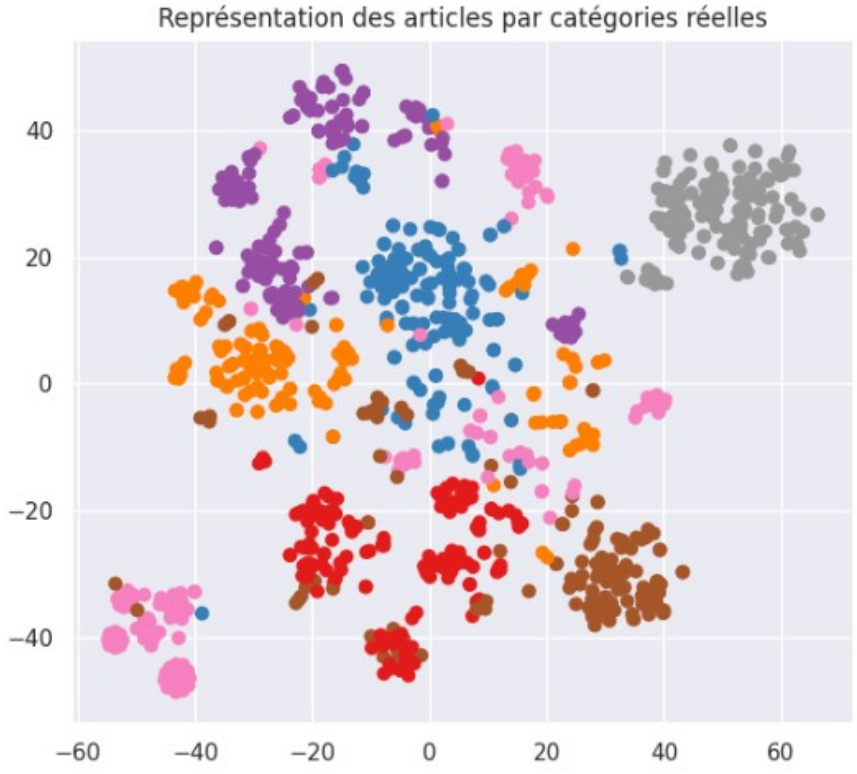
- Bag of words
  - Score ARI : 0.418

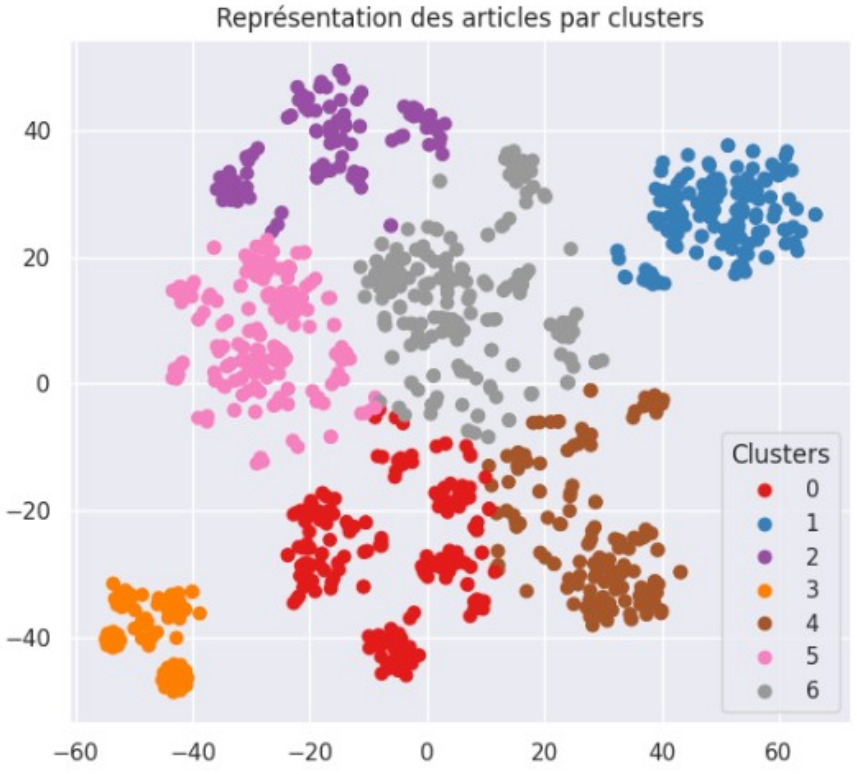




Données textuelles

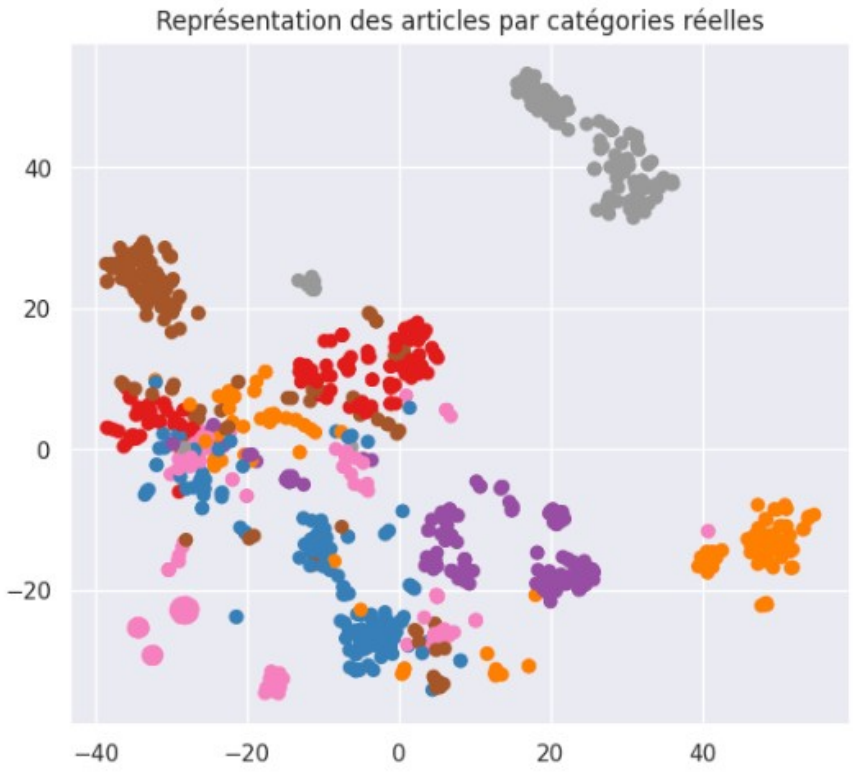
- TF-IDF
  - Score ARI : 0.5068

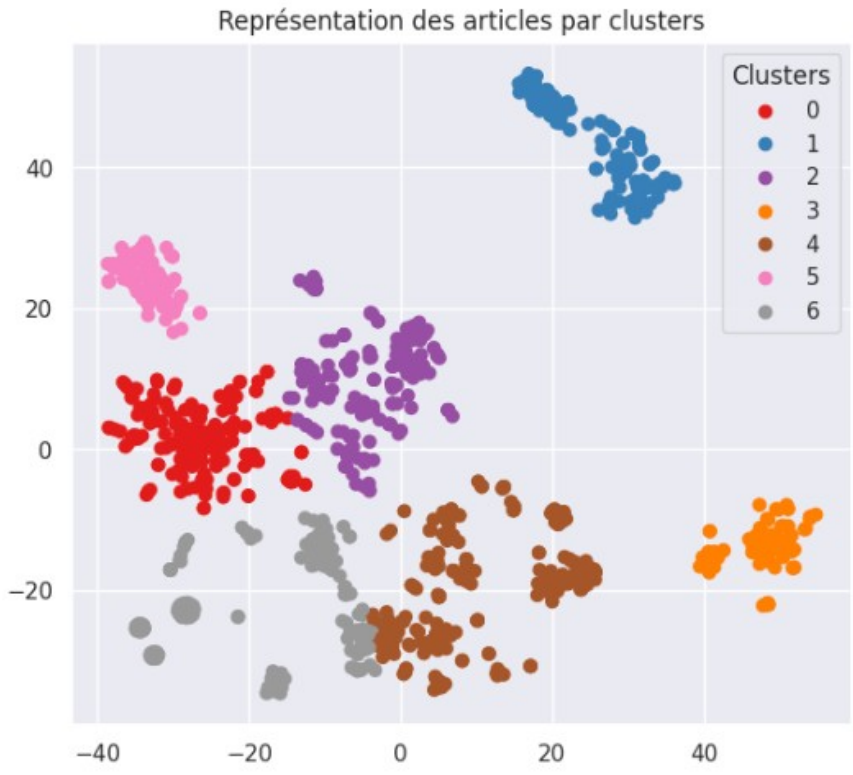




Données textuelles

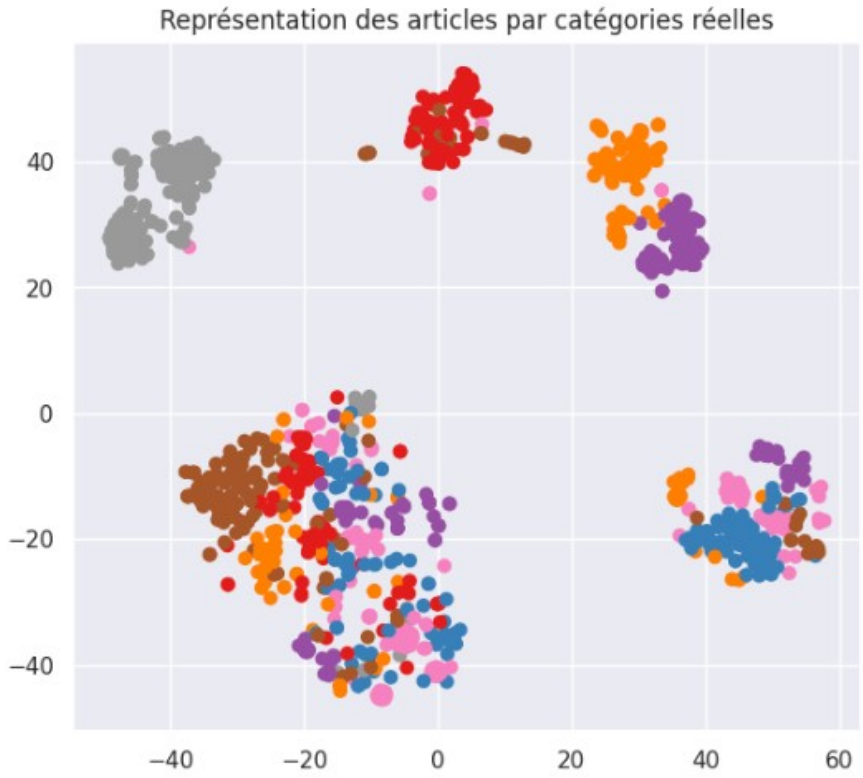
- Word2Vec
  - Score ARI : 0.3935

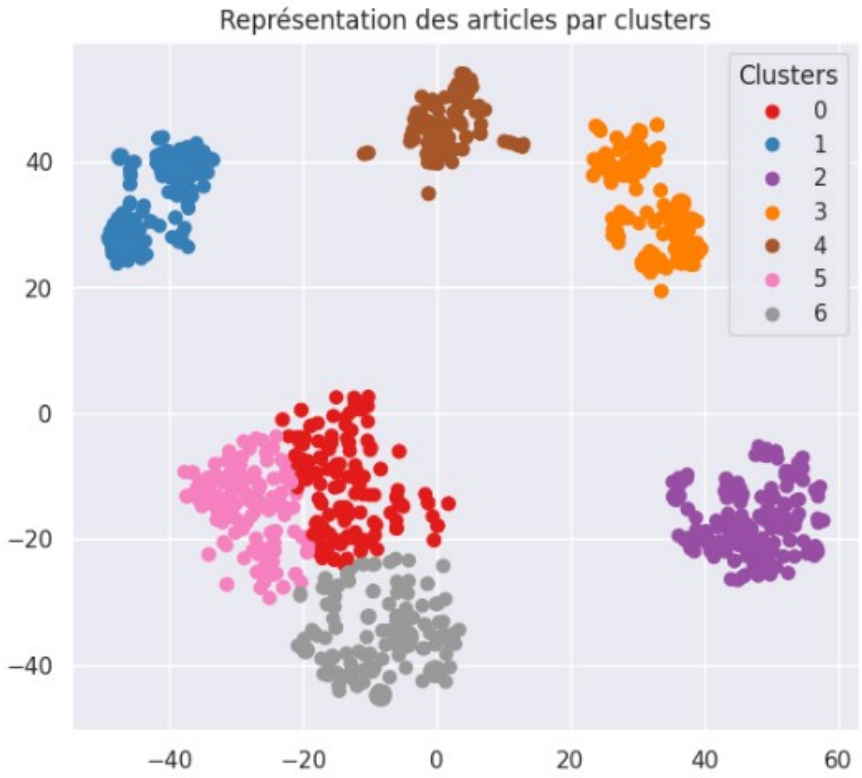




Données textuelles

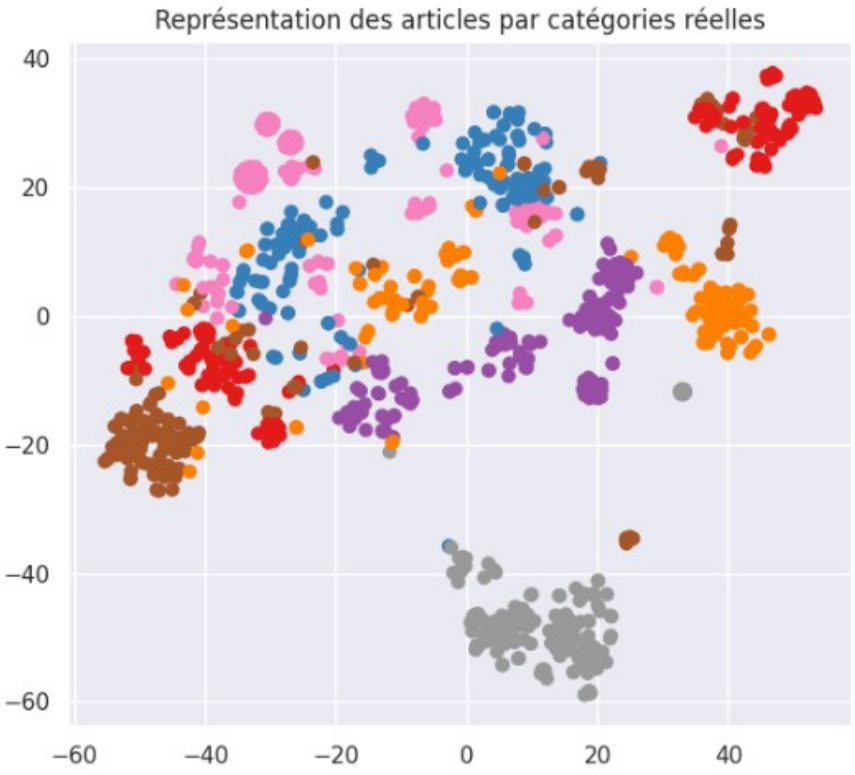
- BERT
  - Score ARI : 0.3127

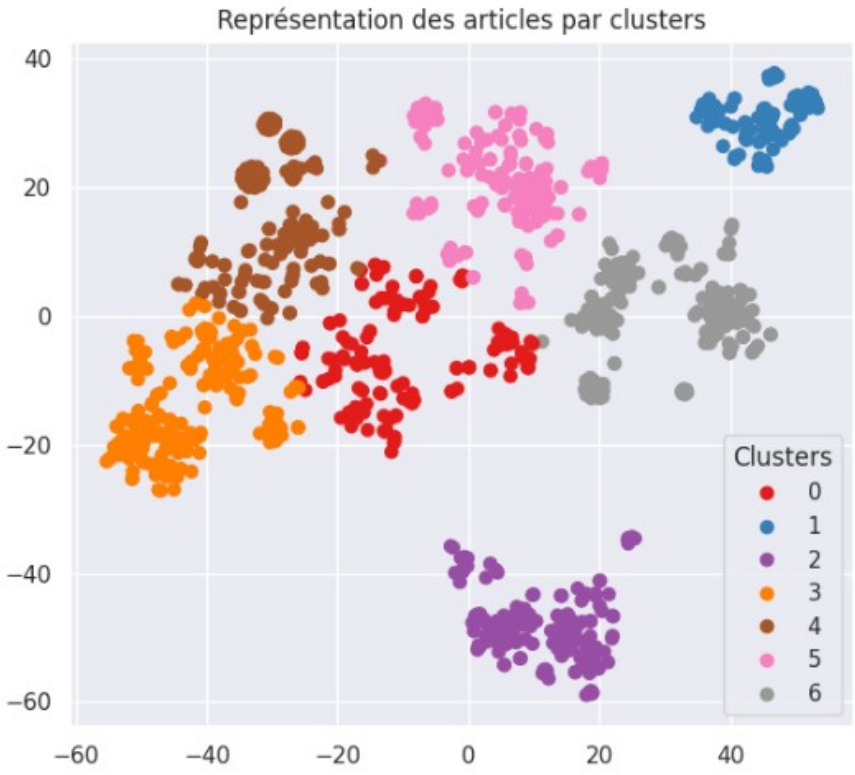




Données textuelles

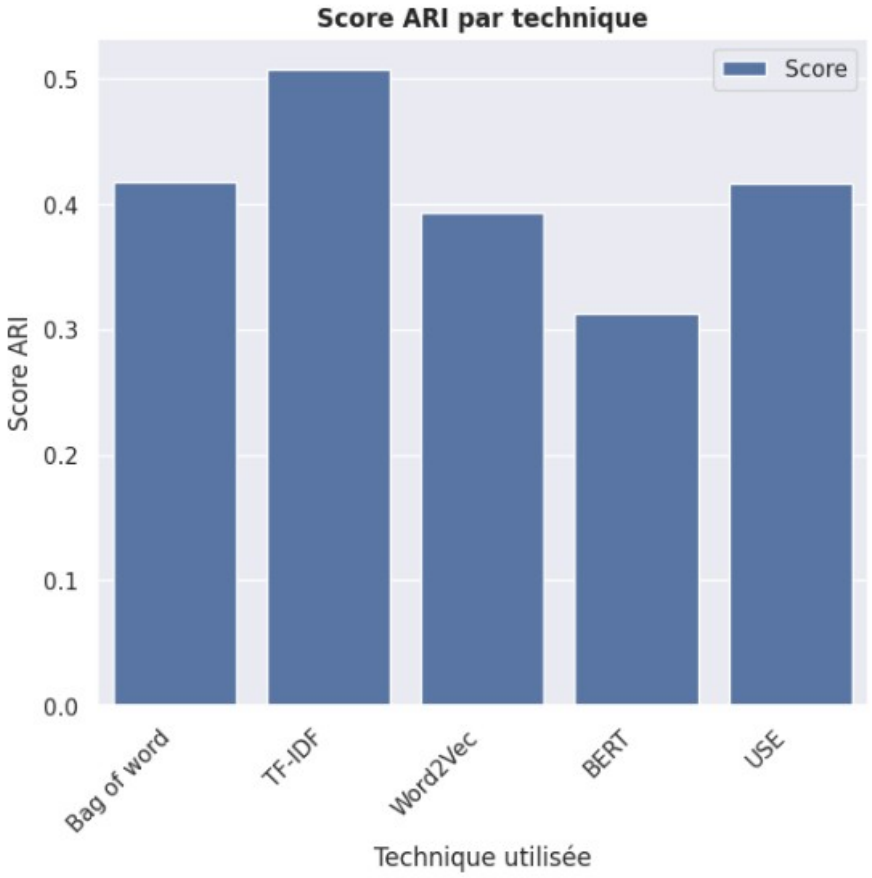
- USE
  - Score ARI : 0.416

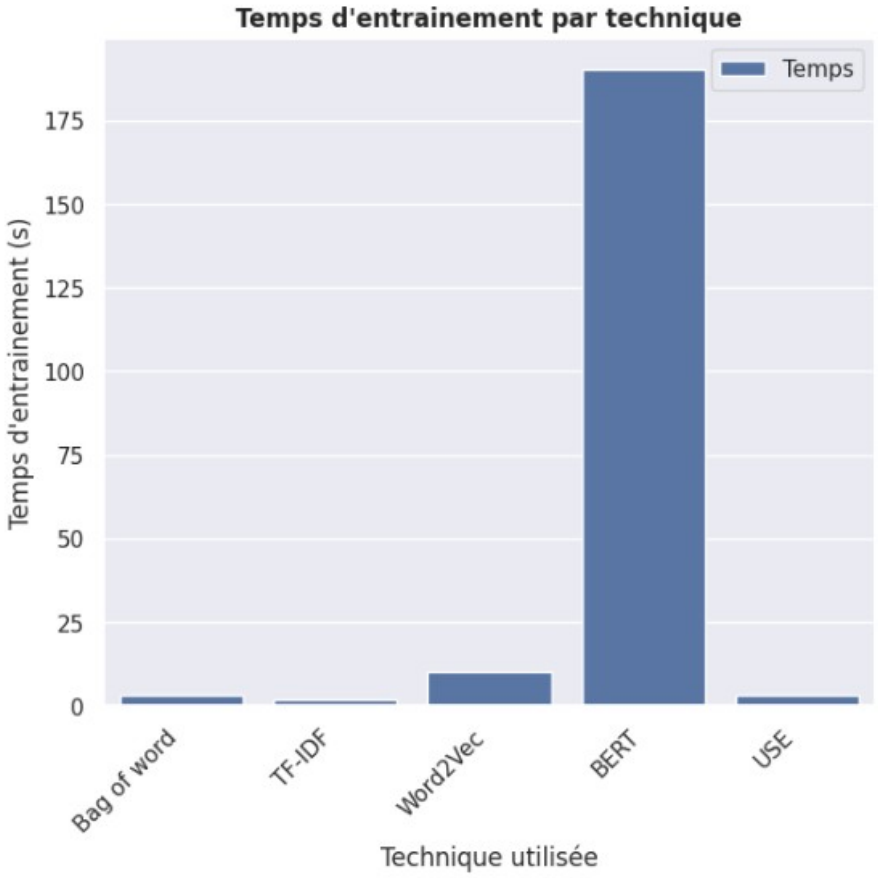




Données textuelles

- Comparaison des résultats
  - Meilleur résultats : TF-IDF





Données images

- Préprocessing
  - Niveau de gris => réduction de dimensions
  - Exposition => correction luminosité
  - Contraste => égalisation histogramme
  - Suppression du bruit

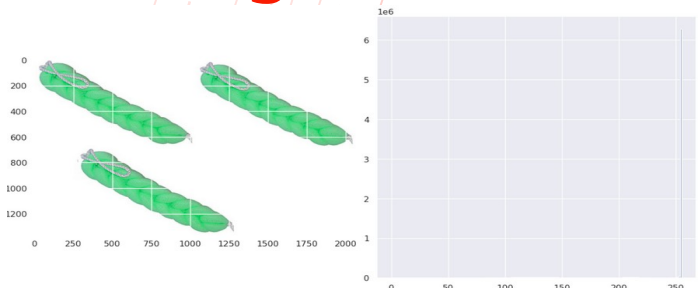
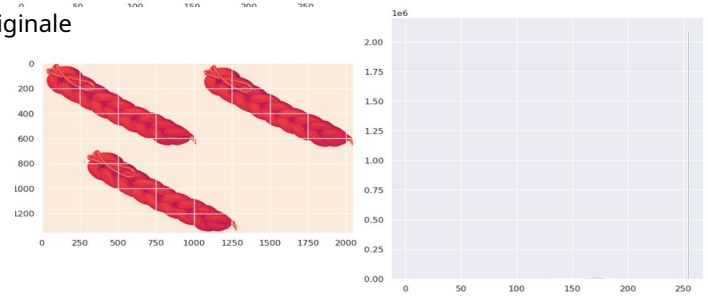
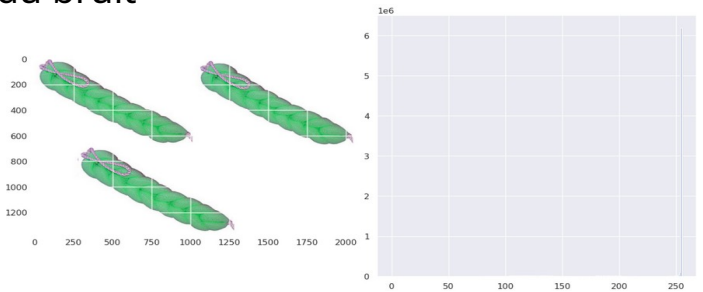


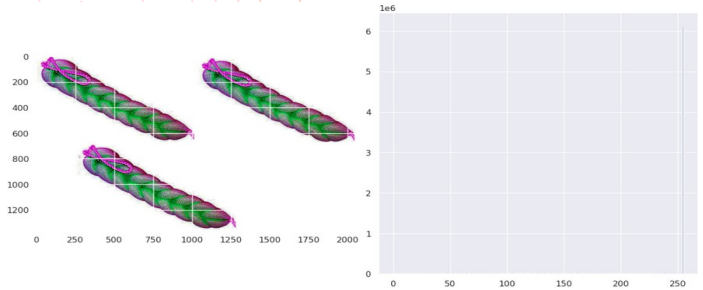
Image originale



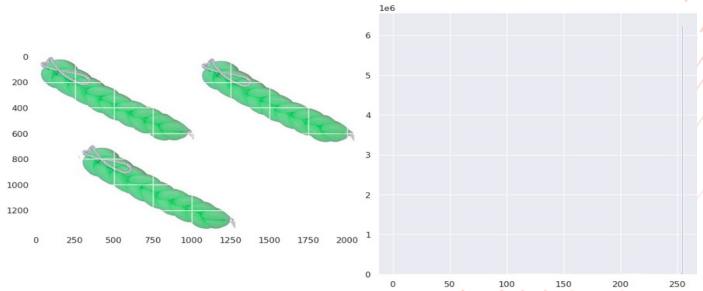
Niveau de gris



Correction de l'exposition



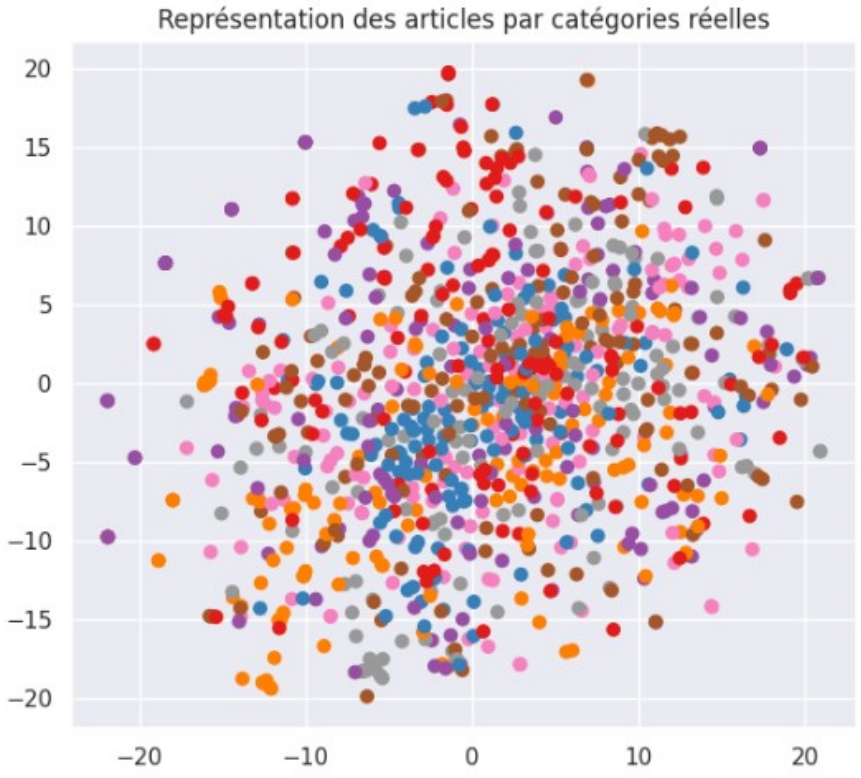
Amélioration du contraste

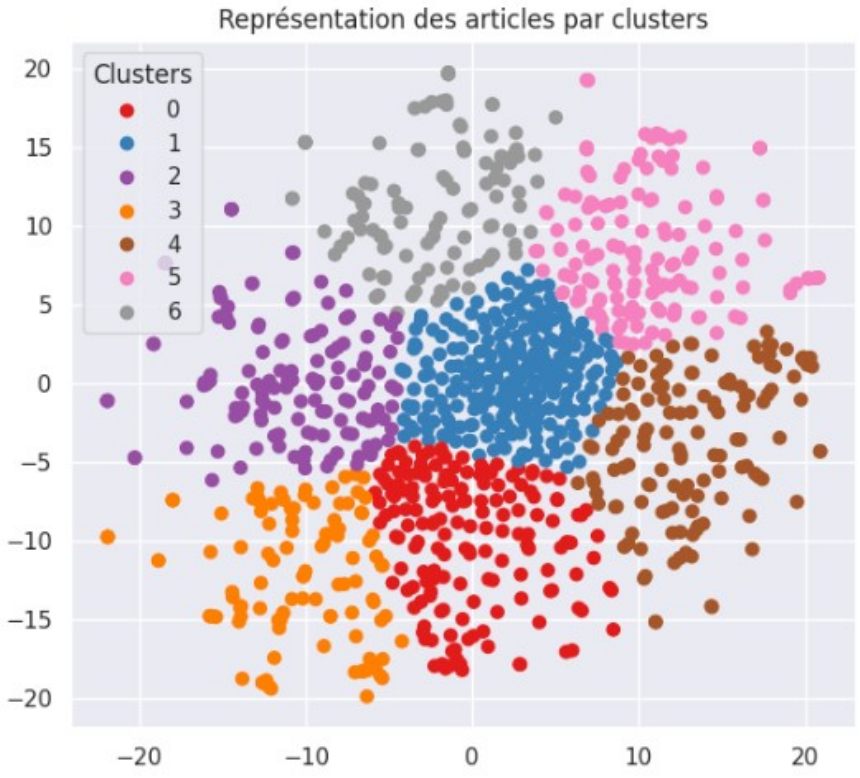


Élimination du brut

Données images

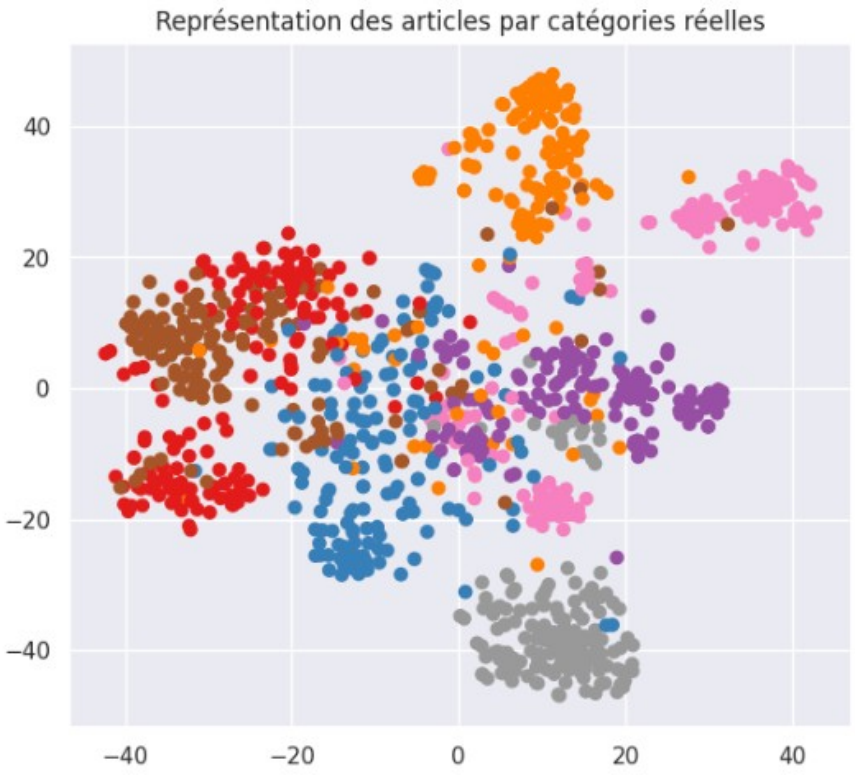
- SIFT
  - Score ARI : 0.0329

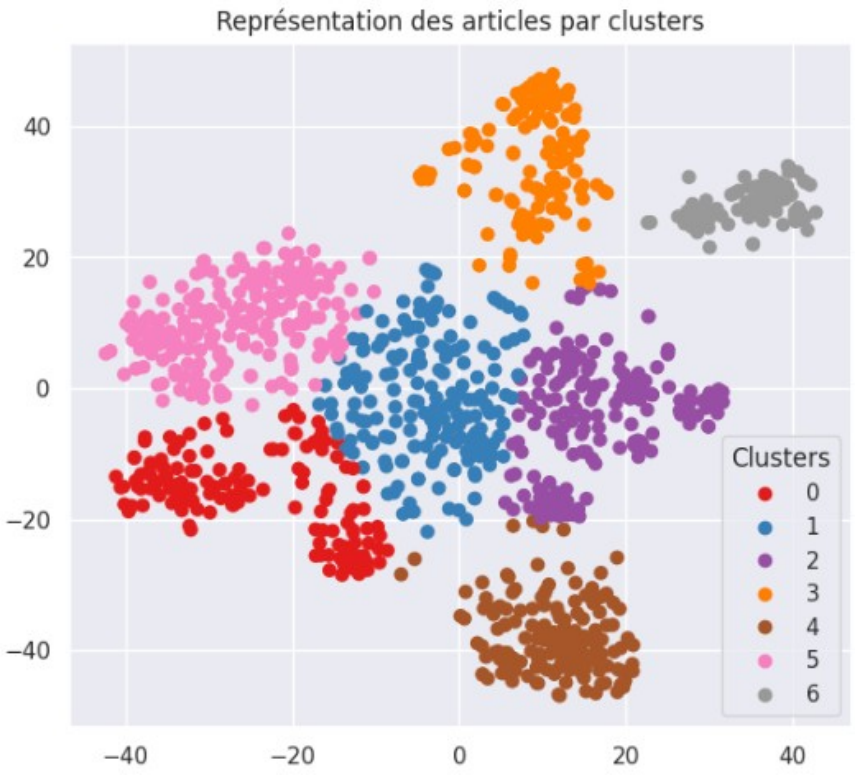




Données images

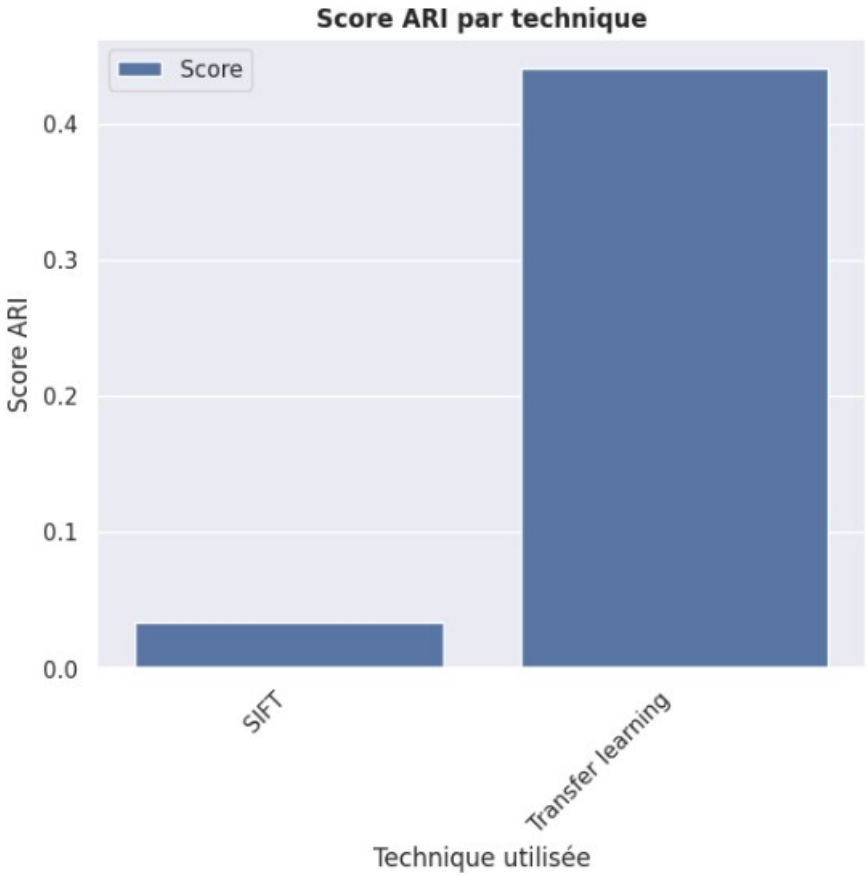
- Transfer learning
  - Score ARI : 0.4405

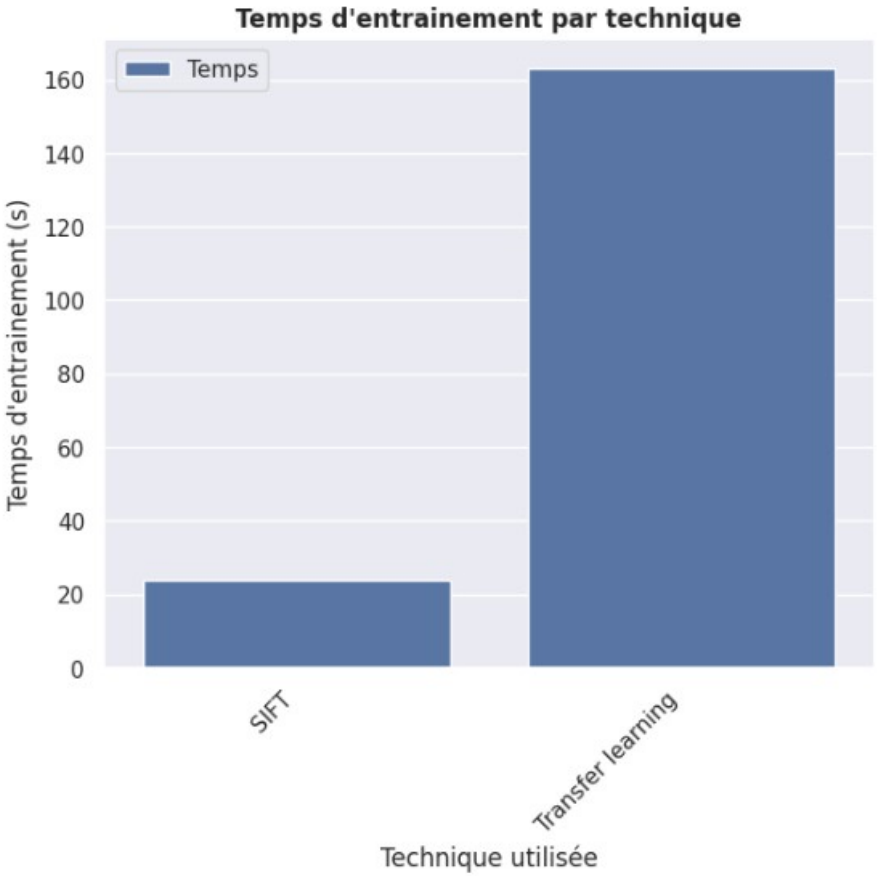




Données images

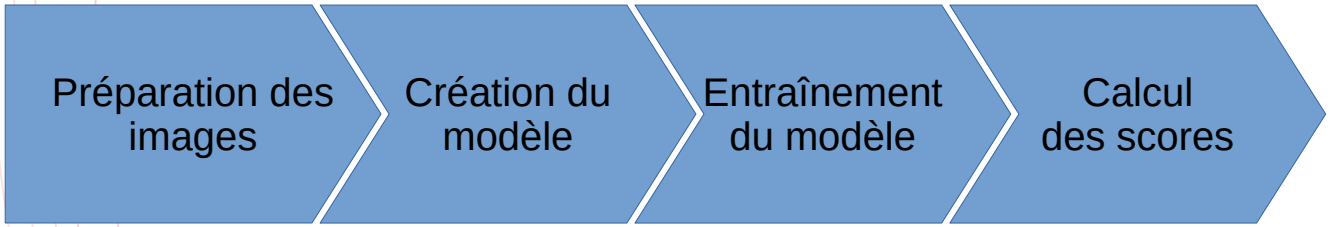
- Comparaison des résultats
  - Meilleur résultats : Transfer learning





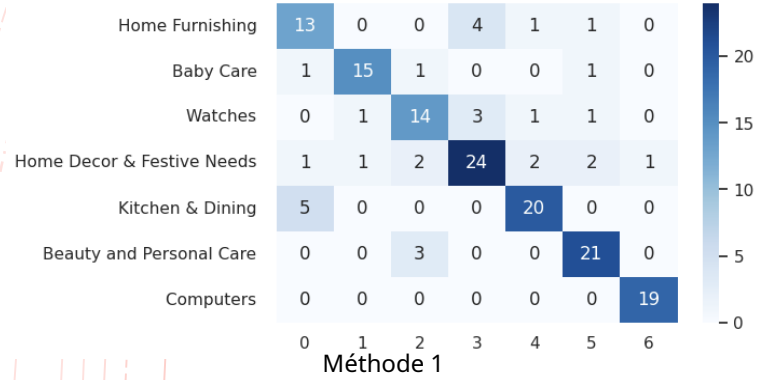
Classification supervisée

- Méthodologie
- Quatre approches :
  - Préprocessing simple
  - Data augmentation
  - Dataset
  - Dataset avec data augmentation intégrée au modèle

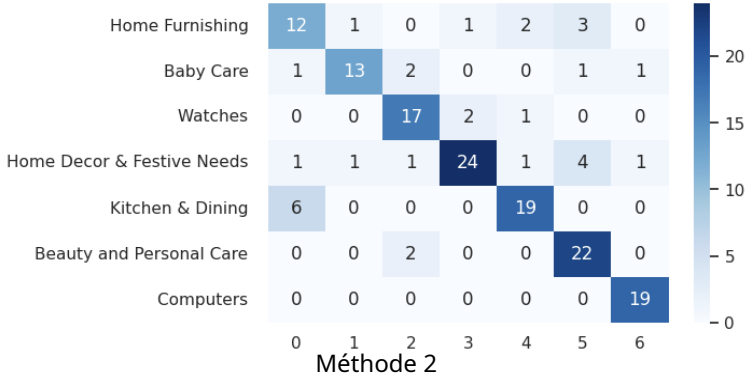


Classification supervisée

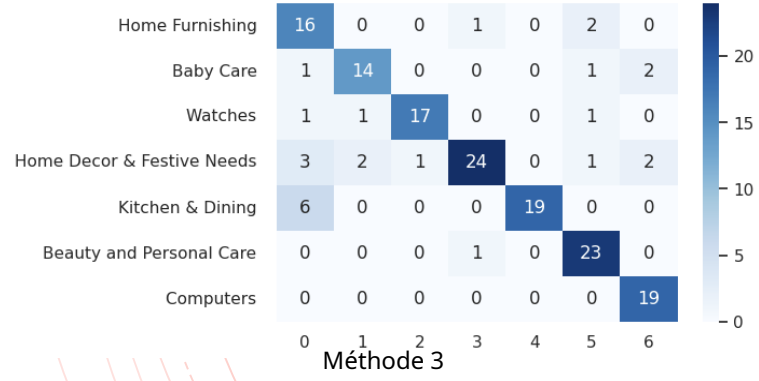
- Matrices de confusion



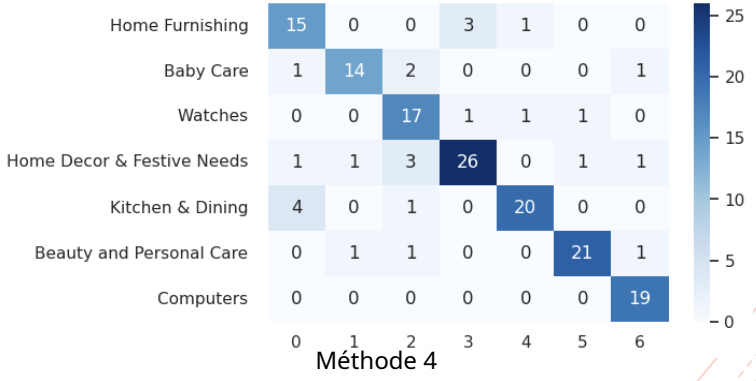
Méthode 1



Méthode 2



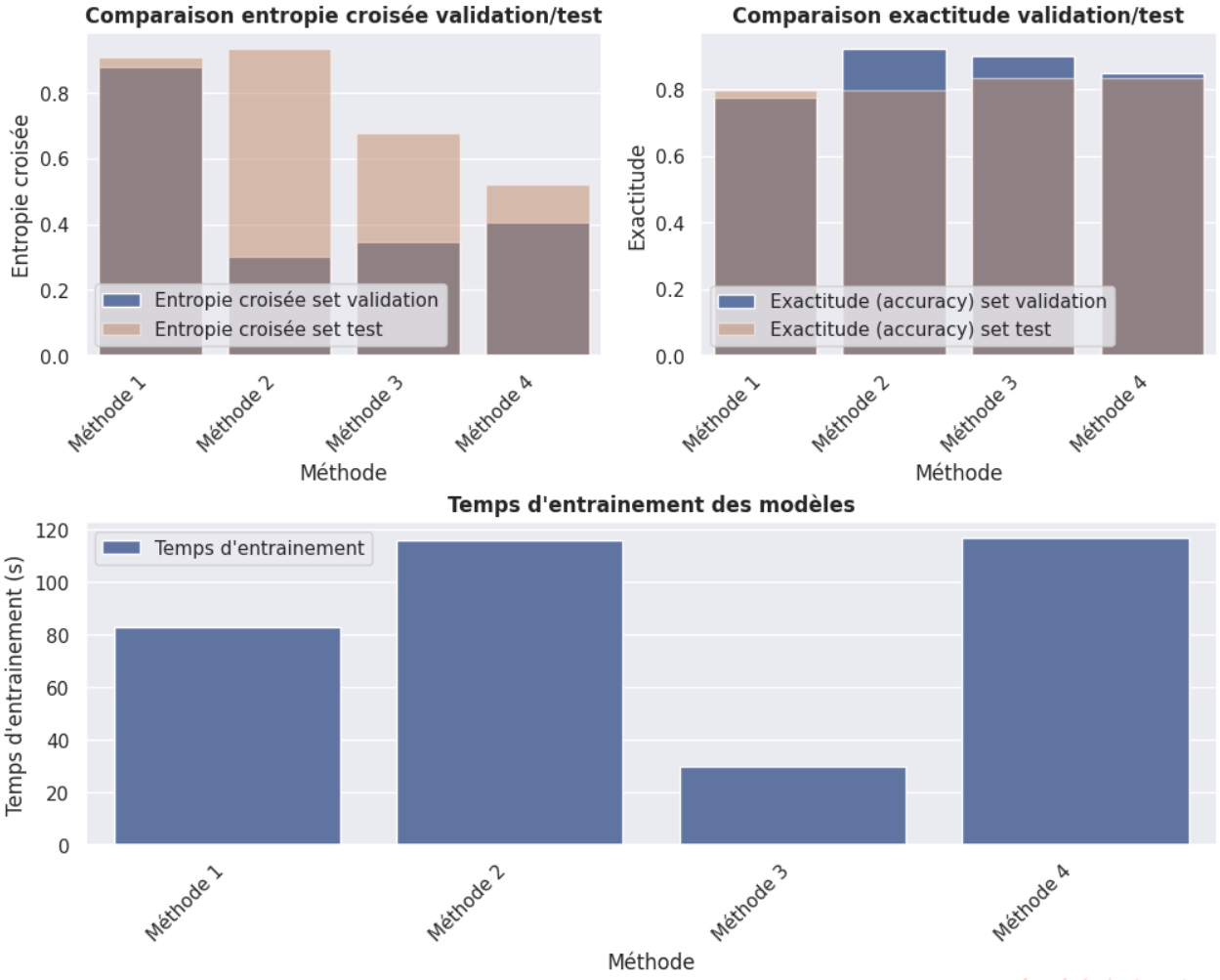
Méthode 3



Méthode 4

Classification supervisée

- Comparaison des résultats



API

- Requête HTTPS GET
- Fonction search
- Header personnalisé
- Paramètres
  - "action": "process"
  - "tagtype_0": "categories"
  - "tag_contains_0": "contains"
  - "tag_0": "champagne"
  - "json": "true"
- Réponse au format json converti en dataframe

Merci

Q & A