# Projet 6 :
# Classifiez automatiquement des biens de consommation

Formation DATA SCIENTIST - Victor BARBIER

# Sommaire

# Contexte et problématique

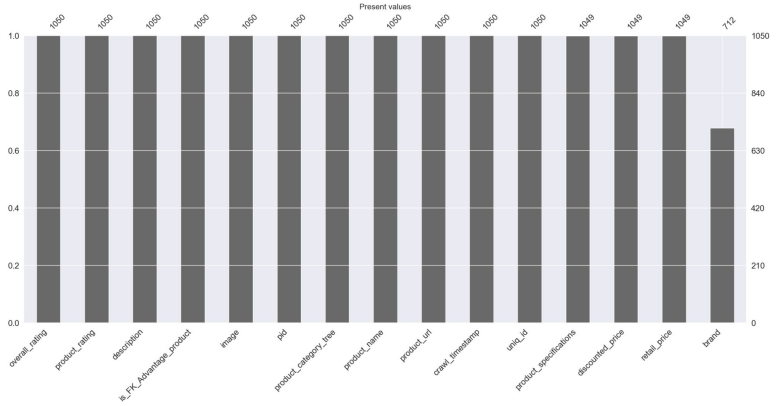
# Contexte et problématique


- Entreprise :
  - Place de Marché
  - Plateforme de e-commerce
  - Articles définis par une description et une image
- Mission :
  - Étude de faisabilité
  - Classification automatique
  - Basé sur une description et une image
- Méthode :
  - Réaliser un pré-traitement des descriptions et des images
  - Utiliser différents algorithmes d'analyse d'images et de texte
  - Effectuer un clustering et/ou des prédictions
  - Afficher une visualisation en 2D

# Préparation des données

# Préparation des données

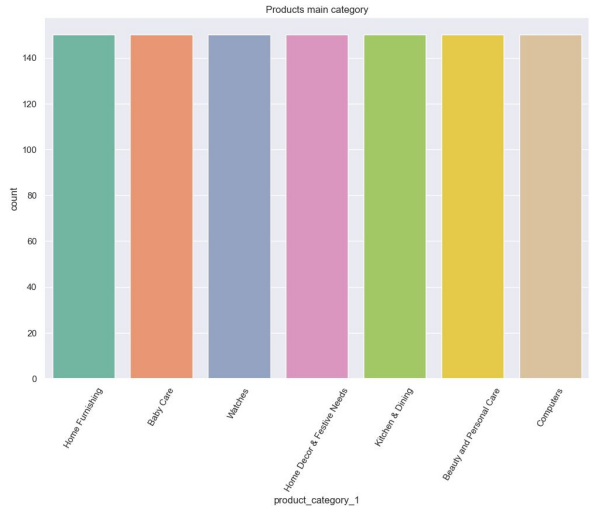
- **1050** articles
- **15** variables

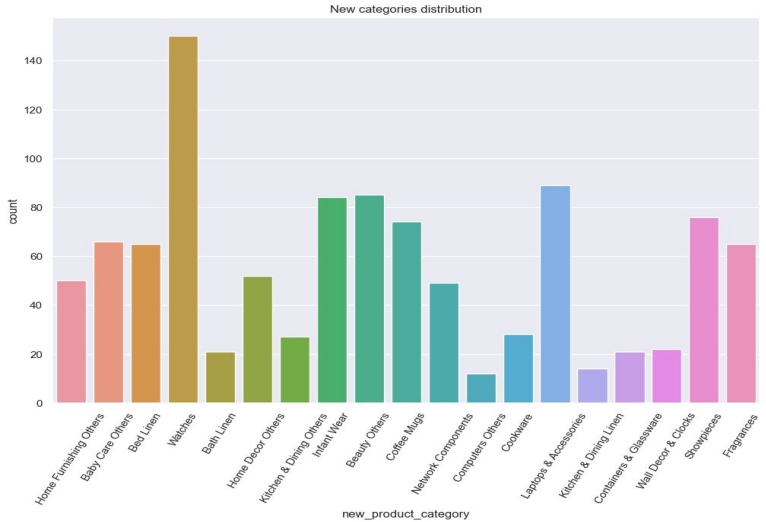



| VARIABLE | product_name | product_category_tree | description | image | ... |
|---|---|---|---|---|---|
| EXEMPLE | Multicolor Door Curtain | Home Furnishing >> Curtains & Accessories >> Curtains | This curtain enhances the look of the interiors. This curtain is made from 100% high quality polyester fabric... | 55b85ea15a1536d46b7190ad6fff8ce7.jpg | ... |

# Nouvelles catégories

- Division de la variable **product_category_tree**
- Création de nouvelles catégories pour affiner classification



# Étude du texte

# Algorithmes utilisés

| Comptage simple | TF-IDF | fastText | BERT | USE |
|---|---|---|---|---|
| Bag of words | Bag of words | Word embedding | Sentence embedding | Sentence embedding |
| Vecteurs de fréquence d'apparition des différents mots utilisés dans la description<br><br>Unigrams<br><br>Bigrams | Vecteurs de fréquence d'apparition des mots pondérés par leur importance dans le corpus<br><br>poids=fréquence du terme×indicateur similarité | Librairie open-source<br><br>Représente les mots en n-grams<br><br>Matrice de vecteurs de mots<br><br>Classification par logistic regression | Pré-entrainé<br><br>Réseaux de transformers<br><br>12 encodeurs<br><br>512 tokens maximum<br><br>Représentation des mots avec contexte | Pré-entrainé<br><br>Réseaux de transformers<br><br>Pas de longueur de texte limite<br><br>512 dimensions en sortie |

# Pré-traitement du texte

- Librairie **spaCy**

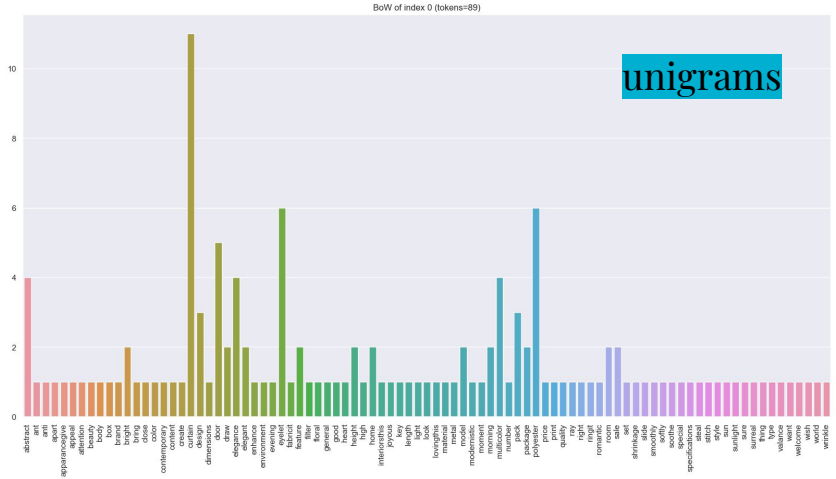
- Tokenization
- Suppression de la ponctuation et des nombres
- Suppression de certains mots :
  - Stopwords
  - Contiennent ponctuation / nombres
  - Noms de moins de deux lettres
- Lemmatization

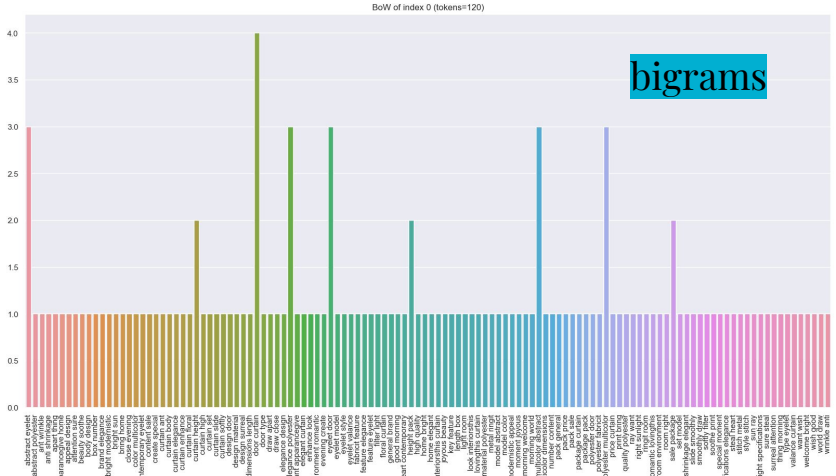
```
Specifications of Eternity Handcrafted unique Mosaic Glass Table Lamp (35 cm, Blue) In The Box Sales Package 1 Table Lamp Number of Contents in Sales Package Pack of 1 General Type Buffet, Desk Color Blue Lamp Body Material Glass Lamp Base Material Metal Bulb Type 1x15 LED, CFSL ,Bulb Light Color Blue Assembly Required No Model Name Handcrafted unique Mosaic Glass Model Number Eternity001007 Power Features Power Requirement 110-240V, 50/60Hz Power Source Ac Additional Features Handcrafted lamps Dimensions Width 15 cm Height 35 cm Weight 1000 g
```

```
['specification', 'eternity', 'handcrafted',
'unique', 'mosaic', 'glass', 'table', 'lamp',
'blue', 'box', 'sales', 'package', 'table', 'lamp',
'number', 'contents', 'sales', 'package', 'pack',
'general', 'type', 'buffet', 'desk', 'color',
'blue', 'lamp', 'body', 'material', 'glass',
'lamp', 'base', 'material', 'metal', 'bulb',
'type', 'led', 'cfsl', 'bulb', 'light', 'color',
'blue', 'assembly', 'required', 'model',
'handcraft', 'unique', 'mosaic', 'glass', 'model',
'number', 'power', 'features', 'power',
'requirement', 'power', 'source', 'additional',
'features', 'handcrafted', 'lamp', 'dimensions',
'width', 'height', 'weight']
```

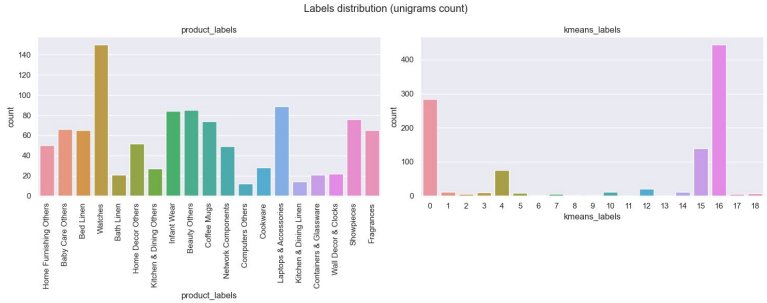
# Comptage simple

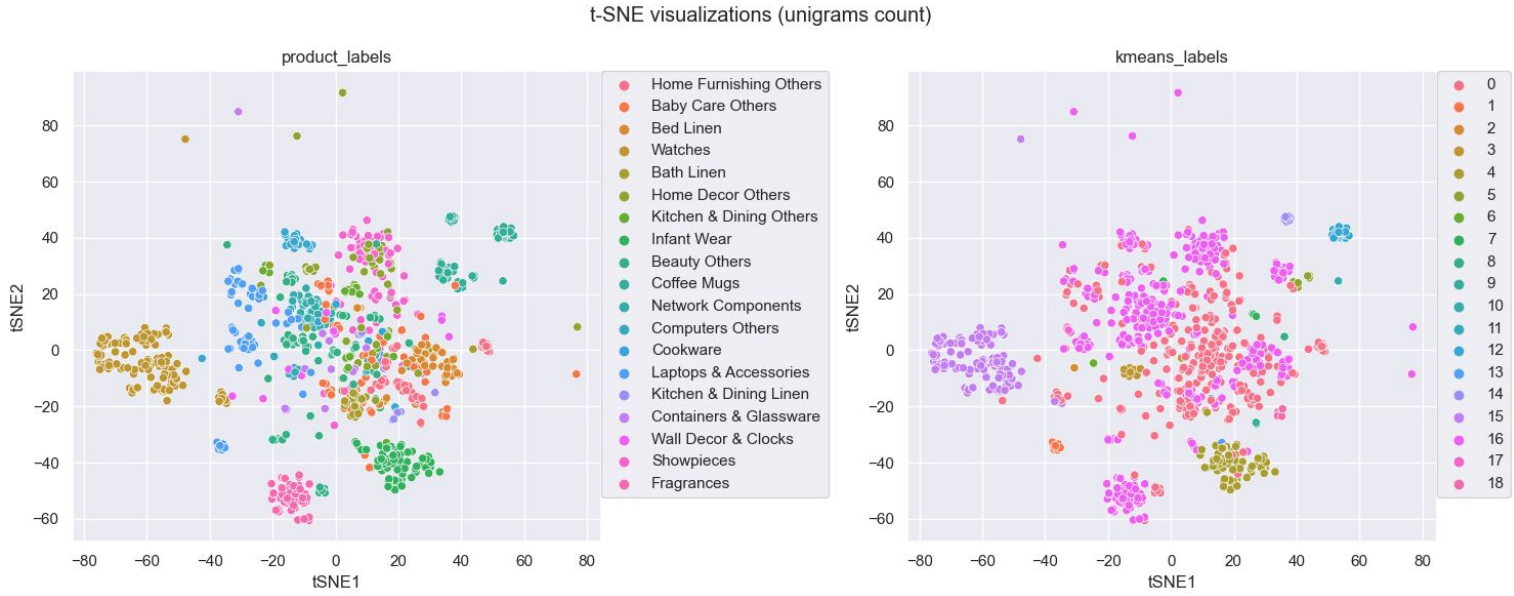
- Prétraitement du texte (**spaCy**)
- Features extraction :
  - **CountVectorizer** *(sklearn)*
  - Unigrams et bigrams
- Clustering avec **KMeans**
- Visualisation avec **t-SNE**



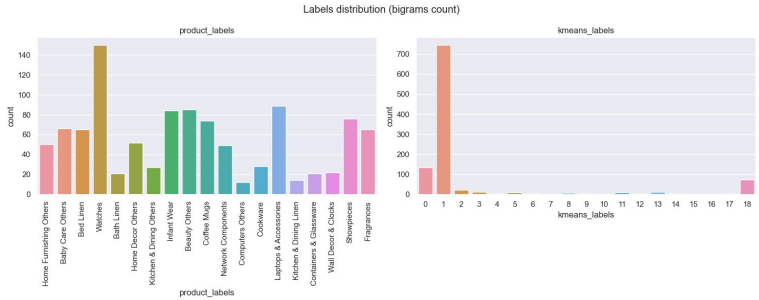



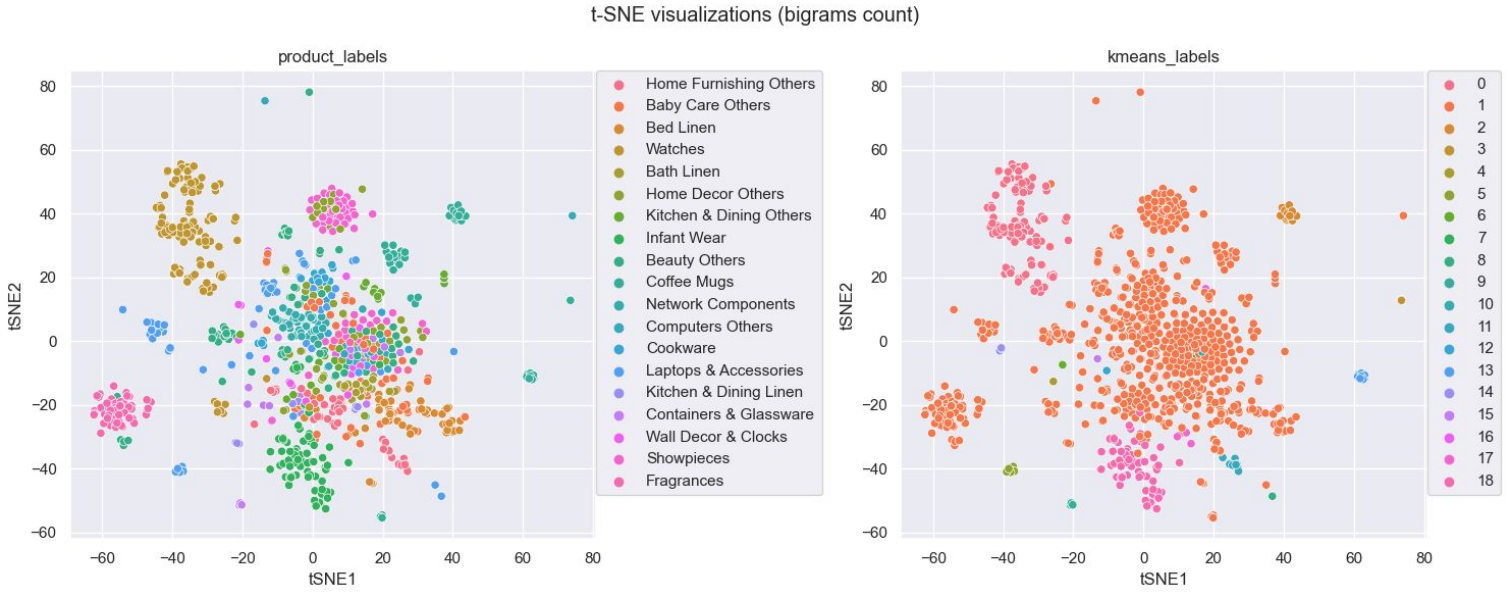
# Comptage simple : unigrams



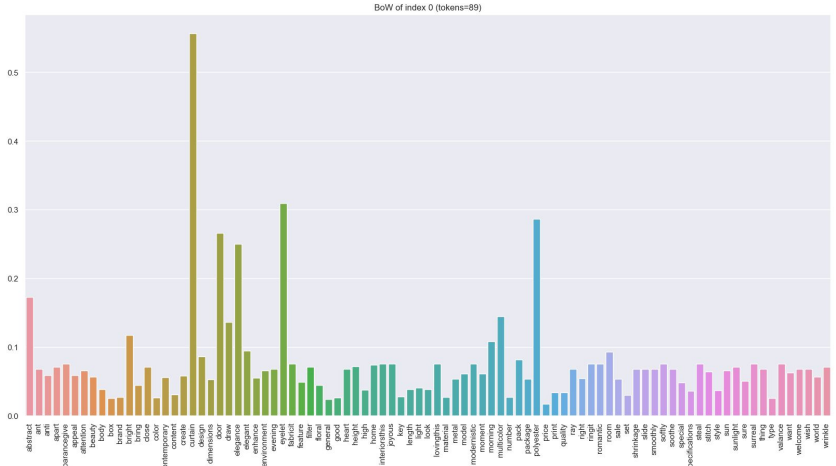
# Comptage simple : bigrams



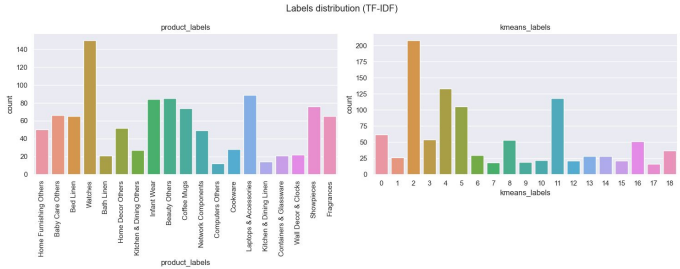



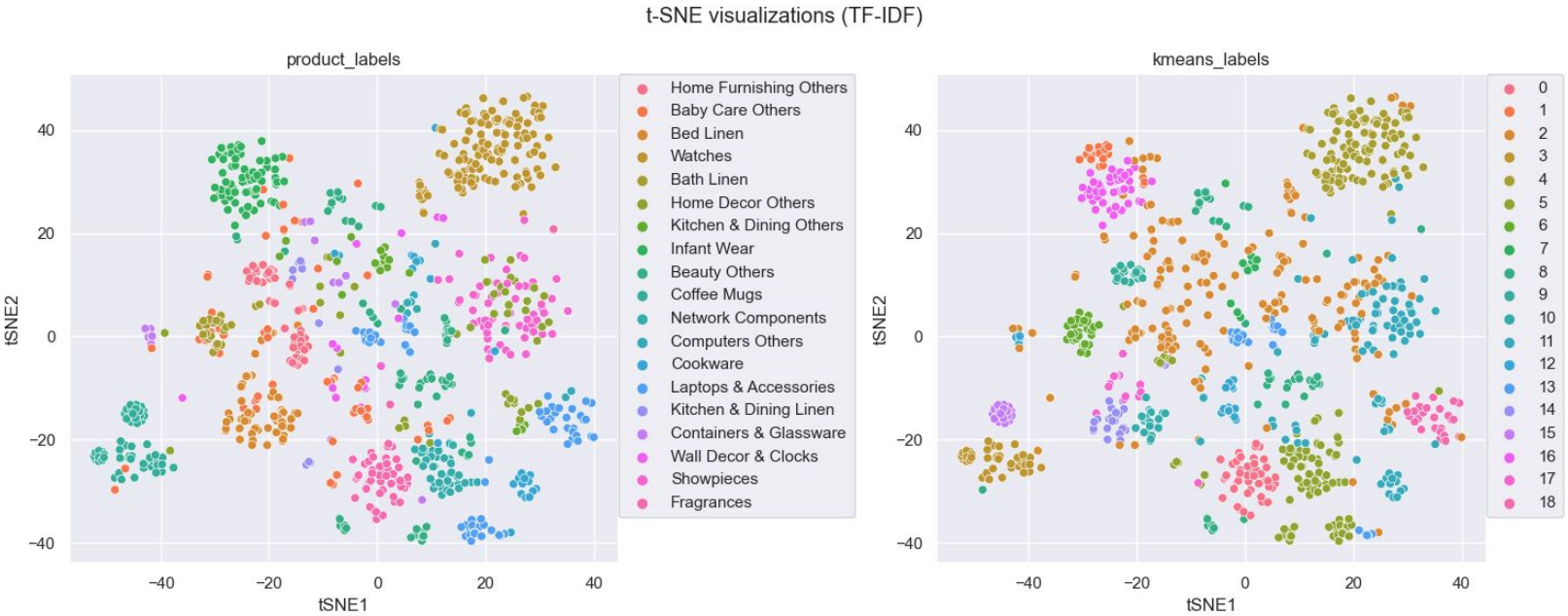
# TF-IDF

- Prétraitement du texte (**spaCy**)
- Features extraction :
  - **TfidfVectorizer** *(sklearn)*
  - Mesure d'importances des mots
- Clustering avec **KMeans**
- Visualisation avec **t-SNE**



# TF-IDF


Labels distribution (TF-IDF)
product_labels
count
Home Furnishing Others
Baby Care Others
Bed Linen
Watches
Bath Linen
Home Decor Others
Kitchen & Dining Others
Infant Wear
Beauty Others
Coffee Mugs
Network Components
Computers Others
Cookware
Laptops & Accessories
Kitchen & Dining Linen
Containers & Glassware
Wall Decor & Clocks
Showpieces
Fragrances
kmeans_labels
count
kmeans_labels


t-SNE visualizations (TF-IDF)
product_labels
Home Furnishing Others
Baby Care Others
Bed Linen
Watches
Bath Linen
Home Decor Others
Kitchen & Dining Others
Infant Wear
Beauty Others
Coffee Mugs
Network Components
Computers Others
Cookware
Laptops & Accessories
Kitchen & Dining Linen
Containers & Glassware
Wall Decor & Clocks
Showpieces
Fragrances
tSNE1
tSNE2
kmeans_labels
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
tSNE1
tSNE2

# fastText

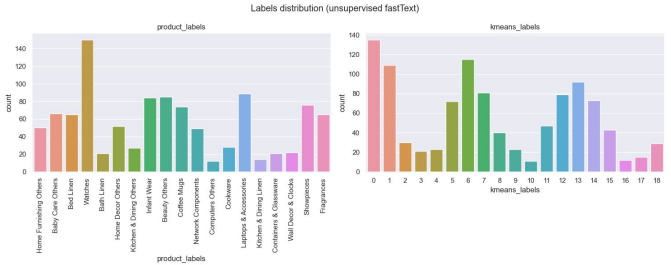
## Non supervisé :

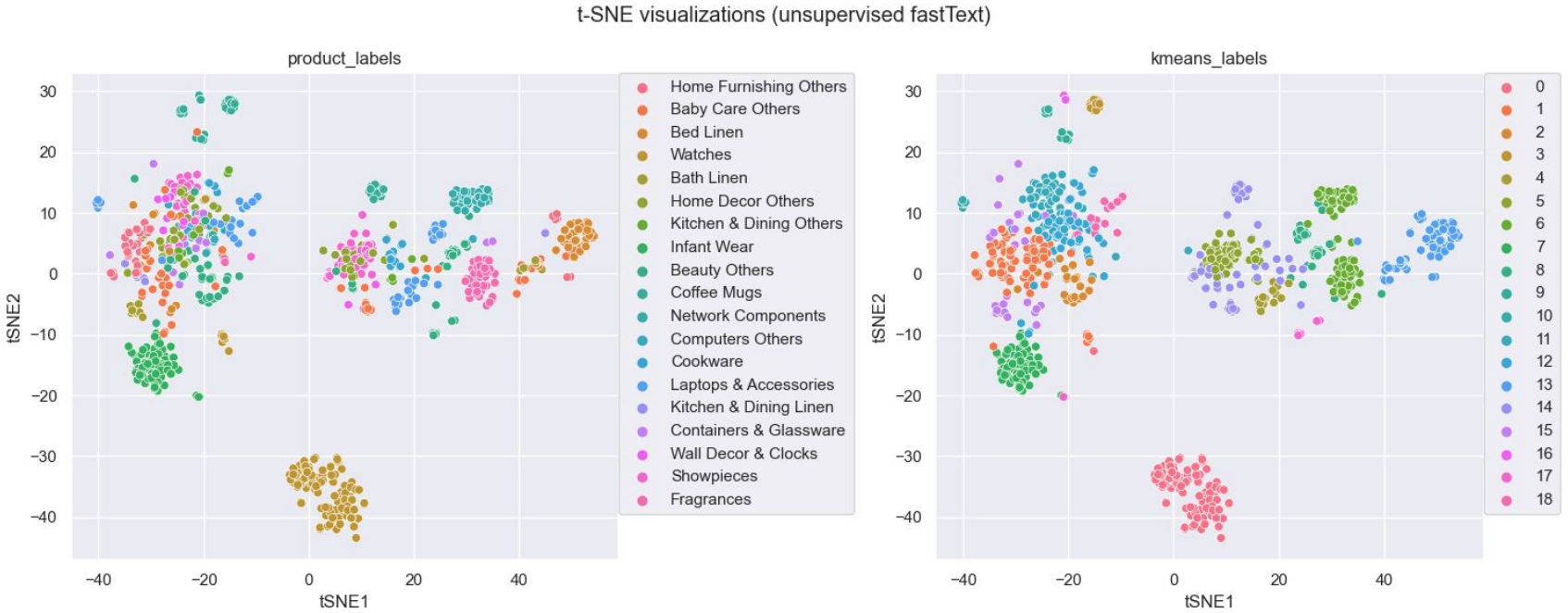
- Prétraitement du texte (**spaCy**)
- Features extraction :
  - Algorithme **fastText**
  - Words embedding
- Clustering avec **KMeans**
- Visualisation avec **t-SNE**

## Supervisé :

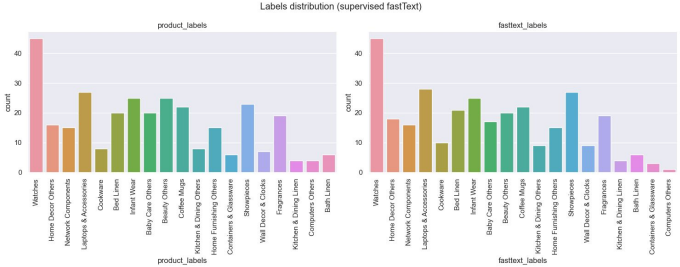
- Prétraitement du texte (**spaCy**)
- Apprentissage **fastText**
  sur jeu d'entraînement
  *(70% - 735 produits)*
- Prédictions sur jeu de test
  *(30% - 315 produits)*
  - Features extraction
  - Prédiction des catégories
- Visualisation avec **t-SNE**

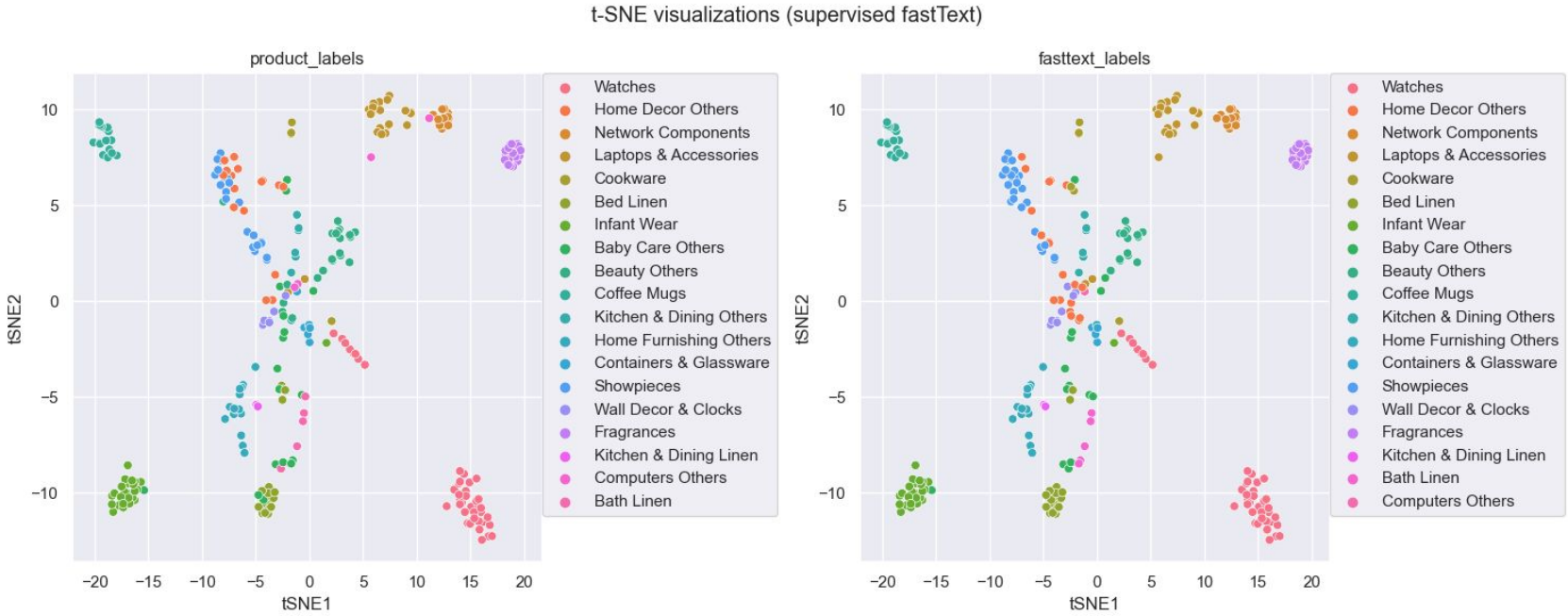
# fastText : non supervisé



# fastText : supervisé

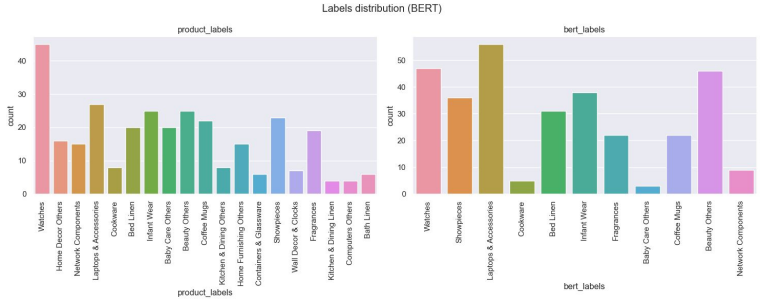


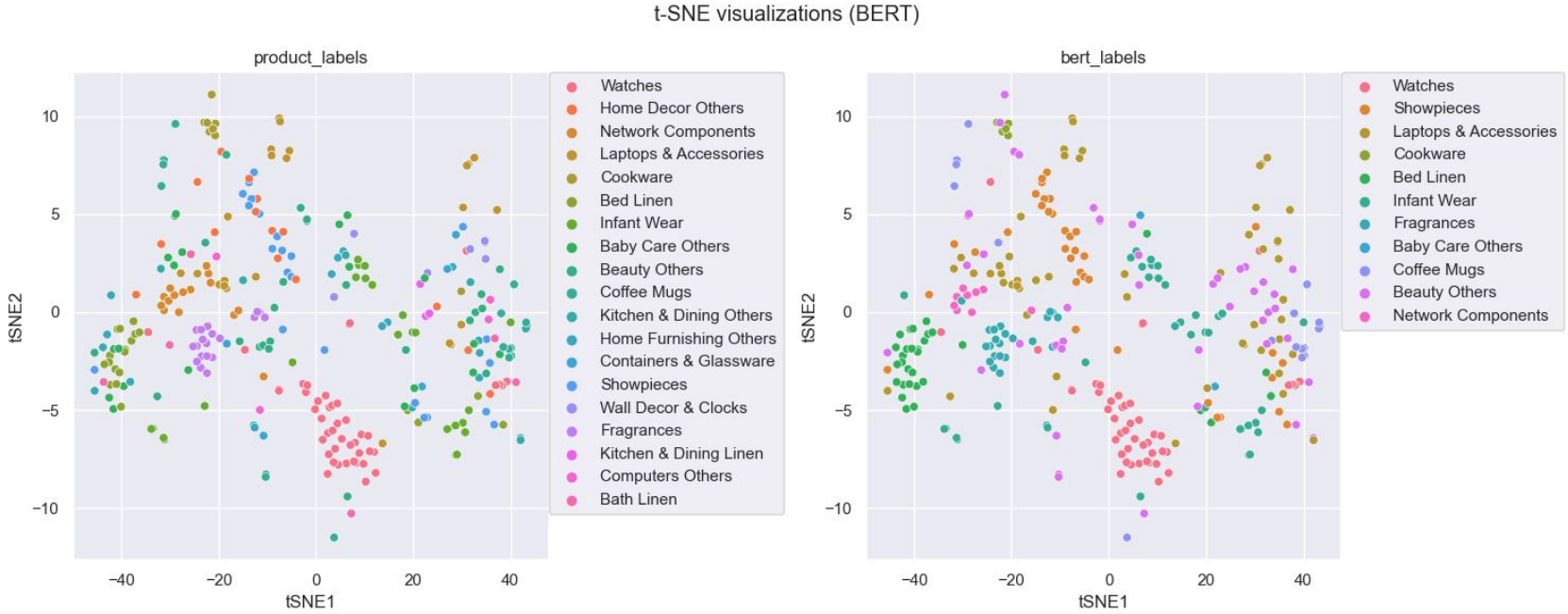



# BERT

- Algorithme **BERT** :
  - Pre-trained model
- Fine tuning sur jeu d'entraînement
- Prédictions sur jeu de test
  - Features extraction
  - Prédiction des catégories
- Visualisation avec **t-SNE**

# BERT

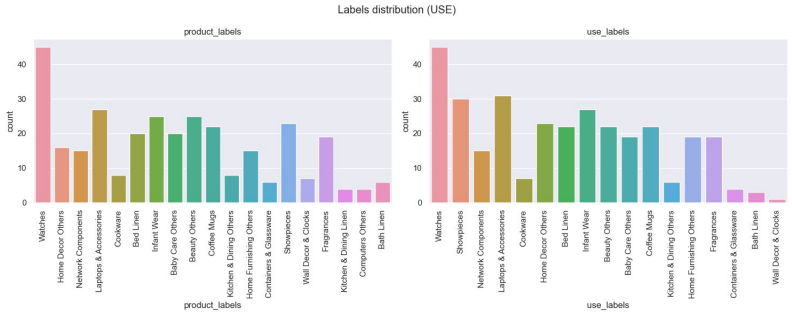

Labels distribution (BERT)
product_labels
count
Watches
Home Decor Others
Network Components
Laptops & Accessories
Cookware
Bed Linen
Infant Wear
Baby Care Others
Beauty Others
Coffee Mugs
Kitchen & Dining Others
Home Furnishing Others
Containers & Glassware
Showpieces
Wall Decor & Clocks
Fragrances
Kitchen & Dining Linen
Computers Others
Bath Linen
bert_labels
count
Watches
Showpieces
Laptops & Accessories
Cookware
Bed Linen
Infant Wear
Fragrances
Baby Care Others
Coffee Mugs
Beauty Others
Network Components

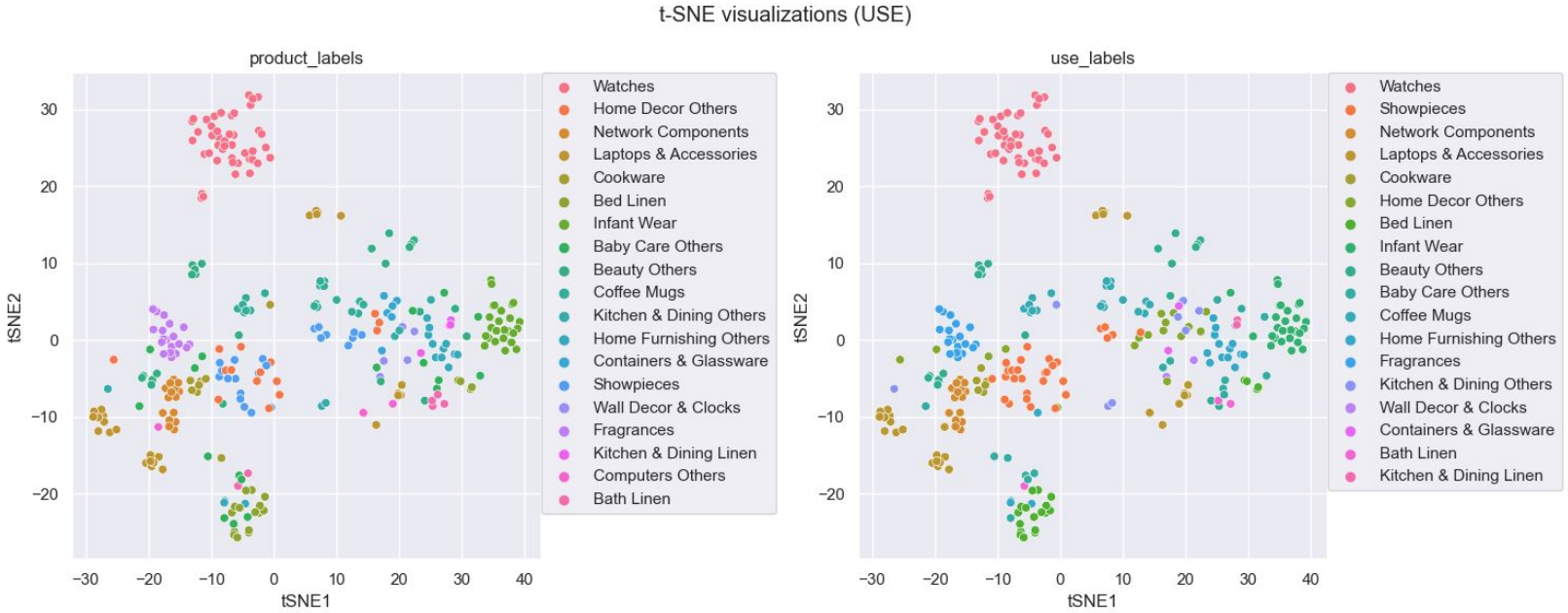

t-SNE visualizations (BERT)
product_labels
tSNE1
tSNE2
Watches
Home Decor Others
Network Components
Laptops & Accessories
Cookware
Bed Linen
Infant Wear
Baby Care Others
Beauty Others
Coffee Mugs
Kitchen & Dining Others
Home Furnishing Others
Containers & Glassware
Showpieces
Wall Decor & Clocks
Fragrances
Kitchen & Dining Linen
Computers Others
Bath Linen
bert_labels
tSNE1
tSNE2
Watches
Showpieces
Laptops & Accessories
Cookware
Bed Linen
Infant Wear
Fragrances
Baby Care Others
Coffee Mugs
Beauty Others
Network Components

# Universal Sentence Encoder

- Algorithme **USE** :
  - Pre-trained model
- Fine-tuning sur jeu d'entraînement
- Prédictions sur jeu de test
  - Features extraction
  - Prédiction des labels
- Visualisation avec **t-SNE**

# Universal Sentence Encoder



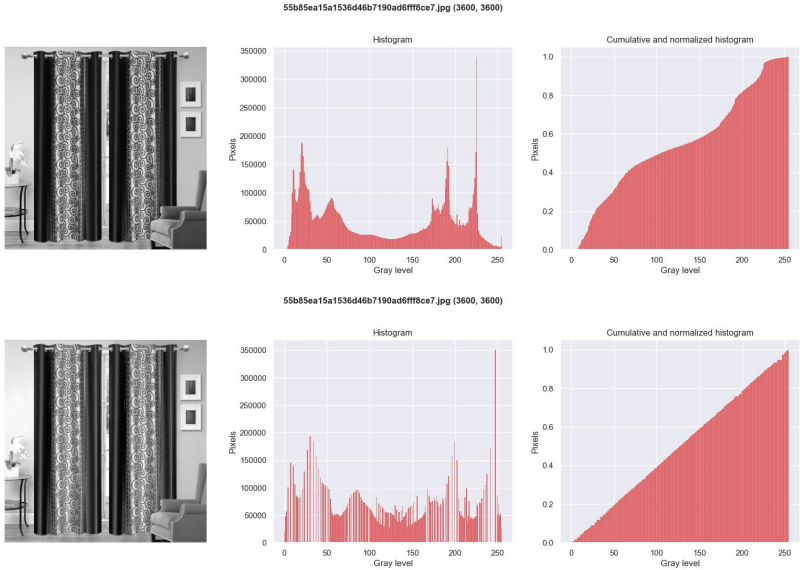
# Étude des images

# Algorithmes utilisés

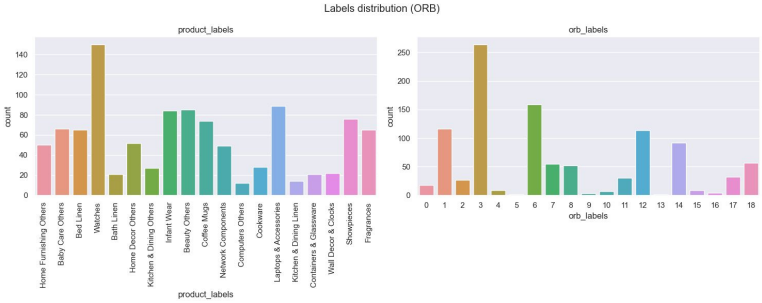
| ORB | ResNet50 |
| --- | --- |
| Algorithme OpenCV<br><br>Détection des points d'intérêt<br><br>Assigne des informations numériques (descriptors) aux points d'intérêt | Réseau neuronal convolutif (50 couches)<br><br>Pré-entraîné sur des millions d'images<br><br>Taille d'image : 224 par 224 pixels |

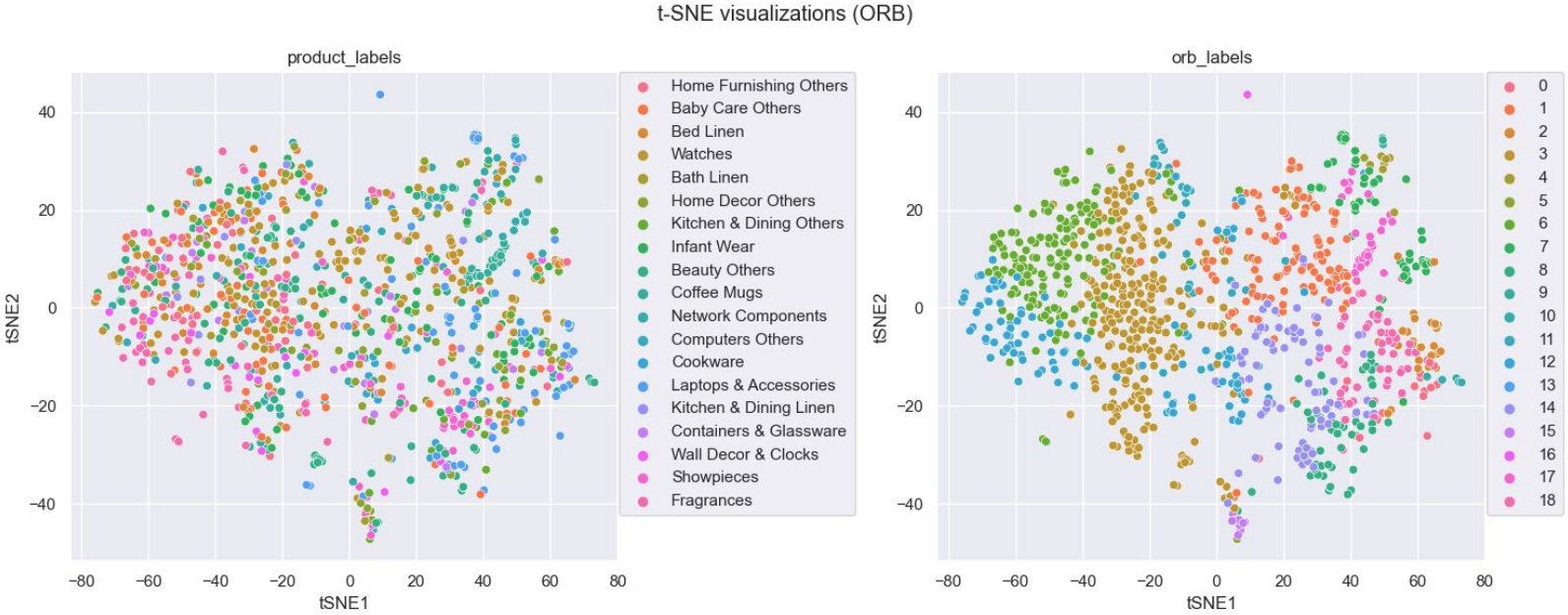
# ORB

- Prétraitement des images
  - Histograms
- Algorithme **ORB :**
  - Détection des *descriptors*
- Clustering des *descriptors* pour définir des *visual words*
- Création de bag of visual words
- Clustering avec **KMeans** *(sklearn)*
- Visualisation avec **t-SNE**




# ORB

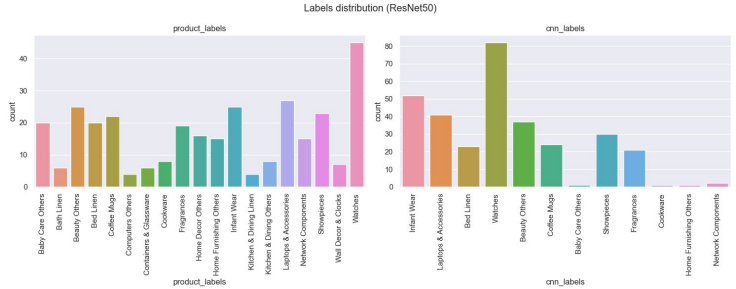

Labels distribution (ORB)
product_labels
count
Home Furnishing Others
Baby Care Others
Bed Linen
Watches
Bath Linen
Home Decor Others
Kitchen & Dining Others
Infant Wear
Beauty Others
Coffee Mugs
Network Components
Computers Others
Cookware
Laptops & Accessories
Kitchen & Dining Linen
Containers & Glassware
Wall Decor & Clocks
Showpieces
Fragrances
product_labels
orb_labels
count
orb_labels

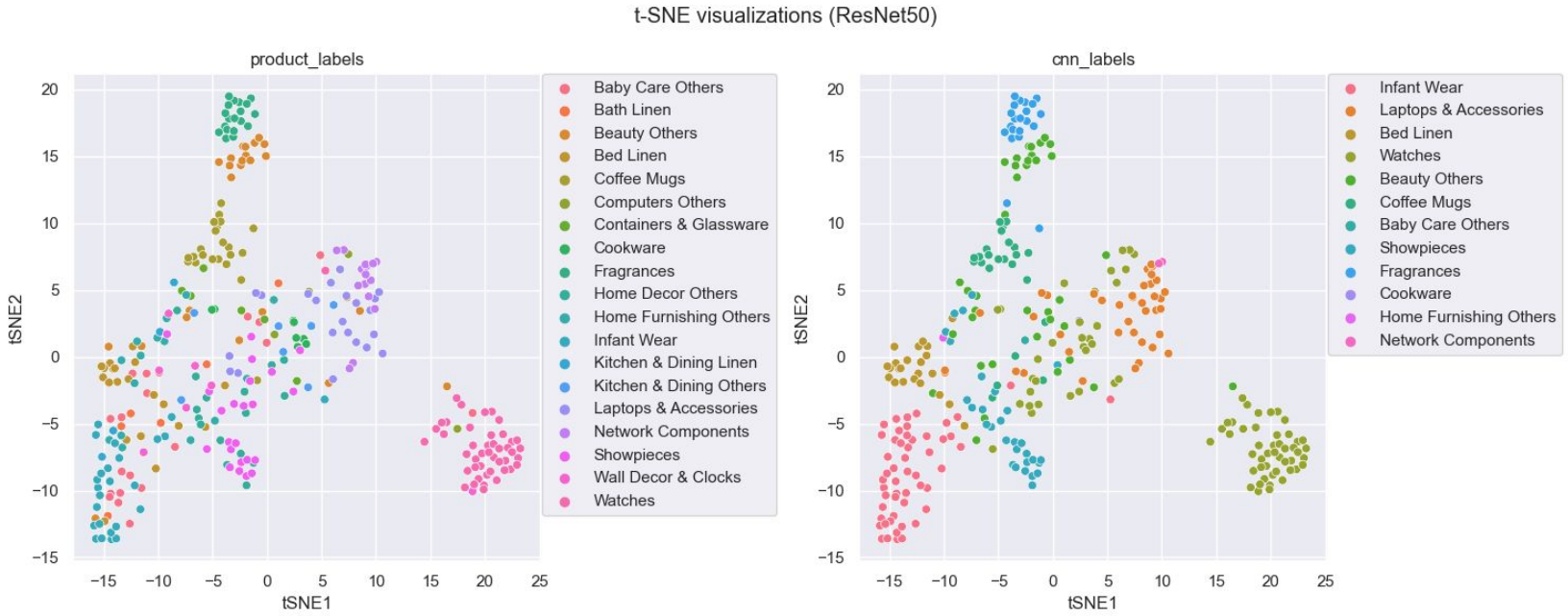

t-SNE visualizations (ORB)
product_labels
tSNE2
tSNE1
Home Furnishing Others
Baby Care Others
Bed Linen
Watches
Bath Linen
Home Decor Others
Kitchen & Dining Others
Infant Wear
Beauty Others
Coffee Mugs
Network Components
Computers Others
Cookware
Laptops & Accessories
Kitchen & Dining Linen
Containers & Glassware
Wall Decor & Clocks
Showpieces
Fragrances
orb_labels
tSNE2
tSNE1
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

# CNN transfer learning

- Algorithme **ResNet50** :
  - Pre-trained model
- Fine-tuning sur jeu d'entraînement
- Prédictions sur jeu de test
  - Features extraction
  - Prédiction des labels
- Visualisation avec **t-SNE**

# CNN transfer learning

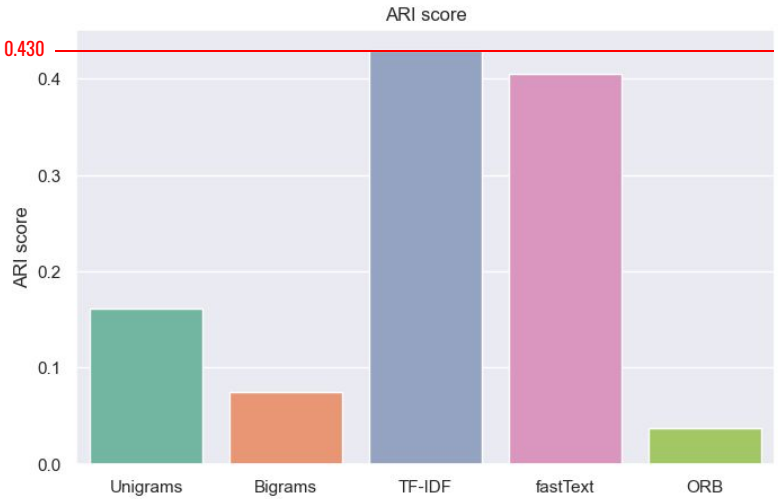


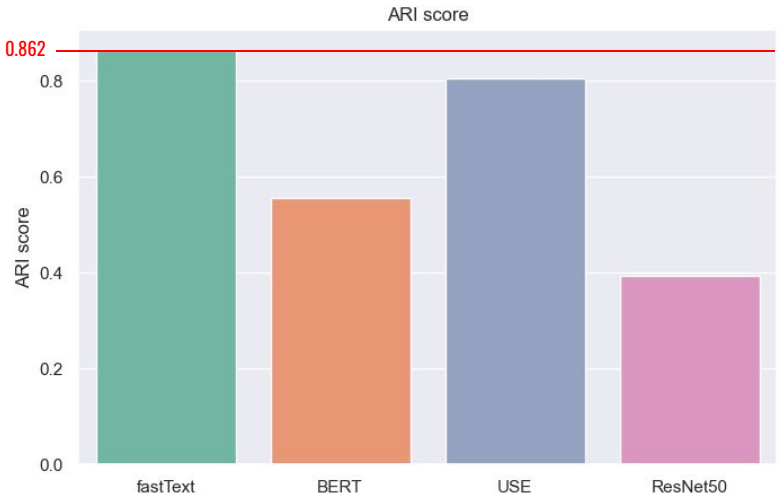
# Faisabilité et conclusion

# Comparaison des algorithmes : ARI score

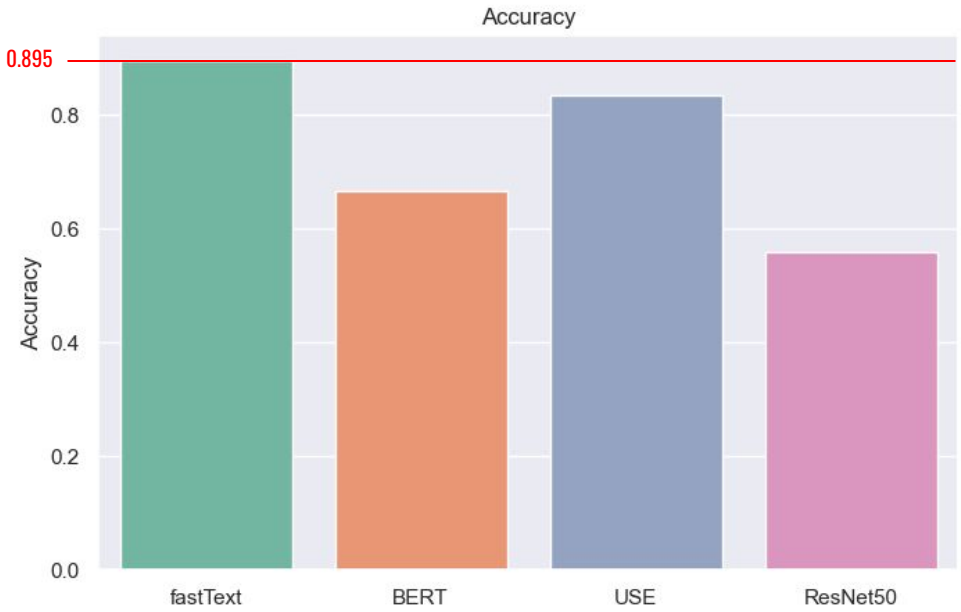
## Non supervisé




## Supervisé

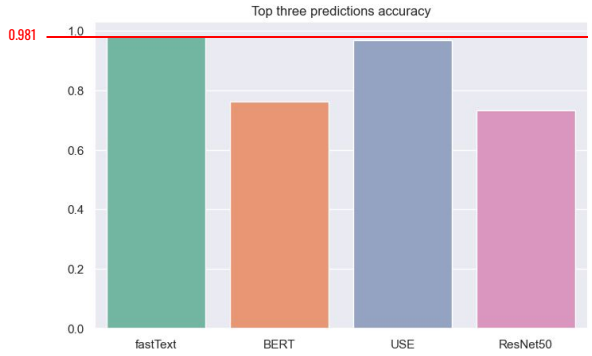


# Comparaison des algorithmes : précision




Prédictions : probabilités d'appartenance à une catégorie

Précision en regardant les trois plus grandes probabilités :



# Conclusion

- Utilisation des descriptions et images
- Réduction de dimensions pertinente
- Test de différentes approches
- ARI scores et précisions concluantes :
    - Un moteur de classification est faisable !

Améliorations :

- Manque de données d'entraînement
- Affiner la catégorisation (biaisée)
- Prédictions textes + images

Merci !
Avez-vous des questions ?