Projet 6 : Classifiez automatiquement des biens de consommation

Eva Rondeau

Présentation

<u>Objectif</u> : faciliter la mise en ligne de nouveaux produits (vendeurs) et la recherche de ces nouveaux produits (acheteurs) en automatisant l'attribution de la catégorie du produit à partir de la description et des images des produits

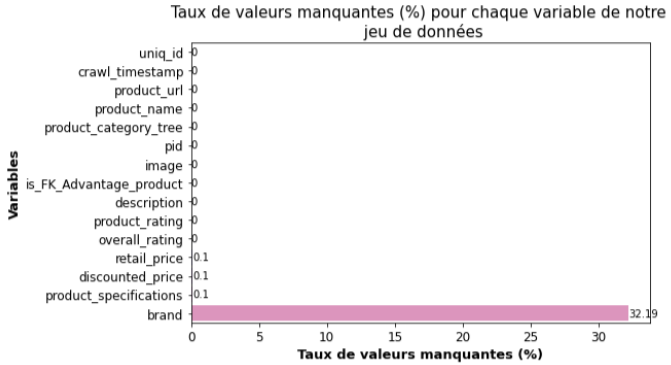
<u>Jeu de données</u> :

- 1050 lignes (produits)
- 15 colonnes (informations complémentaires)

Informations produits
Lien
Nom du produit
Catégories
Prix (rabais ou non)
Nom image
Description
Marque
Note

Analyse exploratoire

1. Valeurs manquantes



2. Doublons

Pas de doublons observés.

⇒ Aucun nettoyage à prévoir

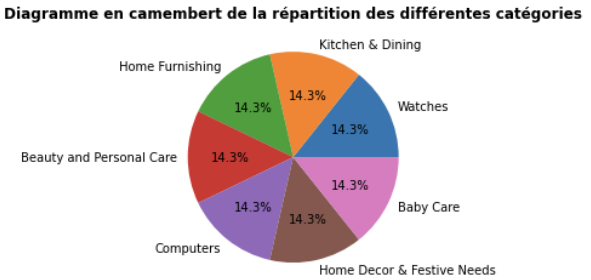
3. Catégories

- 642 catégories uniques

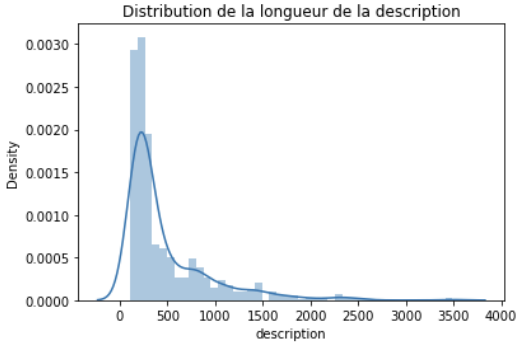
```
["Baby Care >> Baby Bedding >> Baby Blankets >> Offspring Baby Blankets"]
```

Sélection de la catégorie générale :

- 7 catégories générales



4. Description



Majorité des produits contenant un faible nombre de caractères (< 500)

Pré-traitement

1. TEXTE

Processus de transformation afin de préparer les données textuelles à la classification

1.1 Phrase d'exemple

```
'Buy Go Hooked Wheel Pizza Cutter for Rs.199 online. Go Hooked Wheel Pizza Cutter at best prices with FREE shipping & cash on d
elivery. Only Genuine Products. 30 Day Replacement Guarantee.'
```

1.2. Mise en minuscule

```
'buy go hooked wheel pizza cutter for rs.199 online. go hooked wheel pizza cutter at best prices with free shipping & cash on d
elivery. only genuine products. 30 day replacement guarantee.'
```

1.3. Tokenisation

```
['buy', 'go', 'hooked', 'wheel', 'pizza', 'cutter', 'for', 'rs.199', 'online', '.', 'go', 'hooked', 'wheel', 'pizza', 'cutter',
'at', 'best', 'prices', 'with', 'free', 'shipping', '&', 'cash', 'on', 'delivery', '.', 'only', 'genuine', 'products', '.', '3
0', 'day', 'replacement', 'guarantee', '.']
```

1.4. Suppression ponctuation

```
['buy', 'go', 'hooked', 'wheel', 'pizza', 'cutter', 'for', 'rs', '199', 'online', 'go', 'hooked', 'wheel', 'pizza', 'cutter',
'at', 'best', 'prices', 'with', 'free', 'shipping', 'cash', 'on', 'delivery', 'only', 'genuine', 'products', '30', 'day', 'repl
acement', 'guarantee']
```

1.5. Suppression stop-words

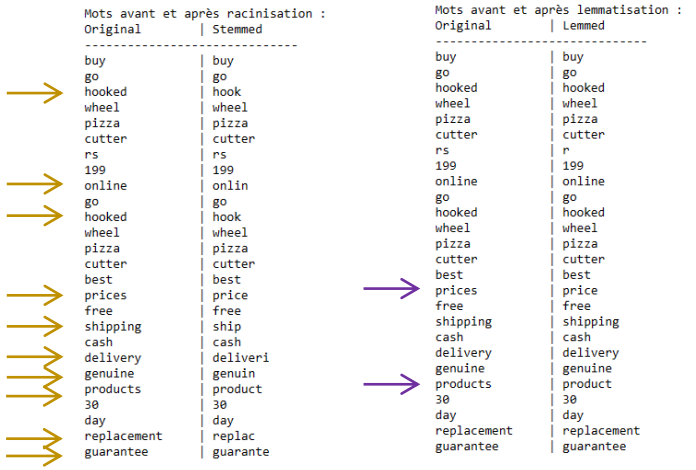
```
['buy', 'go', 'hooked', 'wheel', 'pizza', 'cutter', 'rs', '199', 'online', 'go', 'hooked', 'wheel', 'pizza', 'cutter', 'best',
'prices', 'free', 'shipping', 'cash', 'delivery', 'genuine', 'products', '30', 'day', 'replacement', 'guarantee']
```

1.6. Lemmatisation

```
['buy', 'go', 'hooked', 'wheel', 'pizza', 'cutter', 'r', '199', 'online', 'go', 'hooked', 'wheel', 'pizza', 'cutter', 'best',
'price', 'free', 'shipping', 'cash', 'delivery', 'genuine', 'product', '30', 'day', 'replacement', 'guarantee']
```

Racinisation *vs.* Lemmatisation

- Racinisation (stemming) : réduction du mot à sa racine (radical) après suppression suffixe et préfixe (coupe les mots, plus difficile à interpréter).
- Lemmatisation : réduction du mot à sa forme canonique (verbe à l'infinitif, singulier masculin, …). Prend en compte la signification des mots.



```
Mots avant et après racinisation :
Original        | Stemmed
--------------------------
buy             | buy
go              | go
hooked          | hook
wheel           | wheel
pizza           | pizza
cutter          | cutter
rs              | rs
199             | 199
online          | onlin
go              | go
hooked          | hook
wheel           | wheel
pizza           | pizza
cutter          | cutter
best            | best
prices          | price
free            | free
shipping        | ship
cash            | cash
delivery        | deliveri
genuine         | genuin
products        | product
30              | 30
day             | day
replacement     | replac
guarantee       | guarante
```

```
Mots avant et après lemmatisation :
Original        | Lemmed
--------------------------
buy             | buy
go              | go
hooked          | hooked
wheel           | wheel
pizza           | pizza
cutter          | cutter
rs              | r
199             | 199
online          | online
go              | go
hooked          | hooked
wheel           | wheel
pizza           | pizza
cutter          | cutter
best            | best
prices          | price
free            | free
shipping        | shipping
cash            | cash
delivery        | delivery
genuine         | genuine
products        | product
30              | 30
day             | day
replacement     | replacement
guarantee       | guarantee
```

Après nettoyage :

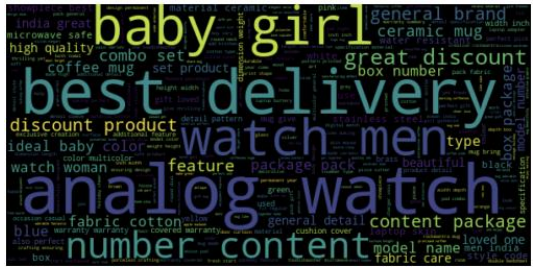
- Phrase d'exemple : 35 tokens ⟹ 26 tokens
- Corpus : 90 712 tokens ⟹ 59 786 tokens

Pré-traitement

1. TEXTE

- WORDCLOUD

Après 2ème traitement



Mots assez généralistes (free, shipping, genuine, products, …)

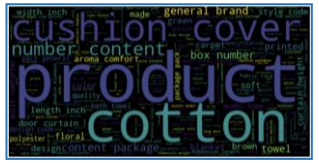
1.7. Suppression mots fréquents

1.8. Suppression mots rares

1.9. Suppression mots courts (< 3 lettres)

Pour chaque catégorie :

Home Furnishing :



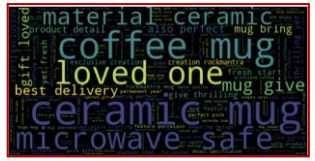
Baby Care

Watches

Computers

Home Decor & Festive Needs

Kitchen & Dining



Beauty and Personal Care

Extraction des features

1. TEXTE

Extraction features : transformation données textuelles en données numériques

« Bag of Words » (sac de mots)

- Texte représenté sous forme de **vecteurs**
- Conversion du texte en **matrice d'occurrence de mots**
- **Ne prennent pas en compte l'ordre et le contexte**

Mot ↓

Document →

	youth	youthful	yuva	zero	zinc	zingalalaa	zipper	zone	zora	zyxel
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0

CountVectorizer	Tf-idf
Nombre d'occurrences de chaque mot dans le document.	Prise en compte de la fréquence de chaque mot dans le document et dans le corpus.

Word Embeddings (incorporation de mots)

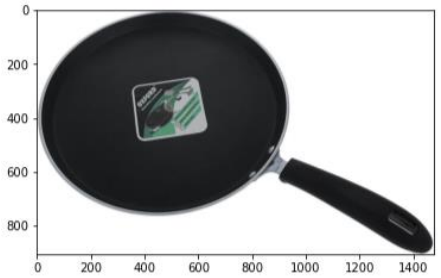
- Texte représenté sous forme de **vecteurs**
- Capturent des informations plus denses: **contexte et ordre des mots** dans un texte

Word2Vec	BERT	USE
Identifie les similarités et relations linguistiques entre les mots. - CBOW : prédit le mot selon le contexte - Skip-gram : prédit le contexte selon le mot	Modèle de langage bidirectionnel : identifie la signification des mots dans un contexte donné: Prend en compte le contexte précédent ET suivant le mot.	Une phrase utilisée pour prédire la phrase suivante ou précédente

Pré-traitement

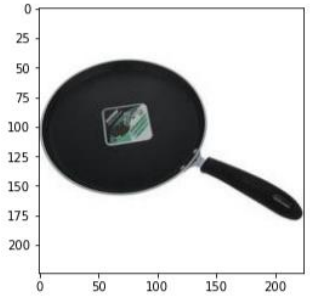
2. IMAGE

2.1. Redimensionnement



Dimensions image : 1478 x 904

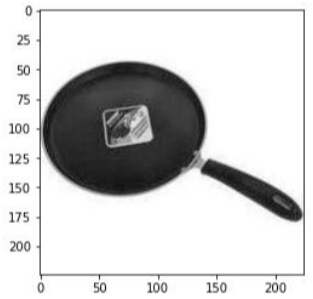
Après redimensionnement

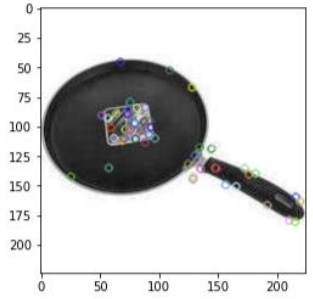


Dimensions image : 224 x 224

2.2. Niveaux de gris + contrastes

Après chargement en niveaux de gris + augmentation des contrastes (histogramme)





58 descripteurs (vecteurs de longueur 128)

Pré-traitement :

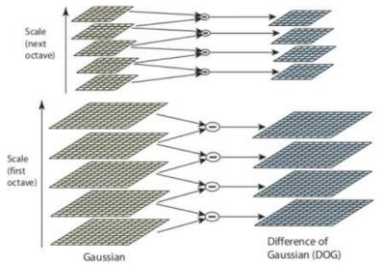
- Niveaux de gris : évite de se concentrer sur les différentes composantes de couleur mais uniquement sur les variations d'intensité
- Contrastes de l'image : mise en évidence des points d'intérêts : histogramme utilisé pour augmenter les contrastes des pixels
- Descripteurs : vecteurs numériques représentant les caractéristiques visuelles locales des points d'intérêts détectés par l'algorithme SIFT (Scale-Invariant Feature Transform)

Extraction des features

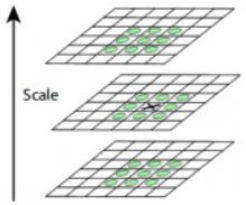
2. IMAGE

Algorithme SIFT (Scale-Invariant Feature Transform)

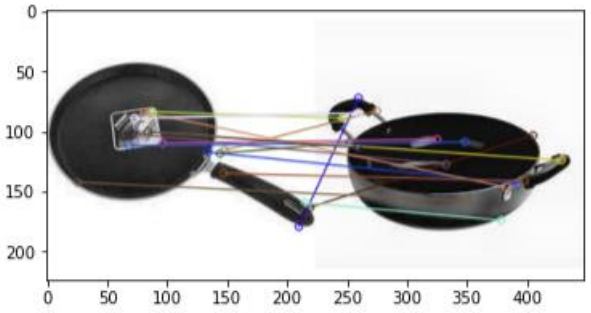
- Utilise la **Différence des Gaussiens** (DoG) calculé en soustrayant les images voisines



- Valeurs du DoG utilisées afin de **localiser** précisément les **points d'intérêts** (*extrema* locaux)



- Points d'intérêts rendus **invariants à l'orientation et à l'échelle** (calculé en fonction des orientations locales)
- **Descripteurs calculés** pour chaque point d'intérêt
- **Correspondances des caractéristiques** entre 2 images :



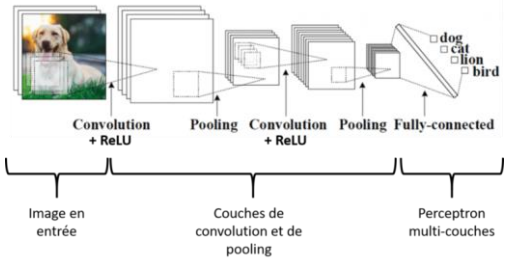
17 correspondances trouvées

Extraction des features

2. IMAGE

Algorithme CNN (Conventional Neural Network)

- Réseaux de neurones convolutifs
- Modèles de classification d'images
- Extraction automatique des features pertinentes
- Couches de neurones :
  - ✓ Convolution : bords, formes, texture
  - ✓ Pooling : réduction dimension spatiale données
  - ✓ Couches denses : prédictions, classifications données



Transfer Learning

- Utilisation modèle pré-entraîné
- Evite de repartir de 0
- Gain de temps

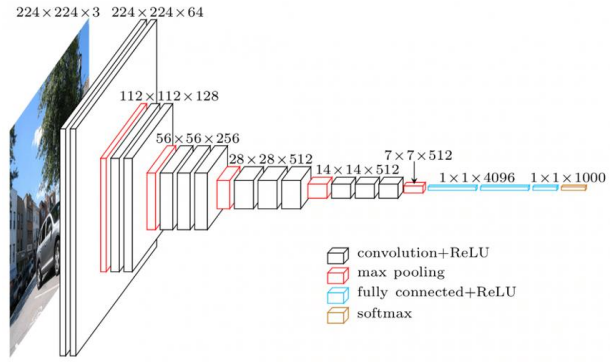
1. VGG16

- Entrainé sur plus d'1 million d'images
- Base de données : **ImageNet**
- Classement des images en 1000 classes d'objets

Architecture :

- 16 couches :
  - ✓ 13 couches de convolution
  - ✓ 3 couches denses avec 4096 neurones chacune

Architecture VGG16 :



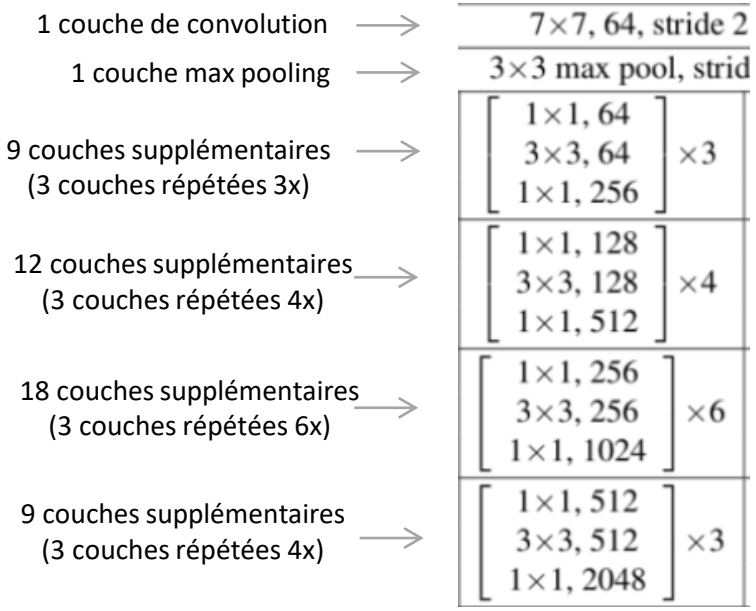
Extraction des features

2. IMAGE

2. ResNet50 (Residual Network 50)

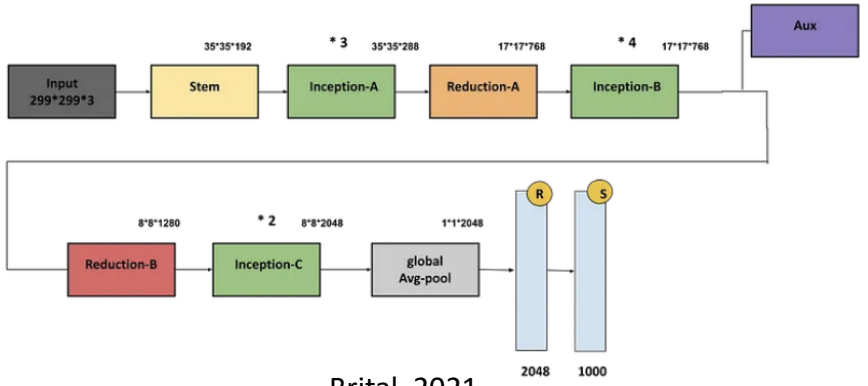
Architecture : blocs résiduels de différentes tailles
1 bloc = couches de convolution, normalisation, mise à l'échelle et pooling

- 50 couches :
  - ✓ Combinaison de convolutions avec différents noyaux et filtres



3. InceptionV3

- 48 couches :
  - ✓ Couches initiales (Stem) : pré-traitement données en entrée
  - ✓ Blocs Inception : passage informations à travers les blocs en évitant les pertes
  - ✓ Blocs Réduction : après blocs Inception, réduction de dimensions spatiales des caractéristiques
  - ✓ Global Average Pooling : agrégation des informations spatiales en calculant la moyenne des caractéristiques
  - ✓ Couches de classification (fully-connected)



Brital, 2021

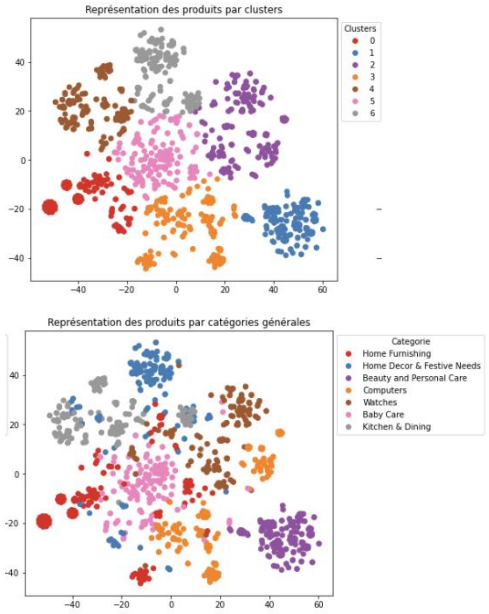
Résultats de la faisabilité

Réduction de dimensions:
- Création de nouvelles features
- Conservation des informations utiles à la classification
- Gain en temps de calcul
- Technique utilisée : t-SNE (t-Distributed Stochastic Neighbor Embedding)

Score ARI (Adjusted Rand Index):
- Score de similitude
- Evalue la qualité de regroupement des clusters
- Compare les clusters obtenus avec les clusters de référence (vraies étiquettes)

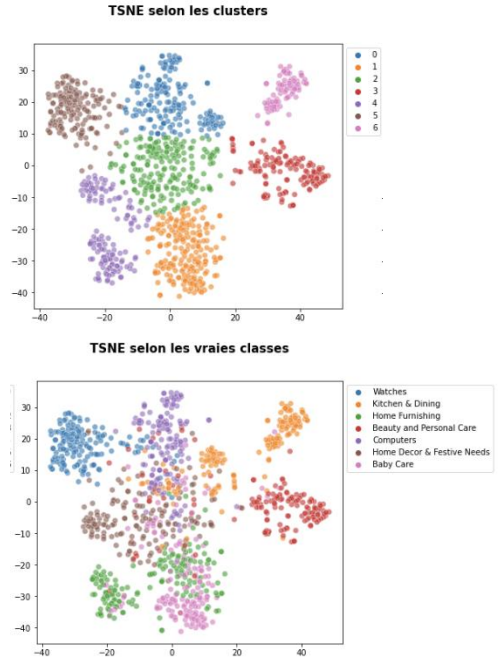
1. TEXTE

Modèle	Score ARI
CountVectorizer	0,497
Tf-idf	0,507
Word2Vec (CBOW)	0,396
Word2Vec (SG)	0,355
BERT	0,321
USE	0,397



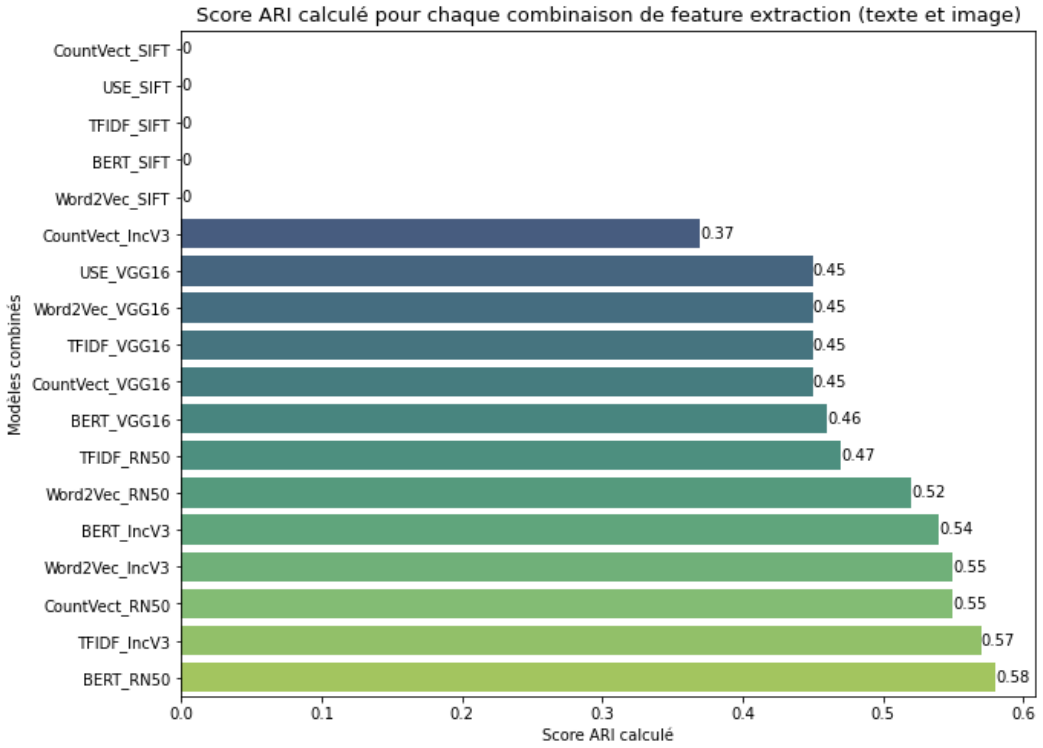
2. IMAGE

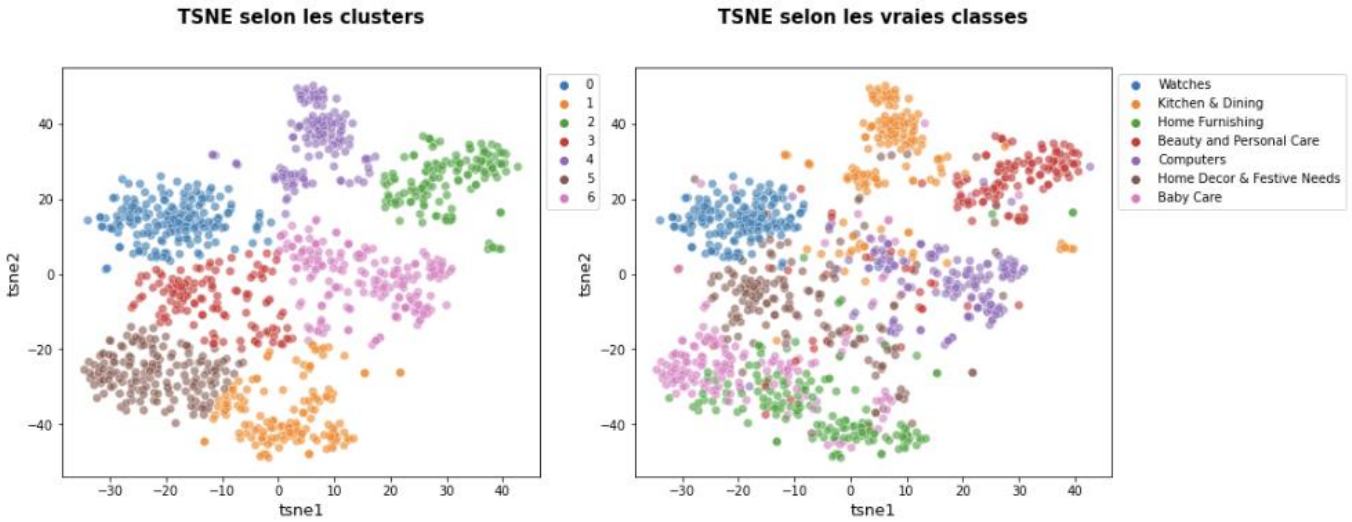
Modèle	Score ARI
SIFT	0,053
VGG16	0,452
ResNet50	0,479
InceptionV3	0,557



Résultats de la faisabilité

3. TEXTE + IMAGE





- Meilleur score ARI : BERT (texte) + ResNet50 (image)
- SIFT : proche de 0
- Modèles d'extraction de données images ayant une influence plus forte par rapport aux modèles d'extraction de données textuelles

→ Faisabilité de regrouper automatiquement des produits d'une même catégorie

Résultats classification supervisée

Classification image

Idée : déterminer si les images présentent des caractéristiques visuelles distinctes justifiant de les classer dans des catégories différentes

1. **Séparation** des données d'entrainement, de test et de validation
2. **Création du modèle** (VGG16, ResNet50, InceptionV3)
3. Utilisation de **métriques** évaluant la performance de chaque modèle:
    - Accuracy : proportion d'images correctement classées par rapport à l'ensemble
    - Loss : erreur entre les prédictions du modèle et les étiquettes de classes réelles
    - Temps d'entrainement : temps de calcul pour l'entrainement du modèle

Résultats:

Model	Data Augmentation	Validation accuracy	Validation loss	Train accuracy	Train Loss	Time
VGG16	No Data augmentation	0.836502	0.717688	0.979670	0.064009	1221.731733
ResNet50	No Data augmentation	0.840304	0.684557	0.963151	0.100062	388.292922
InceptionV3	No Data augmentation	0.844106	0.561488	0.950445	0.152528	644.806028

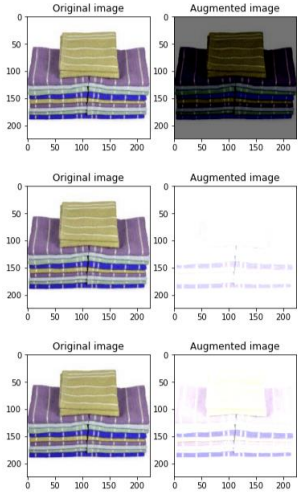
- Différence de **temps de calcul** entre VGG16 (long) et ResNet50 (court)
- **Meilleur apprentissage** sur les données d'entrainement pour **VGG16**
- **Meilleures capacités de généralisation** pour **InceptionV3**

Résultats classification supervisée

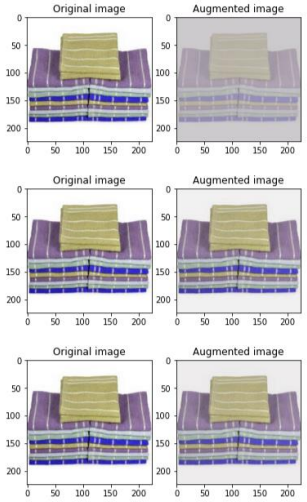
Classification image

Data augmentation : technique de pré-traitement appliquant des transformations aléatoires aux images existantes et créant ainsi de nouvelles versions des images originales

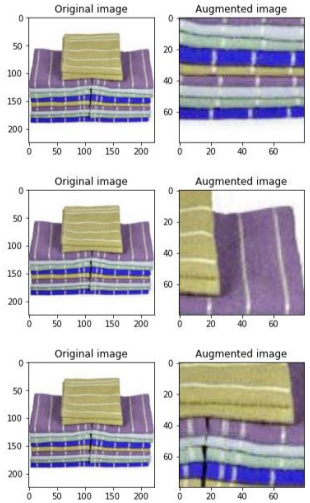
Luminosité



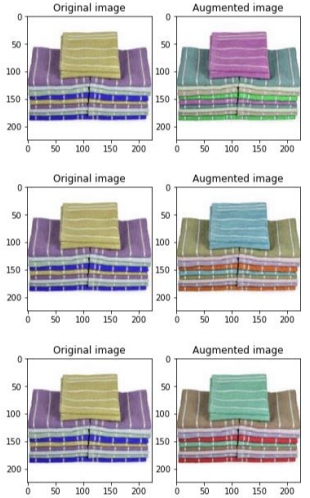
Contrastes



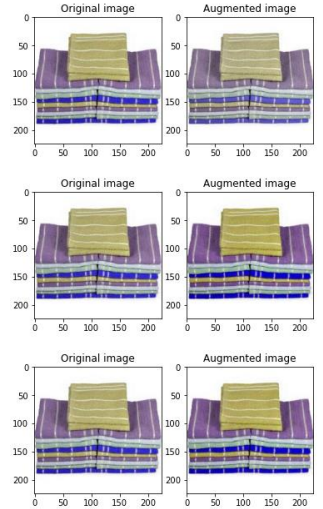
Recadrage



Teinte



Saturation



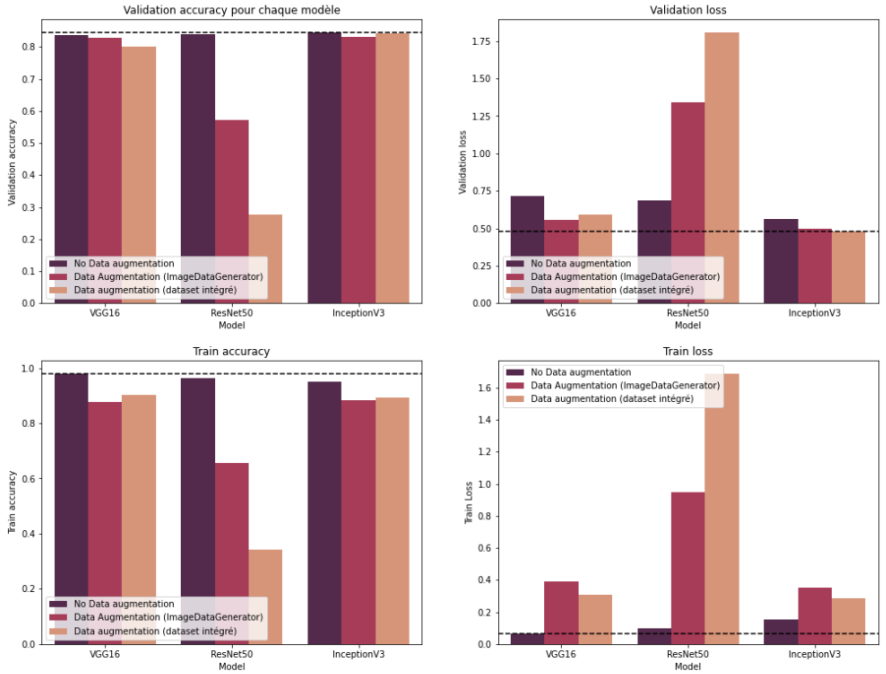
Résultats classification supervisée

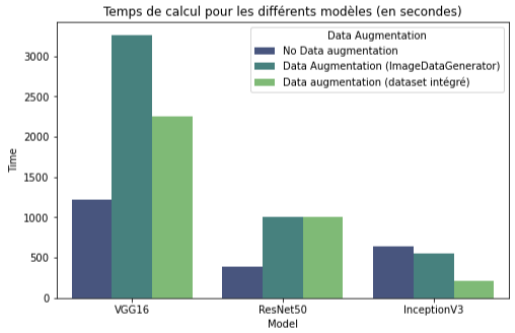
Classification image

Data augmentation

- ImageDataGenerator : augmentations de données images en les modifiant aléatoirement « à la volée »
- Approche utilisant des couches de pré-traitement : transformations appliquées sur toutes les images

Résultats



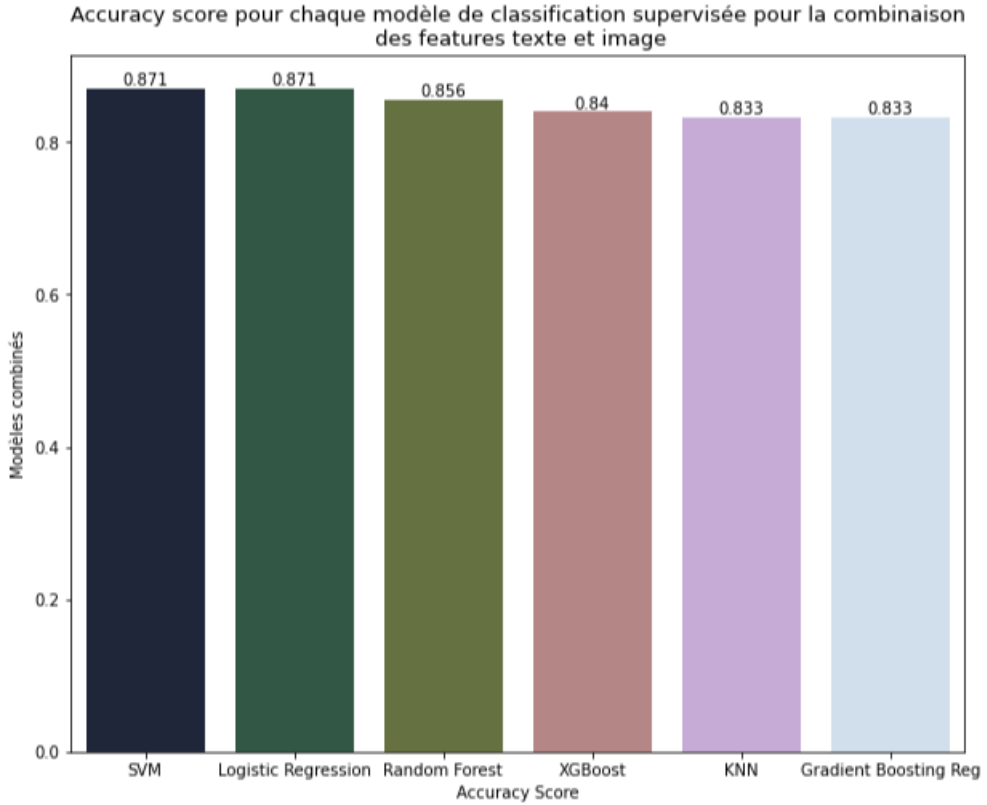


- Pas d'amélioration observée après data augmentation
- Valeurs d'accuracy altérées après data augmentation pour ResNet50
- Meilleurs résultats atteints pour les modèles sans data augmentation
- Temps de calcul élevés pour VGG16
- Temps de calcul augmentés après data augmentation pour VGG16 et ResNet50, mais diminués pour InceptionV3

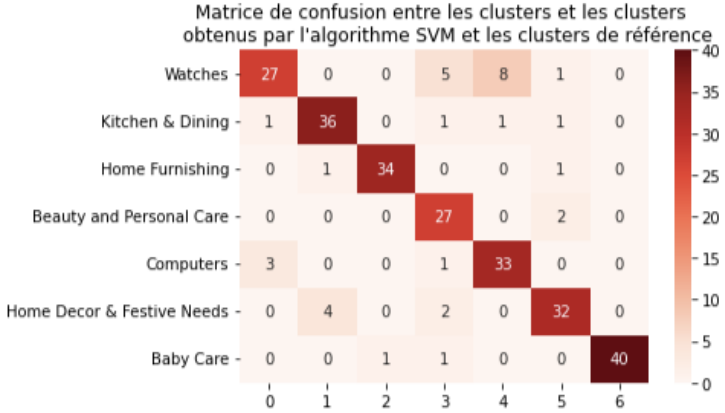
Résultats classification supervisée

Classification texte + image

Combinaison features extraites par BERT (texte) et ResNet50 (images)



- Meilleurs scores accuracy obtenus : SVM et Logistic Regression



- Catégories bien prédites : Baby Care, Beauty and Personal Care, Home Furnishing
- Catégories moins bien prédites : Watches, Home Decor & Festive Needs

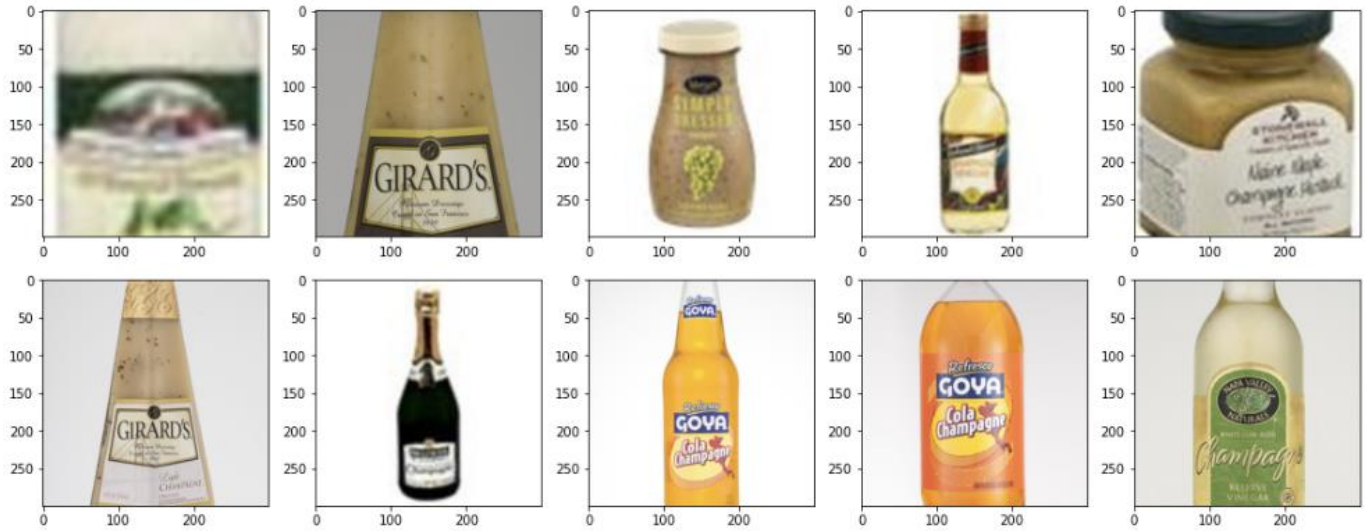
Test de l'API

API Application Programming Interface

Programme permettant à des applications de communiquer et échanger ensemble

Objectif : élargir la gamme de produits dans **l'épicerie fine** et collecte de produits à base de **champagne.**

Récupération des 10 premiers produits à base de champagne afin d'offrir aux clients un plus large choix de produits.



Conclusion

Intérêt(s) d'automatiser l'attribution de la catégorie d'un nouvel article à partir d'une description et/ou une image

Kitchen & Dining

Description de l'article : Buy NIKsales 7 W LED Bulb for Rs.365 online. NIKsales 7 W LED Bulb at best prices with FREE shipping & cash on delivery. Only Genuine Products. 30 Day Replacement Guarantee.

- Gain de temps
- Meilleure précision
- Expérience acheteur améliorée
- Moins de risque d'erreur
- Grand nombre de produits
- Expérience vendeur facilitée