Classifying consumer goods automatically

MAY 2022

NAJWA MOULINE

Presentation Outline

1. Objectives
2. Dataset Preparation
3. Pre-processing and Clustering
4. Classification Feasibility

Objectives

Context

- Company interested to launch an e-commerce marketplace
- Sellers offer items to buyers by posting a photo and description
- Assignment of item categories manually by sellers – tedious and unreliable

Business Problem

- Improving user experience for sellers and buyers through reliable item classification
- Automating the assignment of the category of products for sale

Mission

- Perform a feasibility study of an image or description-based item classification engine to automate:
  - Pre-processing of the dataset (textual and visual)
  - Dimension reduction
  - Clustering and representation in 2D

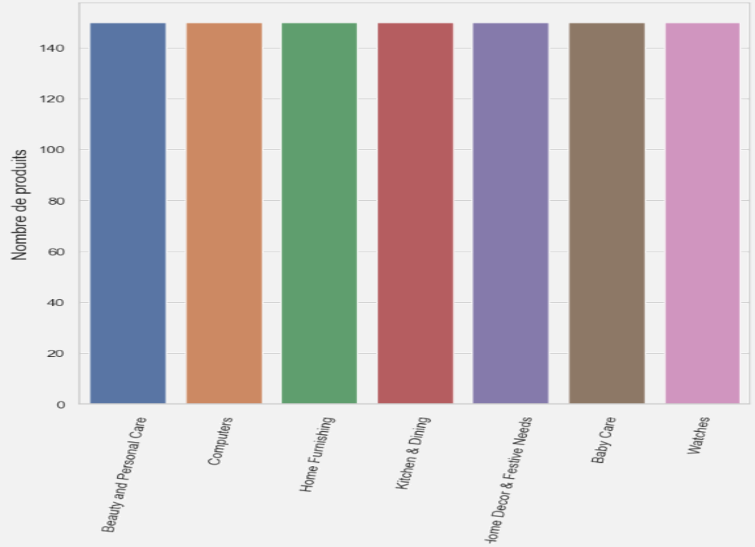
Dataset Preparation

Dataset

Dataset

Dataset	A csv format file and jpg images
Observations	1050 articles
Variables	15 relating to the identifier of the item, description, price, rating, image, ...
Duplicates	None
Filling	Completed for the data considered
Features	Description (text), image (visual)
Target	

- Category class 1 – 7 categories
- Category class 2 – 62 categories
- The dataset is good for our mission
- It is well distributed by category 1, which should make it possible to identify all categories with modeling

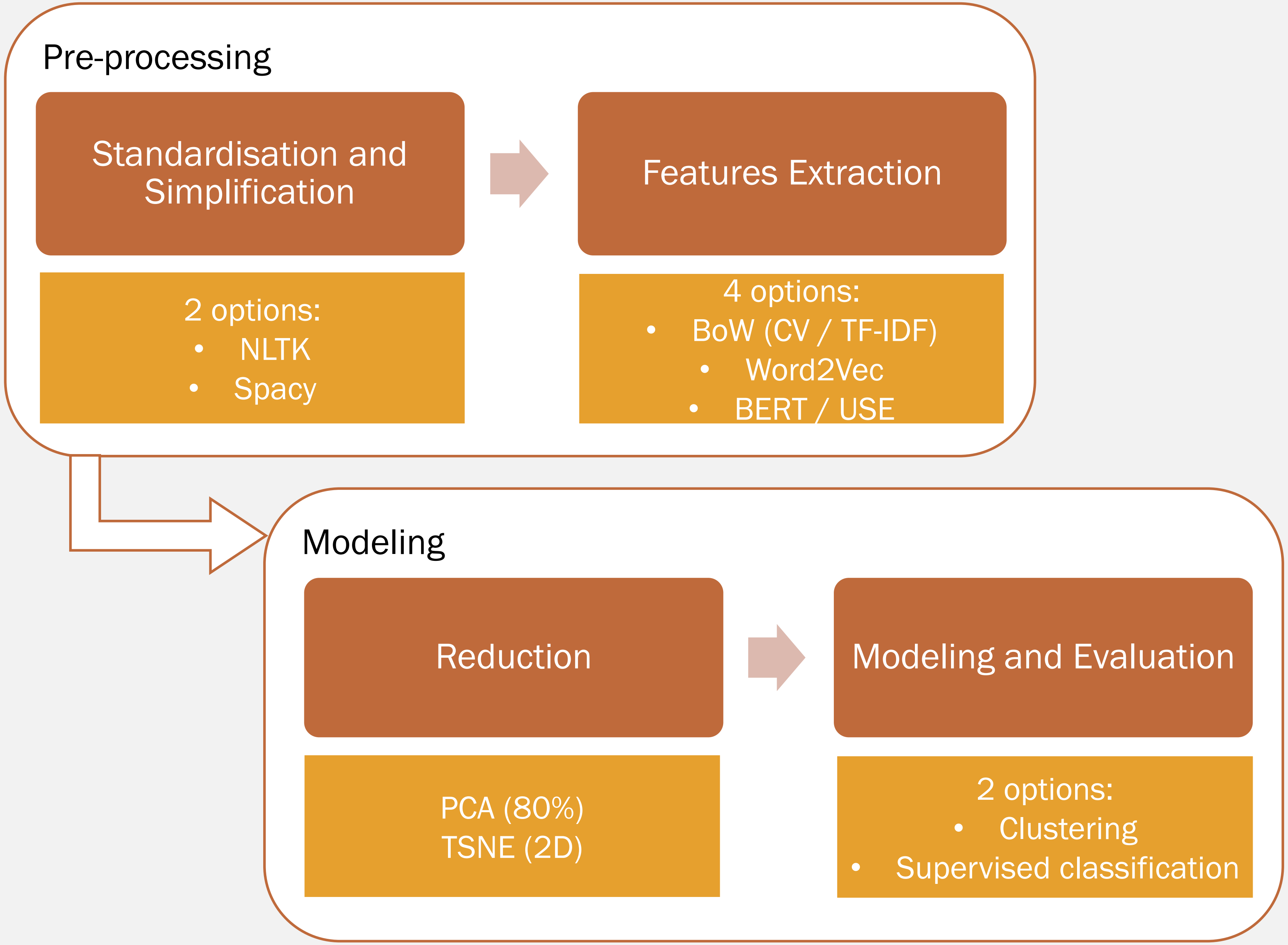


Pre-processing and clustering

Methodology

(textual data)

Methodology – classification of textual data



Pre-processing

Standardisation and Simplification

2 options:
- NLTK
- Spacy

Features Extraction

4 options:
- BoW (CV / TF-IDF)
- Word2Vec
- BERT / USE

Modeling

Reduction

PCA (80%)
TSNE (2D)

Modeling and Evaluation

2 options:
- Clustering
- Supervised classification

Pre-processing and clustering

Pre-processing

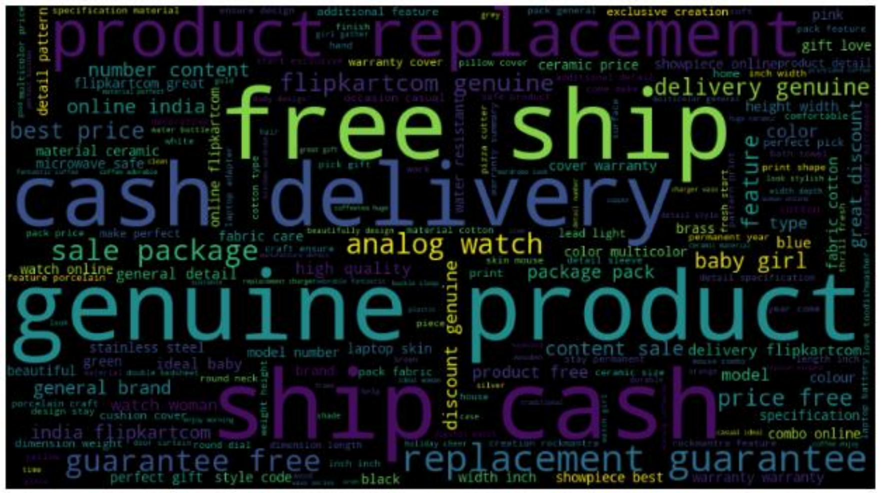
(textual data)

Standardisation and Simplification (NLTK) – from 9591 unique words to 2796

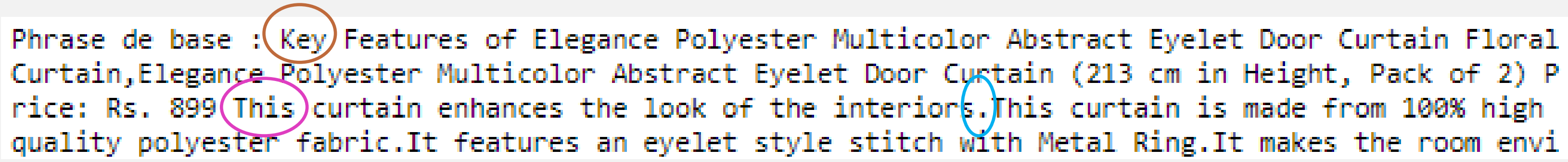
Steps

1. Switching to Lowercase
2. Removal of punctuation
3. Tokenisation
4. Removal of StopWords
5. Lemmatisation
6. Removal of words with less than 3 letters
7. Removal of words occurring once in the corpus

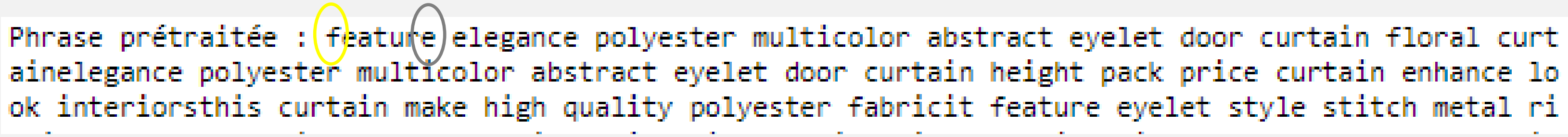
Resulting Worldcloud



Example



```
Phrase de base : 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) P
rice: 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 envi
```



```
Phrase prétraitée : feature elegance polyester multicolor abstract eyelet door curtain floral curt
ainelegance polyester multicolor abstract eyelet door curtain height pack price curtain enhance lo
ok interiorsthis curtain make high quality polyester fabricit feature eyelet style stitch metal ri
```

Pre-processing and clustering

Pre-processing

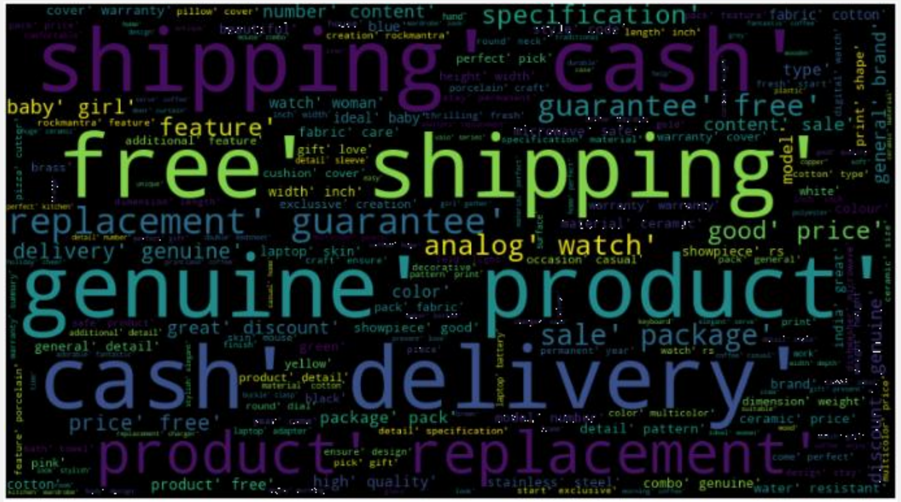
(textual data)

Standardisation and Simplification (Spacy) – from 9591 unique words to 4186

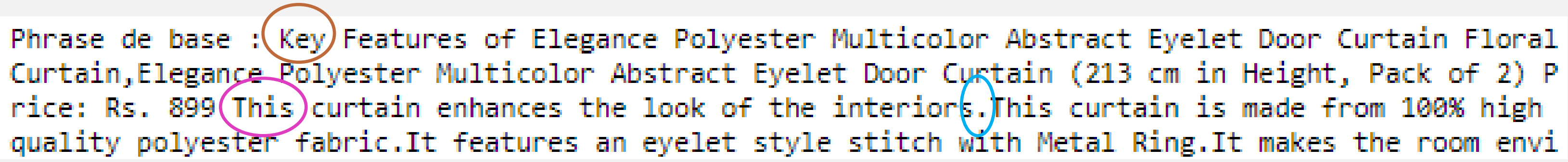
Steps

1. Switching to Lowercase
2. Tokenisation
3. Lemmatisation
4. Keeping words that are pronouns, adj, noun, and verb
5. Removal of StopWords
6. Removal of words with less than 3 letters
7. Removal of words occurring once in the corpus

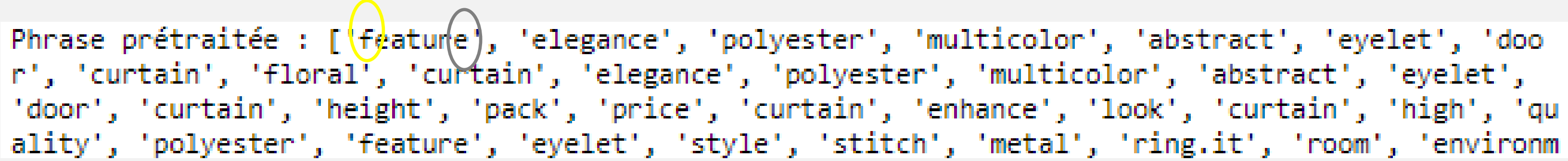
Resulting Worldcloud



Example



```
Phrase de base : 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) P
rice: 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 envi
```



```
Phrase prétraitée : ['feature', 'elegance', 'polyester', 'multicolor', 'abstract', 'eyelet', 'doo
r', 'curtain', 'floral', 'curtain', 'elegance', 'polyester', 'multicolor', 'abstract', 'eyelet',
'door', 'curtain', 'height', 'pack', 'price', 'curtain', 'enhance', 'look', 'curtain', 'high', 'qu
ality', 'polyester', 'feature', 'eyelet', 'style', 'stitch', 'metal', 'ring.it', 'room', 'environm
```

Pre-processing and clustering

Pre-processing

(textual data)

Features Extraction – 5 options considered

Bag of Words - CV

Vector formed from the occurrence of the word in the description.
Vector (1050, number of unique tokens)

Bag of Words – TF-IDF

Vector formed from the occurrence of the word in the description and in the corpus.
Vector (1050, number of unique tokens)

Word2Vec

Vector formed from the "meaning" of the word.
Vector (1050, 300)

BERT

Vector formed from the position and context of the word in the description (transfer learning). Vector (1050, 768)

USE

Vector formed from the position and context of the word in the description (transfer learning). Vector (1050, 512)

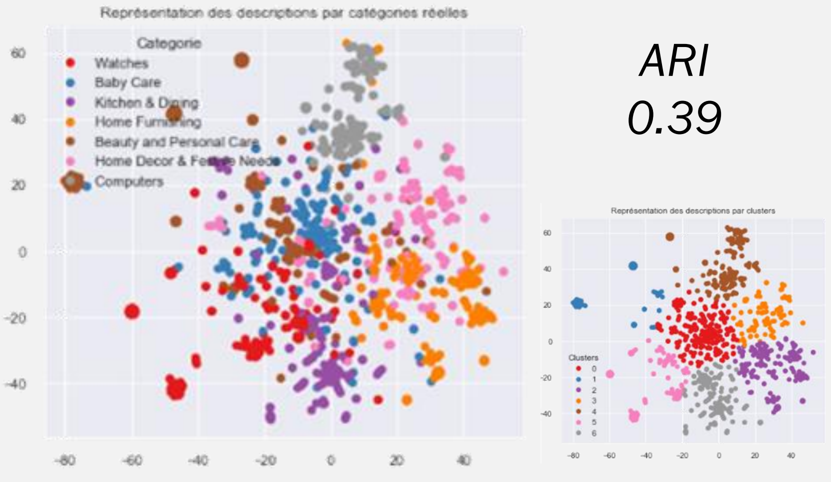
Pre-processing and clustering

Modeling

(textual data)

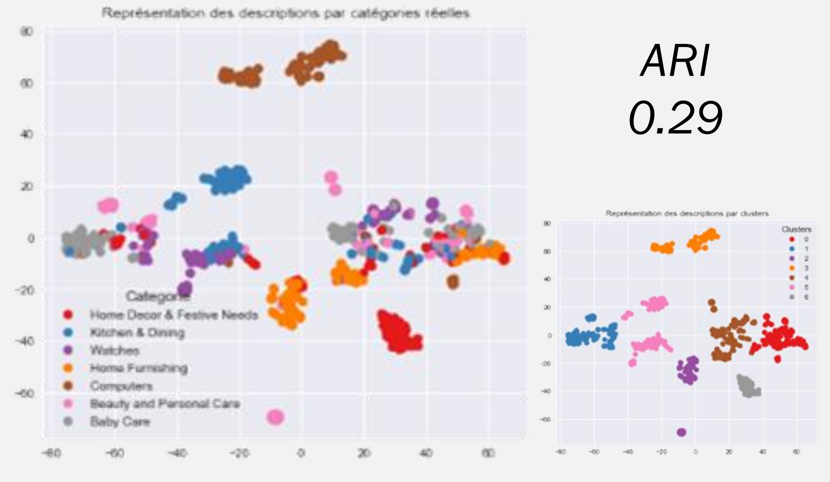
Unsupervised classification with KMeans

Visualisation NLTK / TFIDF



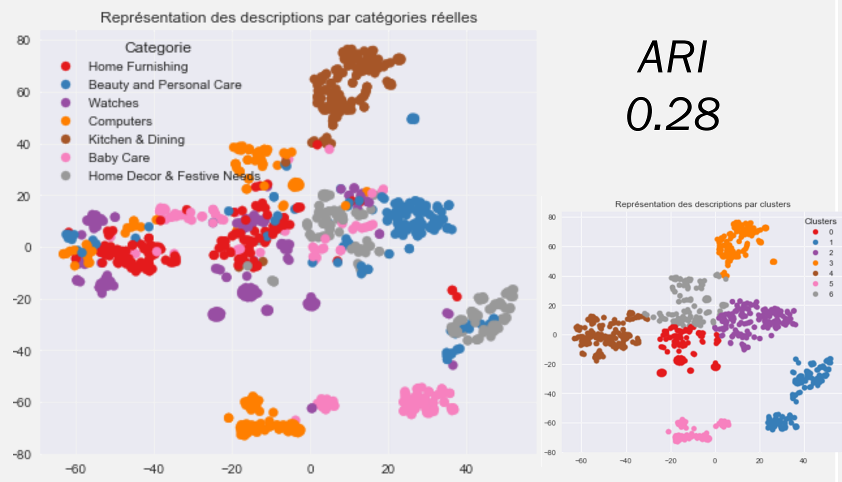
ARI 0.39

Visualisation Spacy / Word2Vec



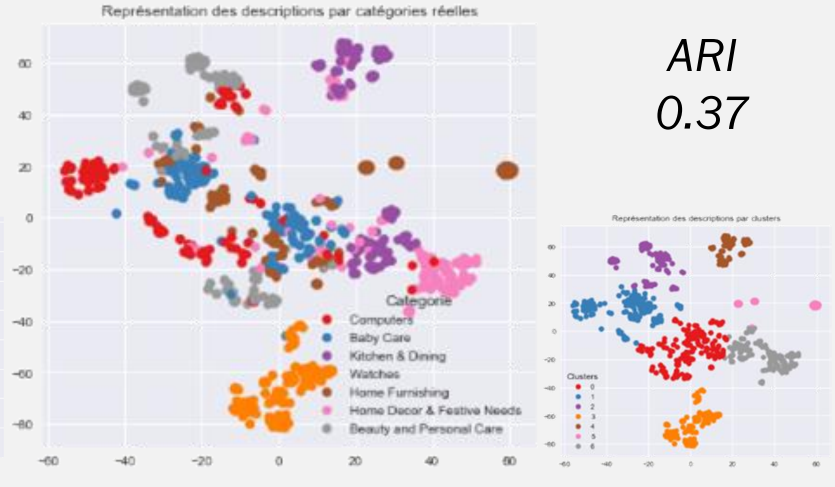
ARI 0.29

Visualisation Spacy / BERT



ARI 0.28

Visualisation NLTK / USE



ARI 0.37

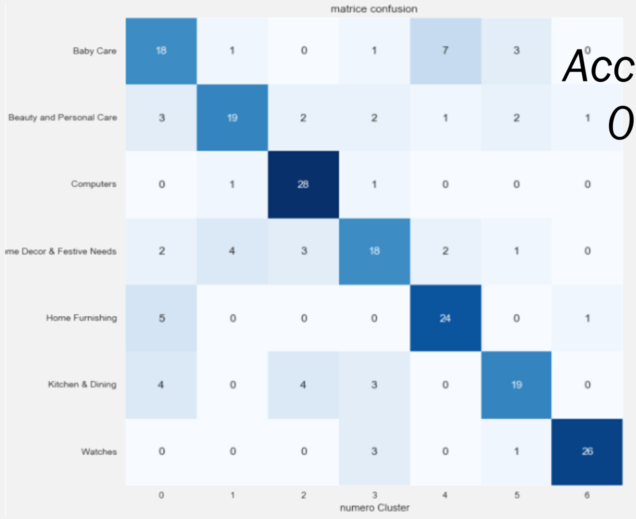
Pre-processing and clustering

Modeling

(textual data)

Supervised classification with KNN

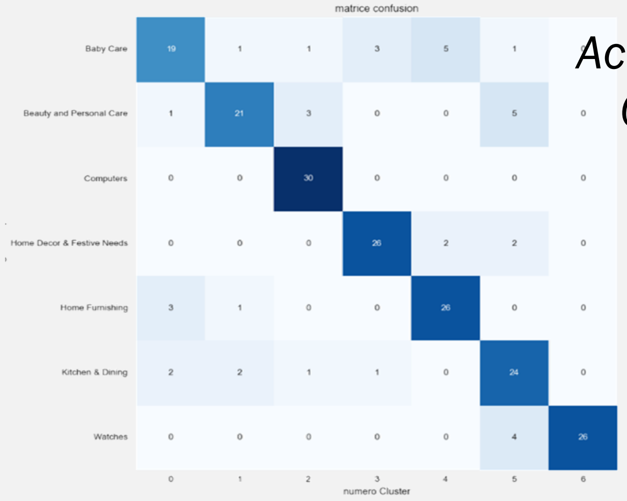
Visualisation NLTK / TFIDF



	0	1	2	3	4	5	6
Baby Care	18	1	0	1	7	3	0
Beauty and Personal Care	3	19	2	2	1	2	1
Computers	0	1	28	1	0	0	0
me Decor & Festive Needs	2	4	3	18	2	1	0
Home Furnishing	5	0	0	0	24	0	1
Kitchen & Dining	4	0	4	3	0	19	0
Watches	0	0	0	3	0	1	26

Accuracy 0.72

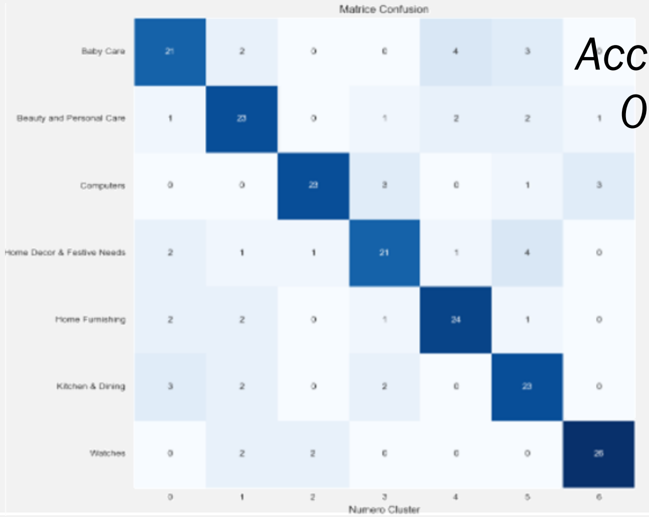
Visualisation Spacy / Word2Vec



	0	1	2	3	4	5	6
Baby Care	19	1	1	3	5	1	0
Beauty and Personal Care	1	21	3	0	0	5	0
Computers	0	0	30	0	0	0	0
Home Decor & Festive Needs	0	0	0	26	2	2	0
Home Furnishing	3	1	0	0	26	0	0
Kitchen & Dining	2	2	1	1	0	24	0
Watches	0	0	0	0	0	4	26

Accuracy 0.82

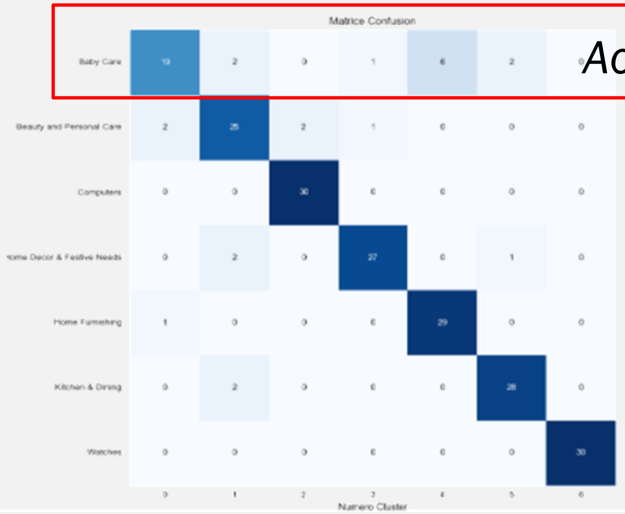
Visualisation Spacy / BERT



	0	1	2	3	4	5	6
Baby Care	21	2	0	0	4	3	0
Beauty and Personal Care	1	23	0	1	2	2	1
Computers	0	0	23	3	0	1	3
Home Decor & Festive Needs	2	1	1	21	1	4	0
Home Furnishing	2	2	0	1	24	1	0
Kitchen & Dining	3	2	0	2	0	23	0
Watches	0	2	2	0	0	0	26

Accuracy 0.77

Visualisation NLTK / USE



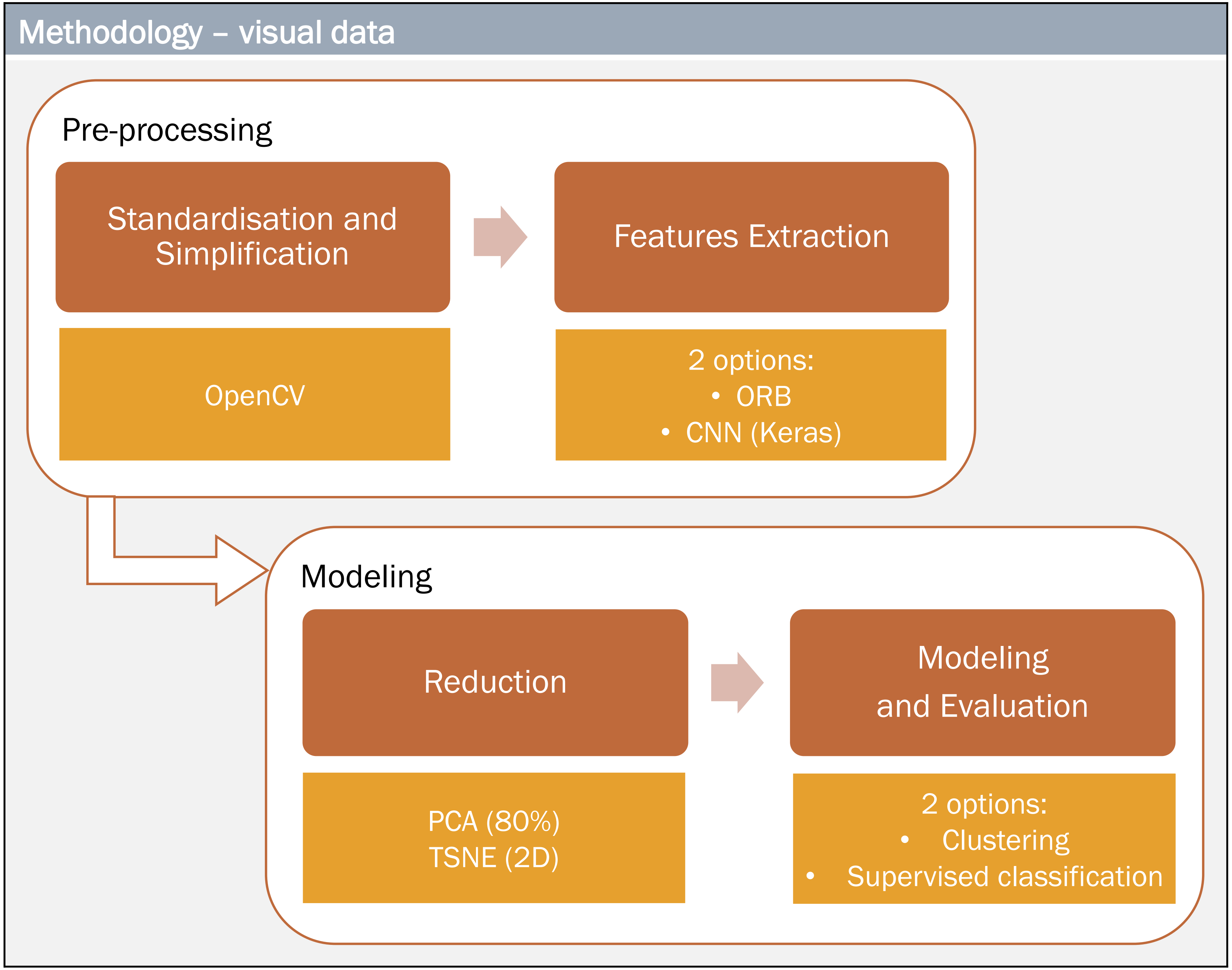
	0	1	2	3	4	5	6
Baby Care	19	2	0	1	6	2	0
Beauty and Personal Care	2	25	2	1	0	0	0
Computers	0	0	30	0	0	0	0
Home Decor & Festive Needs	0	2	0	27	0	1	0
Home Furnishing	1	0	0	0	29	0	0
Kitchen & Dining	0	2	0	0	0	28	0
Watches	0	0	0	0	0	0	30

Accuracy 0.90

Pre-processing and clustering

Methodology

(visual data)



Methodology – visual data

Pre-processing

Standardisation and Simplification

OpenCV

Features Extraction

2 options:
- ORB
- CNN (Keras)

Modeling

Reduction

PCA (80%)
TSNE (2D)

Modeling and Evaluation

2 options:
- Clustering
- Supervised classification

Pre-processing and clustering

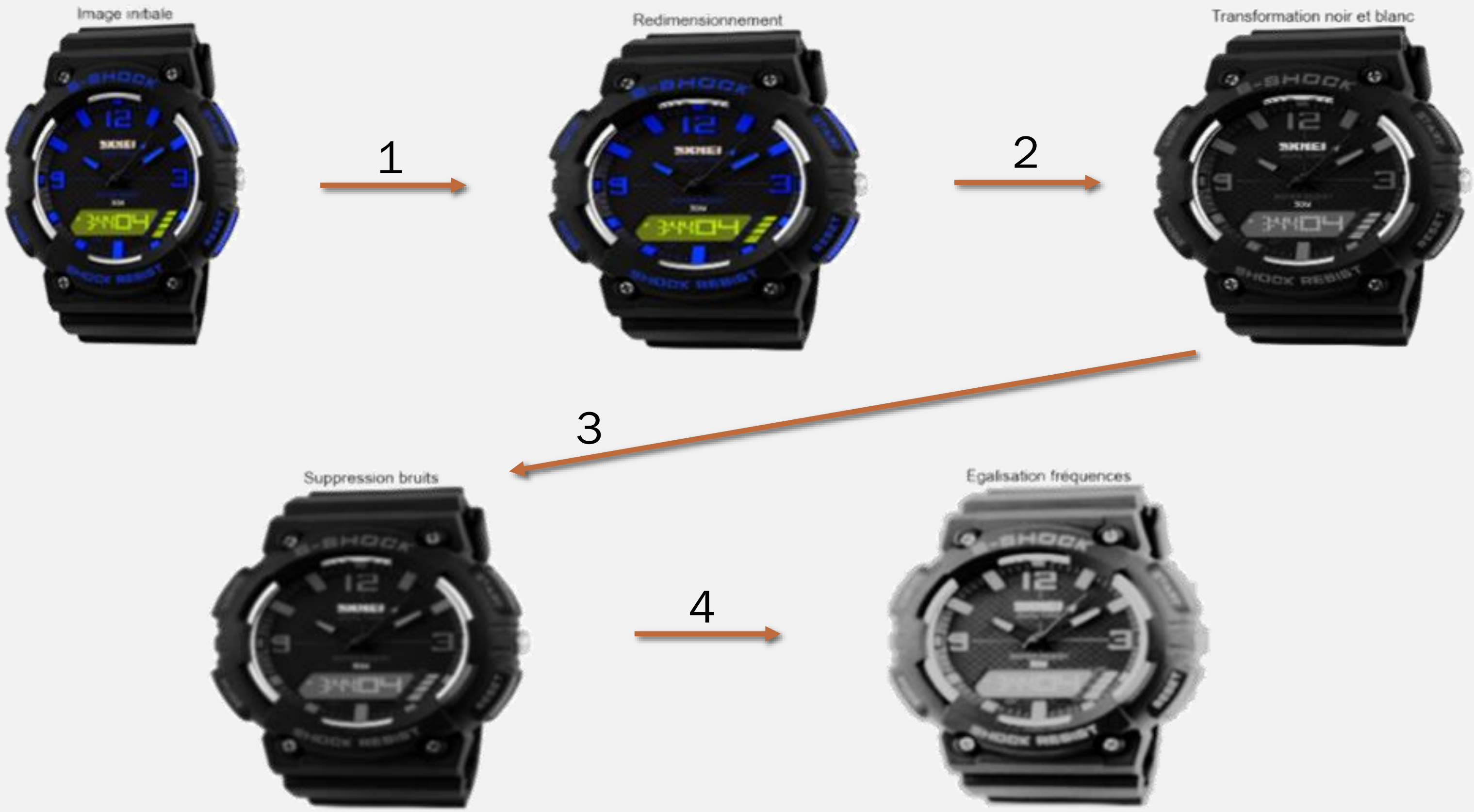
Pre-processing

(visual data)

Standardisation and Simplification (OpenCV)

Steps

1. Image redimensioning
2. Transformation to black and white
3. Noise cancellation
4. Frequencies equalisation



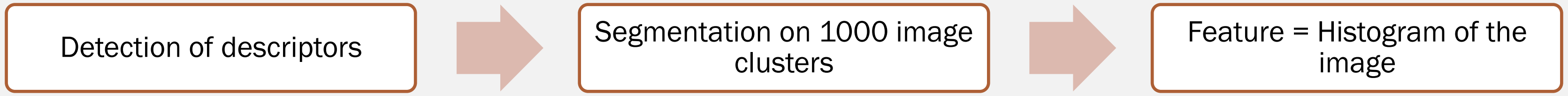
Pre-processing and clustering

Pre-processing

(visual data)

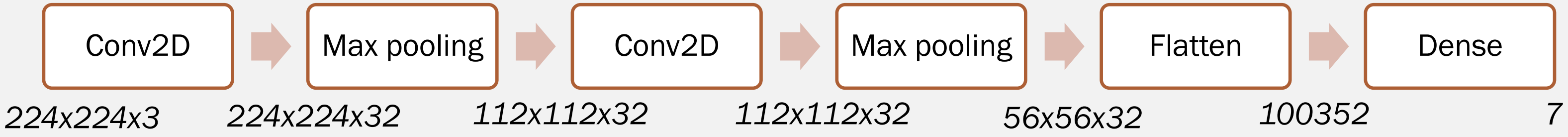
Features Extraction

ORB



Detection of descriptors → Segmentation on 1000 image clusters → Feature = Histogram of the image

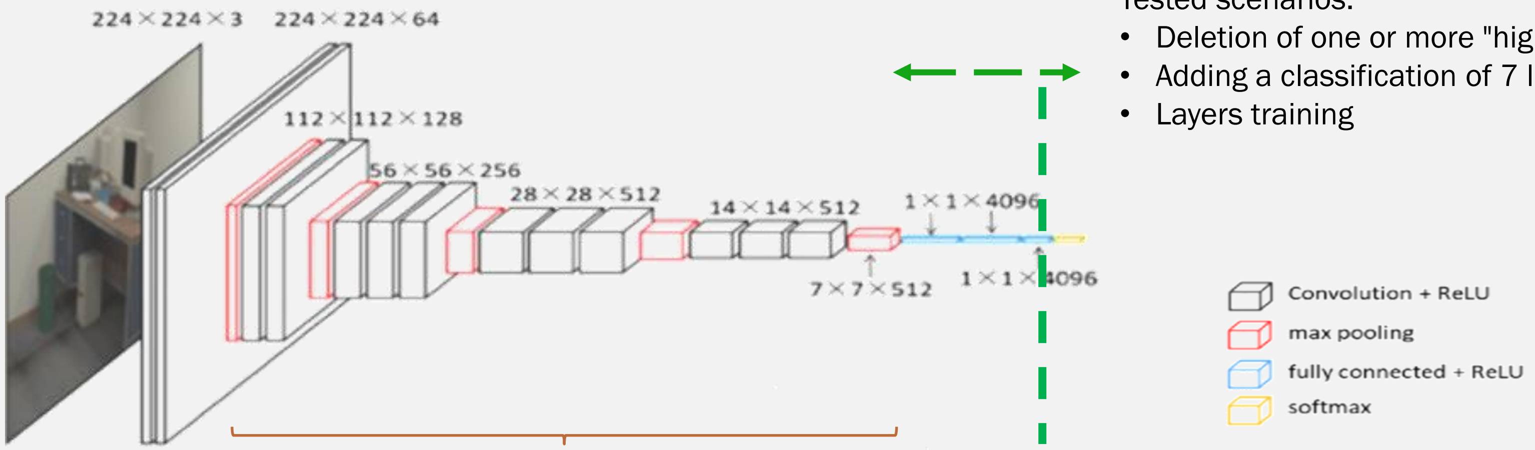
CCN Simple



Conv2D → Max pooling → Conv2D → Max pooling → Flatten → Dense

224x224x3 224x224x32 112x112x32 112x112x32 56x56x32 100352 7

CCN transfer learning - VGG16 model



Convolutional base: pre-trained on ImageNet

Tested scenarios:

- Deletion of one or more "high" layers
- Adding a classification of 7 layers
- Layers training

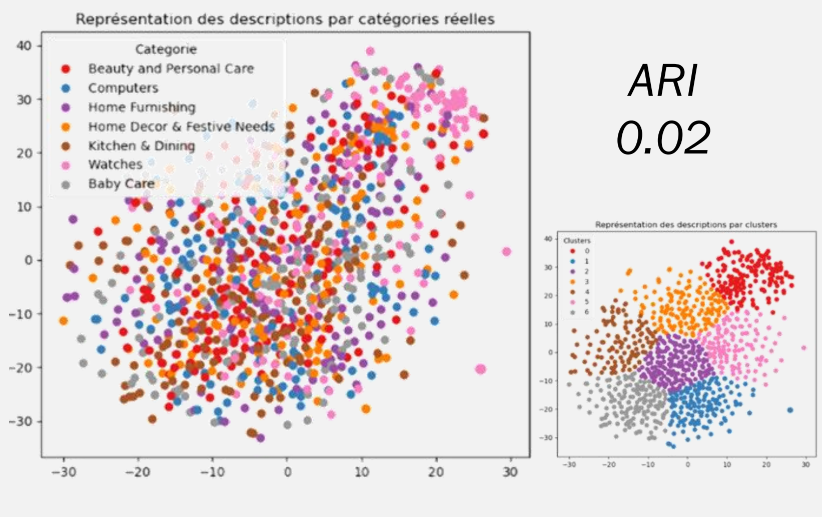
Pre-processing and clustering

Modeling

(visual data)

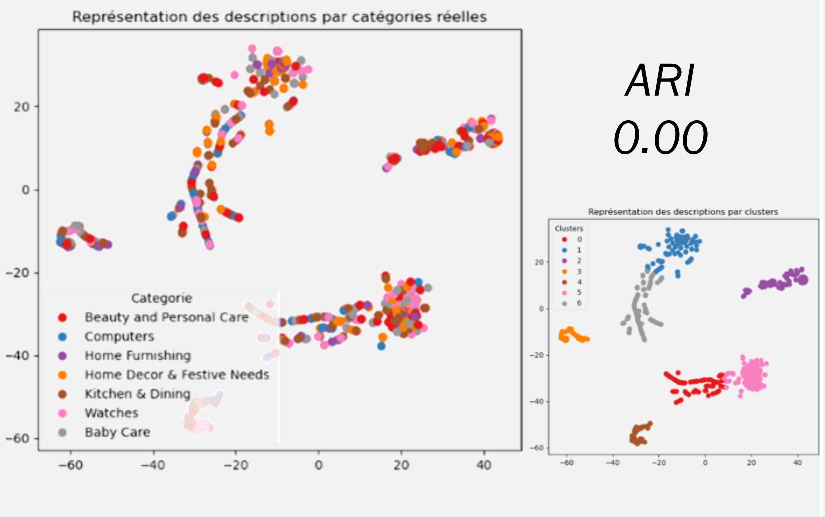
Unsupervised classification with KMeans

Visualisation *ORB*



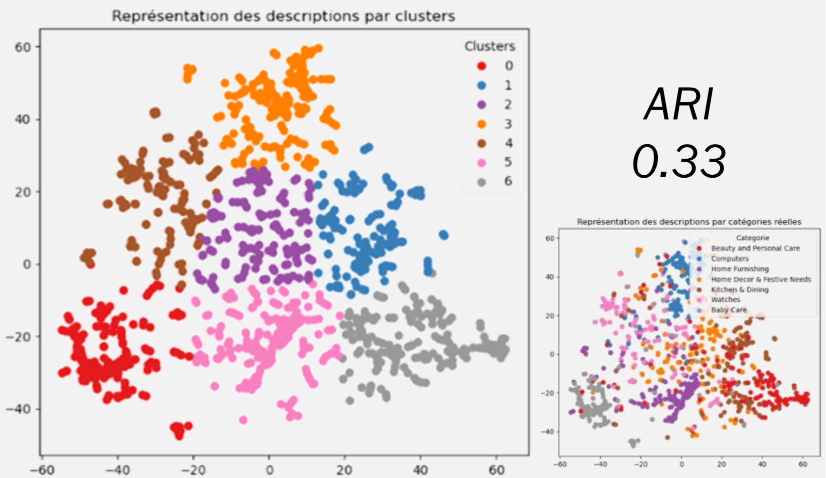
ARI
0.02

Visualisation CNN Simple



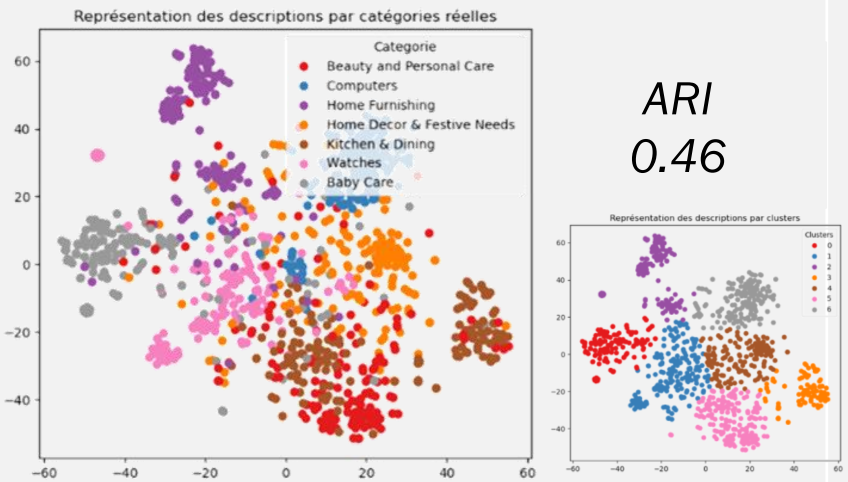
ARI
0.00

Visualisation CNN VGG16



ARI
0.33

Visualisation CNN VGG16 (without Dense)



ARI
0.46

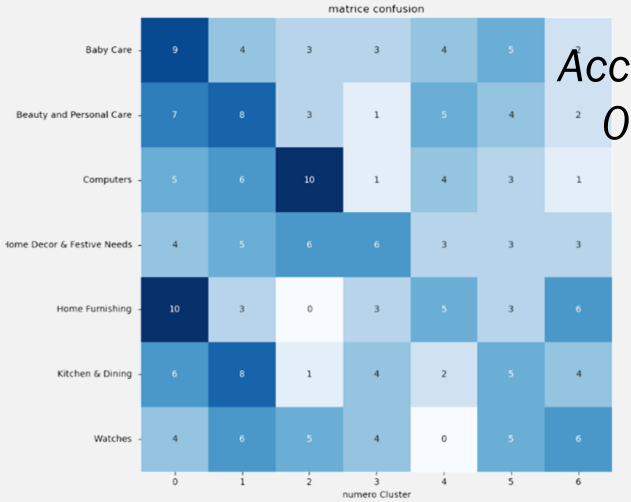
Pre-processing and clustering

Modeling

(visual data)

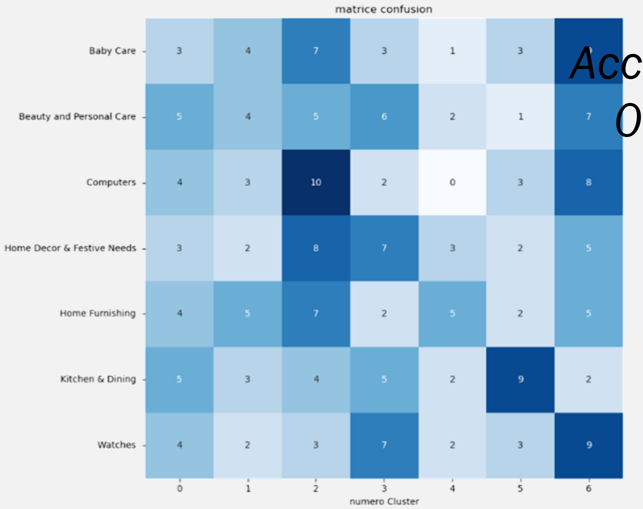
Supervised classification with KNN

Visualisation *ORB*



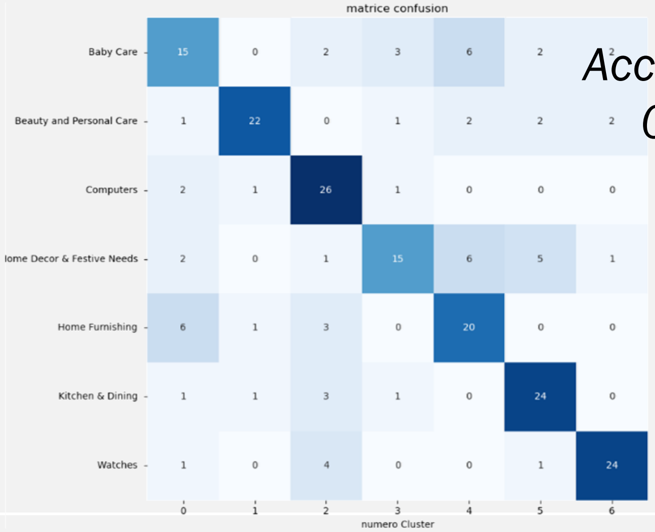
Accuracy 0.23

Visualisation CNN Simple



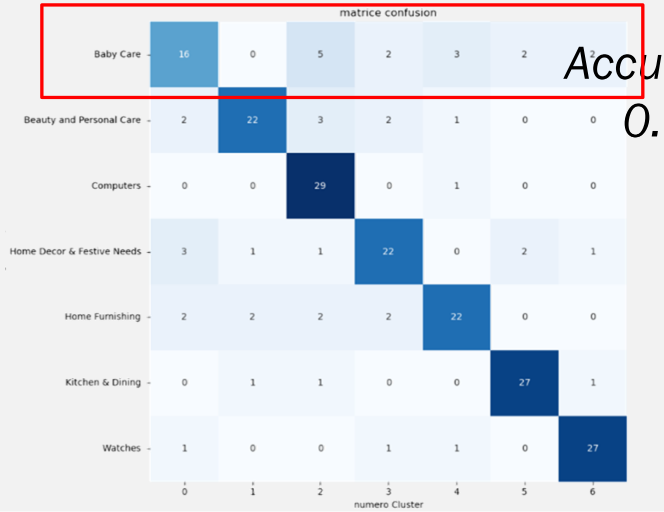
Accuracy 0.22

Visualisation CNN VGG16



Accuracy 0.7

Visualisation CNN VGG16 (without Dense)



Accuracy 0.8

Classification Feasibility

Feasibility and Conclusions

Classification engine feasibility

Feasibility study – unsupervised classification

- Clustering performed by the Kmeans model with 7 clusters
- In some cases, clusters spaced and evenly distributed
- But low ARIs

	ARI
Description - NLTK / CV	0.33
Description - NLTK / TFIDF	0.39
Description - NLTK / Word2Vec	0.28
Description - NLTK / BERT	0.32
Description - NLTK / USE	0.37

	ARI
Image - ORB	0.02
Image – CNN Simple	0.01
Image – CNN VGG16	0.33
Image – CNN VGG16 (Dense 7)	0.03
Image – CNN VGG16 (without Dense)	0.46

=> Inconclusive feasibility study

Classification Feasibility

Feasibility and Conclusions

Classification engine feasibility

Feasibility study – supervised classification

- Prediction of the classification issued from the KNN model trained on the 7 categories
- For the majority of cases, clusters spaced and distributed equally
- With right accuracy for all categories except BabyCare

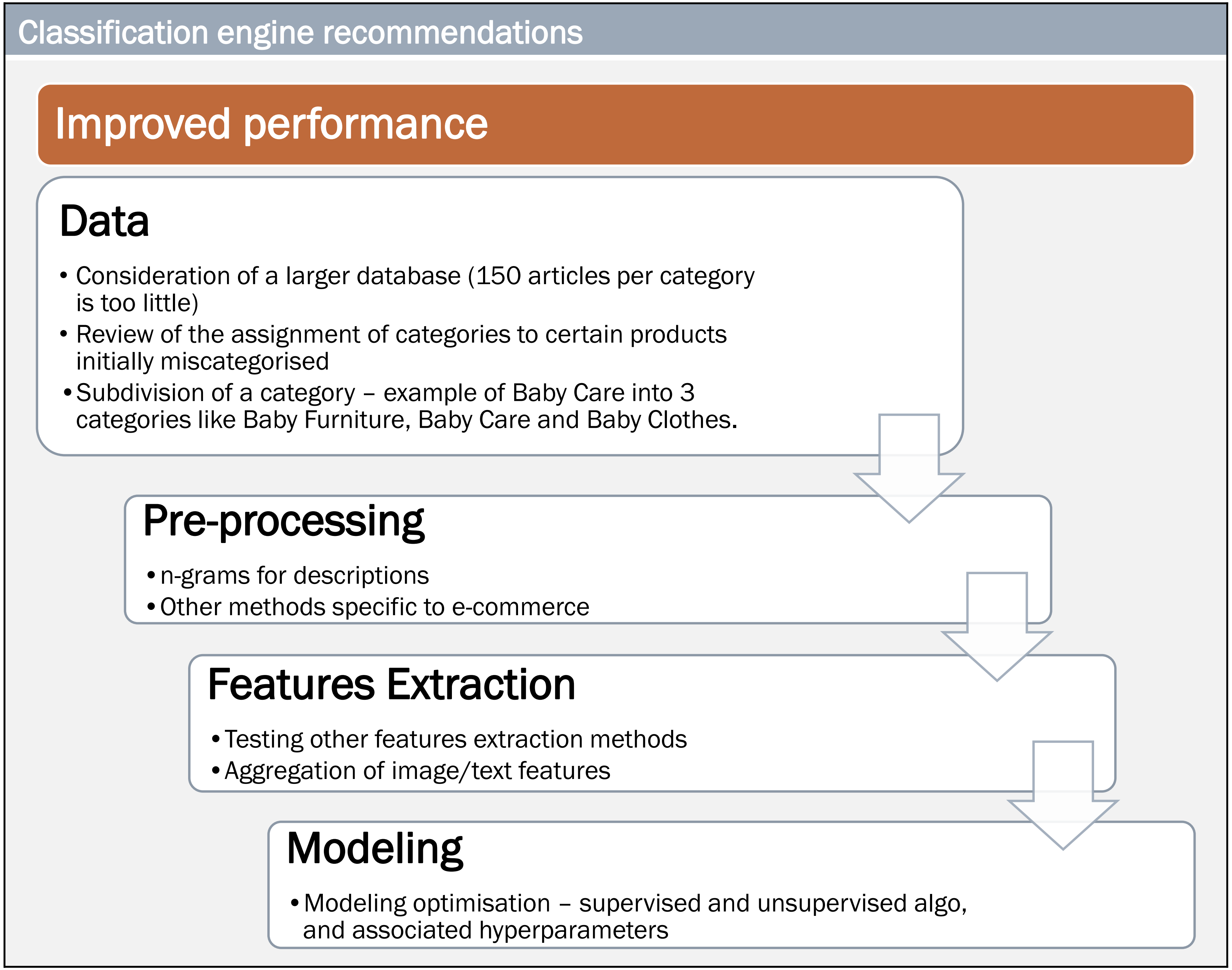
	Acc
Description - NLTK / CV	0.74
Description - Spacy / TFIDF	0.75
Description - Spacy / Word2Vec	0.82
Description - NLTK / BERT	0.84
Description - NLTK / USE	0.90

	Acc
Image - ORB	0.23
Image – CNN Simple	0.20
Image – CNN VGG16	0.70
Image – CNN VGG16 (Dense with 7)	0.27
Image – CNN VGG16 (without Dense)	0.79

=> Conclusive feasibility study

Classification Feasibility

Recommendations



Classification engine recommendations

Improved performance

Data

- Consideration of a larger database (150 articles per category is too little)
- Review of the assignment of categories to certain products initially miscategorised
- Subdivision of a category – example of Baby Care into 3 categories like Baby Furniture, Baby Care and Baby Clothes.

Pre-processing

- n-grams for descriptions
- Other methods specific to e-commerce

Features Extraction

- Testing other features extraction methods
- Aggregation of image/text features

Modeling

- Modeling optimisation – supervised and unsupervised algo, and associated hyperparameters

Thank you for your attention!