P6 : CLASSIFIEZ AUTOMATIQUEMENT DES BIENS DE CONSOMMATION

October 5, 2021

Carlos SACRISTAN

OpenClassrooms

Mission objective

place de marché

Context :

1. L'entreprise "Place de marché" souhaite lancer une marketplace e-commerce
2. Des vendeurs proposent des articles à des acheteurs en postant une photo et une description

Objective :

1. A partir de la photo et de la description, réaliser une première étude de faisabilité d'un moteur de **classification**

classification

Table des matières

Index

Data cleaning

place de marché

1 Basic cleaning

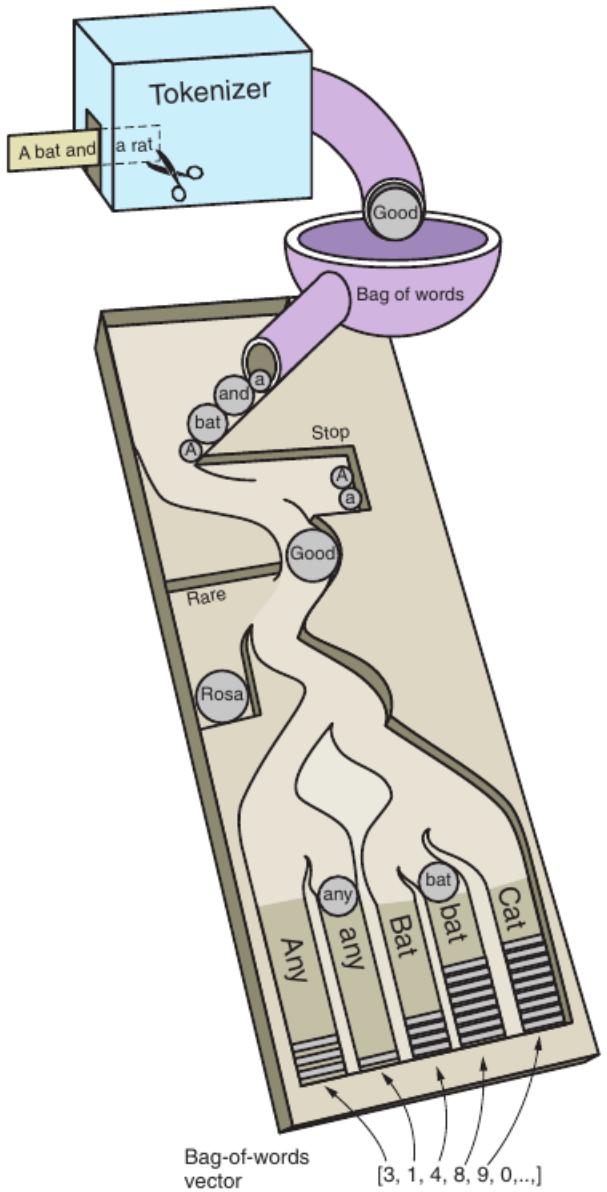
- Import 1 csv (1050 rows $\times$ 14 columns)
- Lowercases of features name
- Convert features to best possible dtypes
- Split : product_category_tree (>>) $\rightarrow$ Level 0 $\rightarrow$ **TARGET**
- Delete of duplicate values
- Delete of missing values

Index

Pre-processing

place de marché

1. **Preprocessing pipeline**
   - **Tokenization**
   - **StopWords**
   - **Lemmatization**
   - **Stemming**
1. Bag of words
   - Singel-gram
   - Range-Gram
1. TF-IDF
   - Singel-gram
   - Range-Gram
1. T-SNE



Pre-processing

place de marché

1. Preprocessing pipeline
   - **Tokenization**
   - StopWords
   - Lemmatization
   - Stemming
1. Bag of words
   - Singel-gram
   - Range-Gram
1. TF-IDF
   - Singel-gram
   - Range-Gram
1. T-SNE

```
ORIGINAL
specifications of sathiyas cotton bath towel _3 bath towel, red, y
ellow, blue_ bath towel features machine washable yes material cot
ton design self design general brand sathiyas type bath towel gsm
500 model name sathiyas cotton bath towel ideal for men, women, bo
ys, girls model id asvtwl322 color red, yellow, blue size mediam d
imensions length 30 inch width 60 inch in the box number of conten
ts in sales package 3 sales package 3 bath towel
TOKENIZATION
specifications of sathiyas cotton bath towel  bath towel  red  ye
llow  blue  bath towel features machine washable yes material cott
on design self design general brand sathiyas type bath towel gsm
model name sathiyas cotton bath towel ideal for men  women  boys
girls model id asvtwl color red  yellow  blue size mediam dimensio
ns length  inch width  inch in the box number of contents in sales
package  sales package  bath towel
```

Removal of special characters/punctuations & case folding

Pre-processing

place de marché

1. Preprocessing pipeline
   - Tokenization
   - **StopWords**
   - Lemmatization
   - Stemming
1. Bag of words
   - Singel-gram
   - Range-Gram
1. TF-IDF
   - Singel-gram
   - Range-Gram
1. T-SNE

```
TOKENIZATION
specifications of sathiyas cotton bath towel   bath towel  red  ye
llow  blue  bath towel features machine washable yes material cott
on design self design general brand sathiyas type bath towel gsm
model name sathiyas cotton bath towel ideal for men  women  boys
girls model id asvtwl color red  yellow  blue size mediam dimensio
ns length  inch width  inch in the box number of contents in sales
package  sales package  bath towel
REMOVE_STOPWORDS
[['specifications', 'sathiyas', 'cotton', 'bath', 'towel', 'bath',
'towel', 'red', 'yellow', 'blue', 'bath', 'towel', 'features', 'ma
chine', 'washable', 'yes', 'material', 'cotton', 'design', 'self',
'design', 'general', 'brand', 'sathiyas', 'type', 'bath', 'towel',
'gsm', 'model', 'name', 'sathiyas', 'cotton', 'bath', 'towel', 'id
eal', 'men', 'women', 'boys', 'girls', 'model', 'id', 'asvtwl', 'c
olor', 'red', 'yellow', 'blue', 'size', 'mediam', 'dimensions', 'l
ength', 'inch', 'width', 'inch', 'box', 'number', 'contents', 'sal
es', 'package', 'sales', 'package', 'bath', 'towel']]
```

Stop words : 'who', 'what', 'when', 'why', 'how', 'which', 'where', 'whom'

Pre-processing

place de marché

1. Preprocessing pipeline
   - Tokenization
   - StopWords
   - **Lemmatization**
   - Stemming
1. Bag of words
   - Singel-gram
   - Range-Gram
1. TF-IDF
   - Singel-gram
   - Range-Gram
1. T-SNE

```
REMOVE_STOPWORDS
[['specifications', 'sathiyas', 'cotton', 'bath', 'towel', 'bath',
'towel', 'red', 'yellow', 'blue', 'bath', 'towel', 'features', 'ma
chine', 'washable', 'yes', 'material', 'cotton', 'design', 'self',
'design', 'general', 'brand', 'sathiyas', 'type', 'bath', 'towel',
'gsm', 'model', 'name', 'sathiyas', 'cotton', 'bath', 'towel', 'id
eal', 'men', 'women', 'boys', 'girls', 'model', 'id', 'asvtwl', 'c
olor', 'red', 'yellow', 'blue', 'size', 'mediam', 'dimensions', 'l
ength', 'inch', 'width', 'inch', 'box', 'number', 'contents', 'sal
es', 'package', 'sales', 'package', 'bath', 'towel']]
LEMMATIZATION
[['specifications', 'sathiyas', 'cotton', 'bath', 'towel', 'bath',
'towel', 'red', 'yellow', 'blue', 'bath', 'towel', 'feature', 'mac
hine', 'washable', 'yes', 'material', 'cotton', 'design', 'self',
'design', 'general', 'brand', 'sathiyas', 'type', 'bath', 'towel',
'gsm', 'model', 'name', 'sathiyas', 'cotton', 'bath', 'towel', 'id
eal', 'men', 'women', 'boys', 'girls', 'model', 'id', 'asvtwl', 'c
olor', 'red', 'yellow', 'blue', 'size', 'mediam', 'dimension', 'le
ngth', 'inch', 'width', 'inch', 'box', 'number', 'content', 'sales
', 'package', 'sales', 'package', 'bath', 'towel']]
```

Ex: computer, computerization or computerize $\longrightarrow$ compute

Pre-processing

place de marché

1. Preprocessing pipeline
   - Tokenization
   - StopWords
   - Lemmatization
   - **Stemming**
1. Bag of words
   - Singel-gram
   - Range-Gram
1. TF-IDF
   - Singel-gram
   - Range-Gram
1. T-SNE

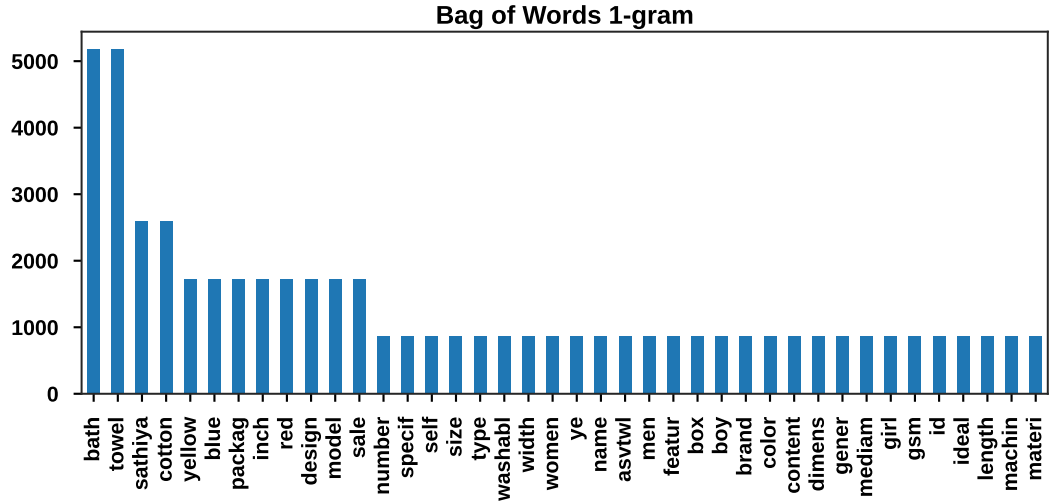
```
LEMMATIZATION
[['specifications', 'sathiyas', 'cotton', 'bath', 'towel', 'bath',
'towel', 'red', 'yellow', 'blue', 'bath', 'towel', 'feature', 'mac
hine', 'washable', 'yes', 'material', 'cotton', 'design', 'self',
'design', 'general', 'brand', 'sathiyas', 'type', 'bath', 'towel',
'gsm', 'model', 'name', 'sathiyas', 'cotton', 'bath', 'towel', 'id
eal', 'men', 'women', 'boys', 'girls', 'model', 'id', 'asvtwl', 'c
olor', 'red', 'yellow', 'blue', 'size', 'mediam', 'dimension', 'le
ngth', 'inch', 'width', 'inch', 'box', 'number', 'content', 'sales
', 'package', 'sales', 'package', 'bath', 'towel']]
STEMMING
[['specif', 'sathiya', 'cotton', 'bath', 'towel', 'bath', 'towel',
'red', 'yellow', 'blue', 'bath', 'towel', 'featur', 'machin', 'was
habl', 'ye', 'materi', 'cotton', 'design', 'self', 'design', 'gene
r', 'brand', 'sathiya', 'type', 'bath', 'towel', 'gsm', 'model', '
name', 'sathiya', 'cotton', 'bath', 'towel', 'ideal', 'men', 'wome
n', 'boy', 'girl', 'model', 'id', 'asvtwl', 'color', 'red', 'yello
w', 'blue', 'size', 'mediam', 'dimens', 'length', 'inch', 'width',
'inch', 'box', 'number', 'content', 'sale', 'packag', 'sale', 'pac
kag', 'bath', 'towel']]
```

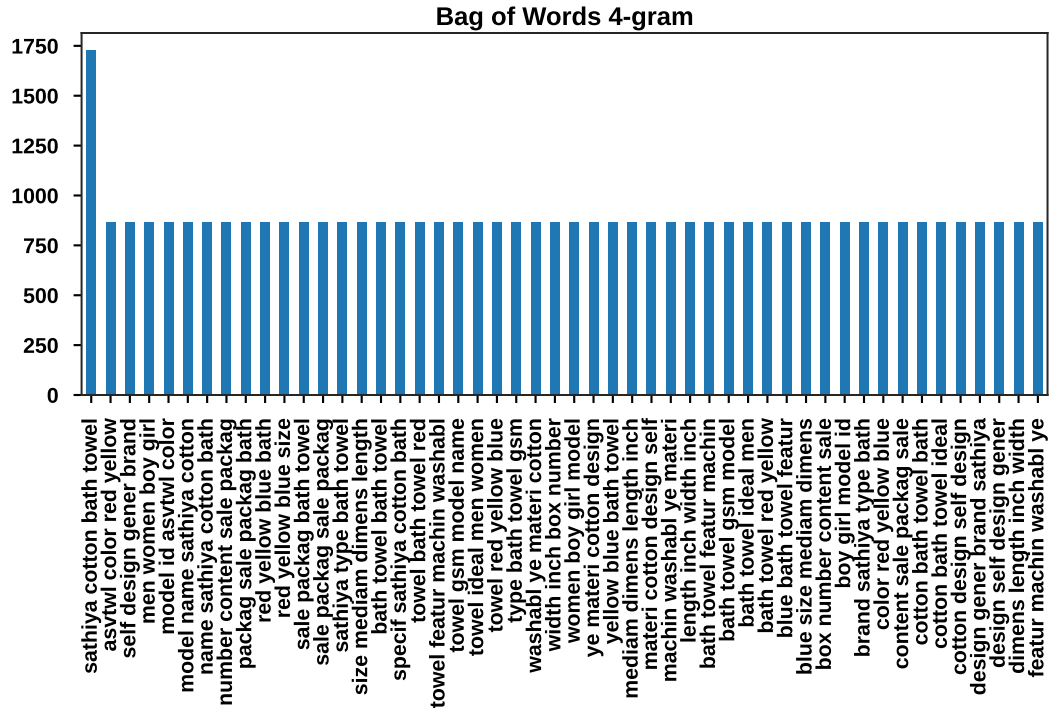
To convert a word to its meaningful base form by removing few characters

Pre-processing

place de marché

1. Preprocessing pipeline
   - Tokenization
   - StopWords
   - Lemmatization
   - Stemming
1. **Bag of words**
   - **Singel-gram**
   - **Range-Gram**
1. TF-IDF
   - Singel-gram
   - Range-Gram
1. T-SNE

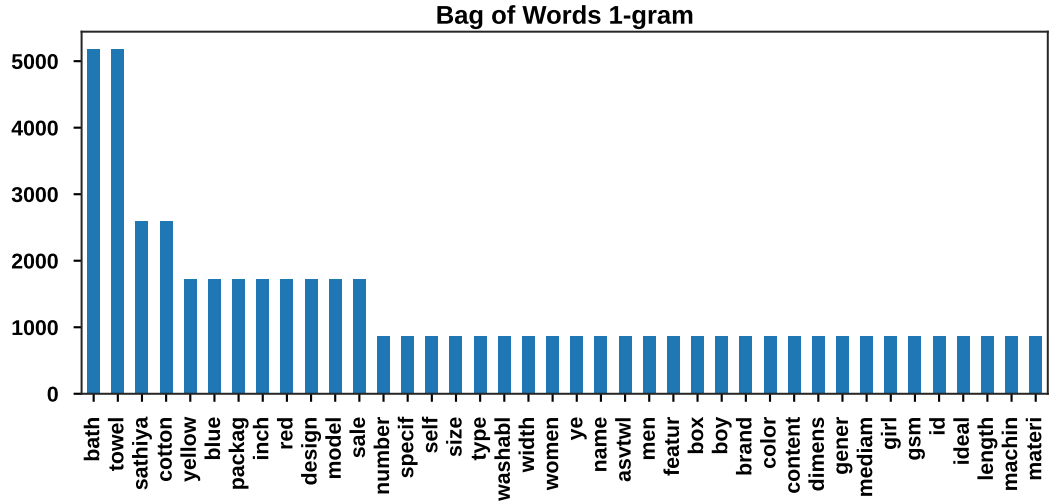


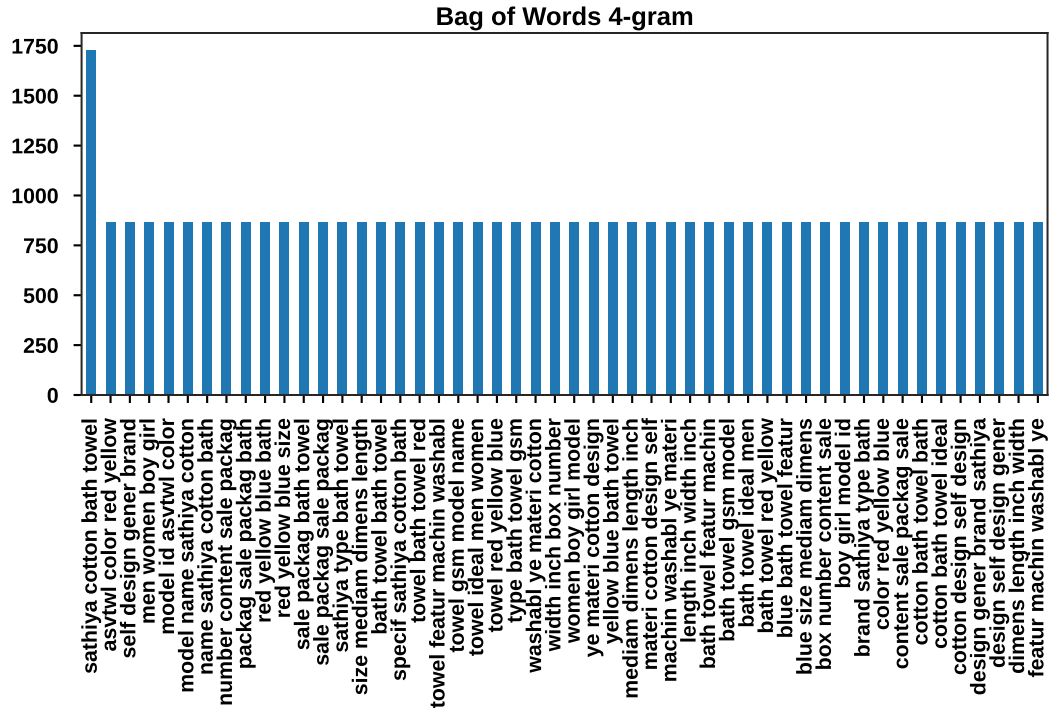


Pre-processing

place de marché

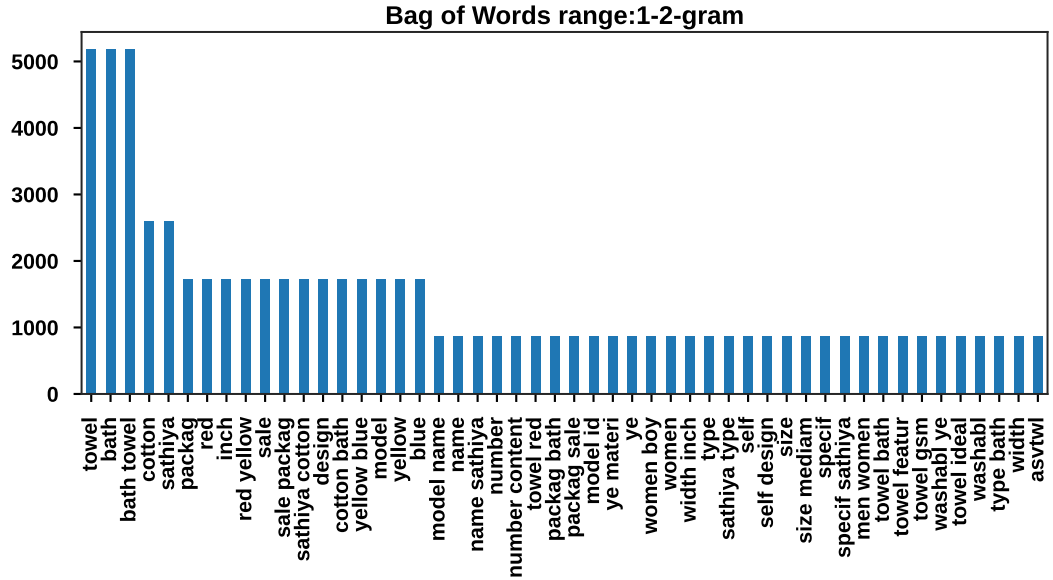
1. Preprocessing pipeline
   - Tokenization
   - StopWords
   - Lemmatization
   - Stemming
1. Bag of words
   - **Singel-gram**
   - Range-Gram
1. TF-IDF
   - Singel-gram
   - Range-Gram
1. T-SNE

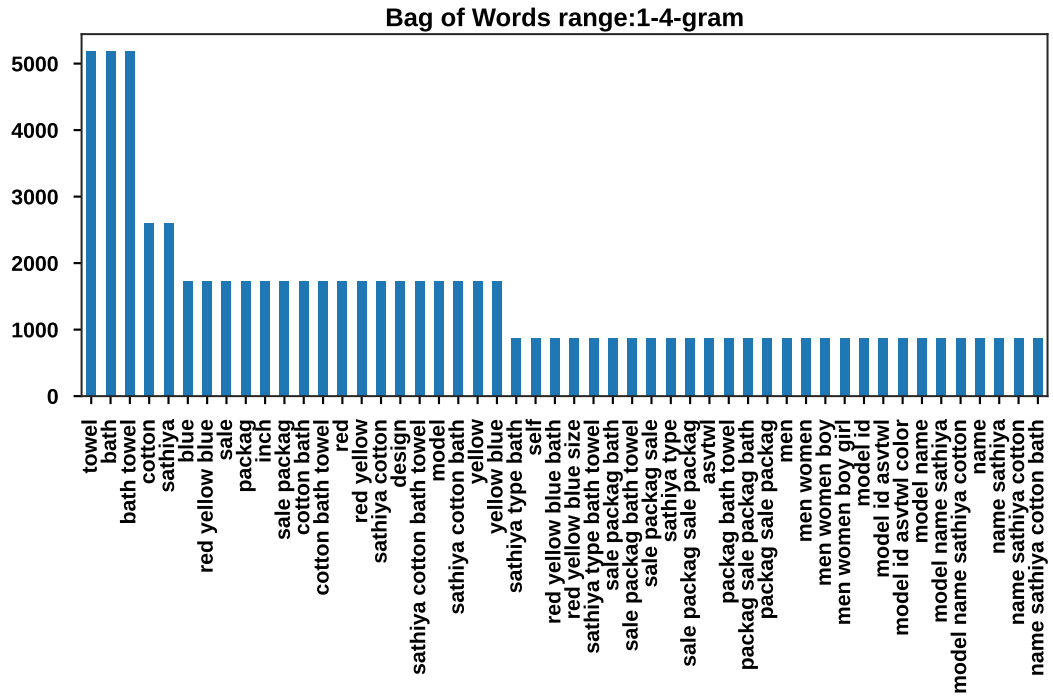




Pre-processing

1. Preprocessing pipeline
   - Tokenization
   - StopWords
   - Lemmatization
   - Stemming
1. Bag of words
   - Singel-gram
   - **Range-Gram**
1. TF-IDF
   - Singel-gram
   - Range-Gram
1. T-SNE

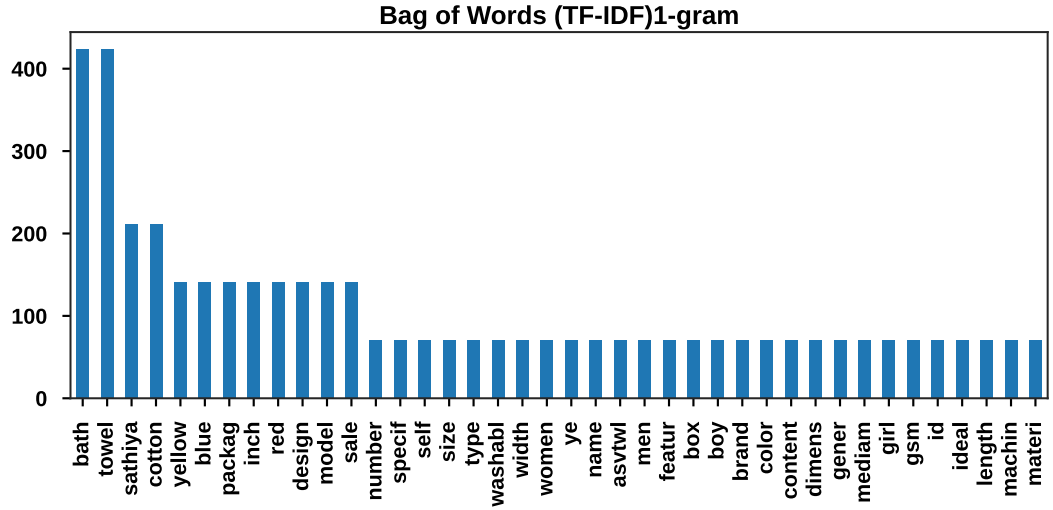


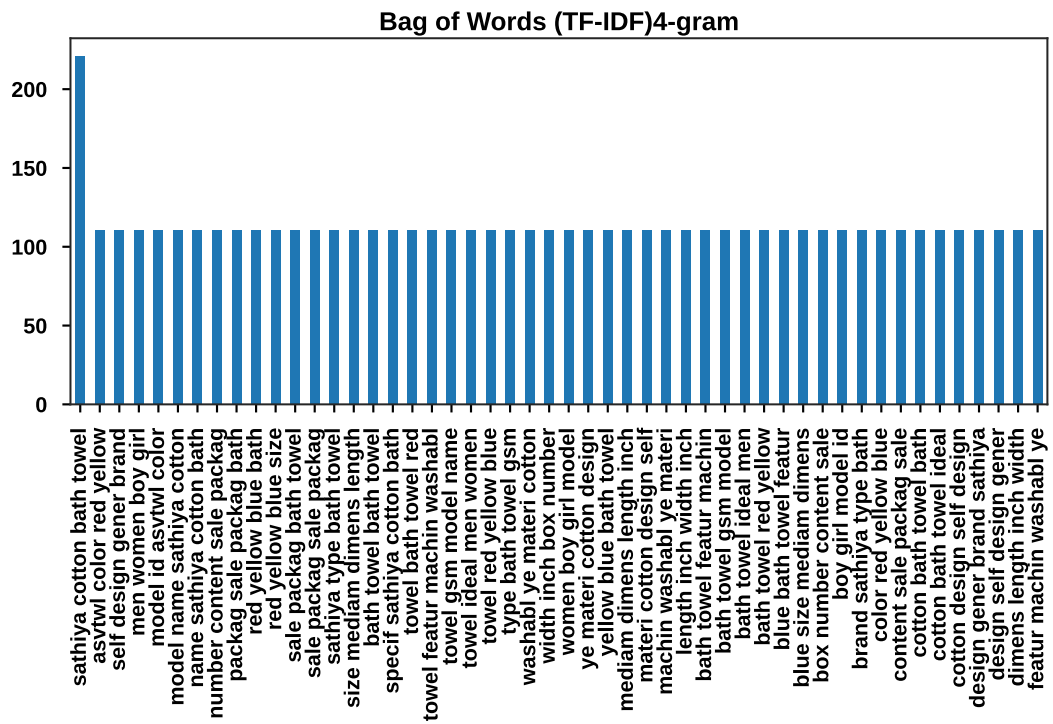


Pre-processing

place de marché

1. Preprocessing pipeline
   - Tokenization
   - StopWords
   - Lemmatization
   - Stemming
1. Bag of words
   - Singel-gram
   - Range-Gram
1. **TF-IDF**
   - **Singel-gram**
   - **Range-Gram**
1. T-SNE

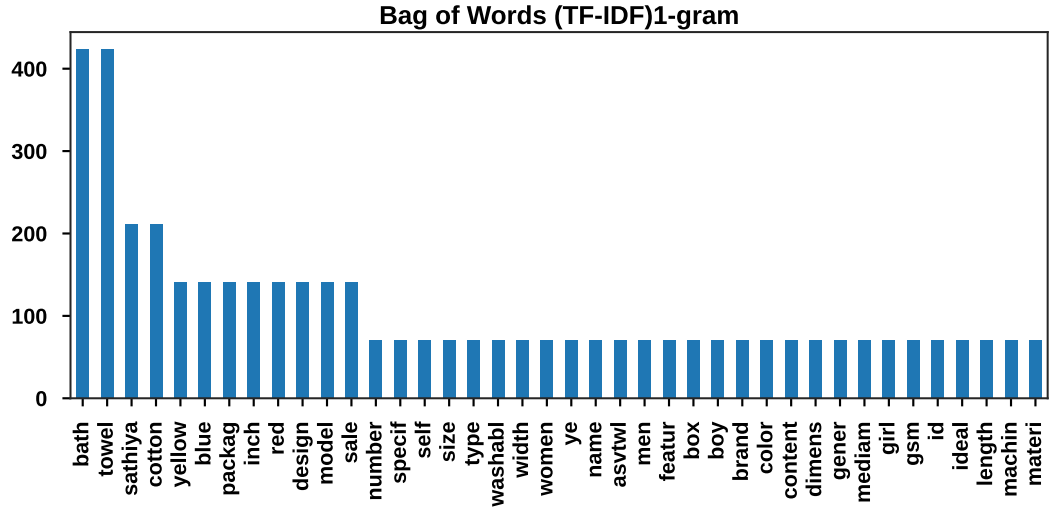


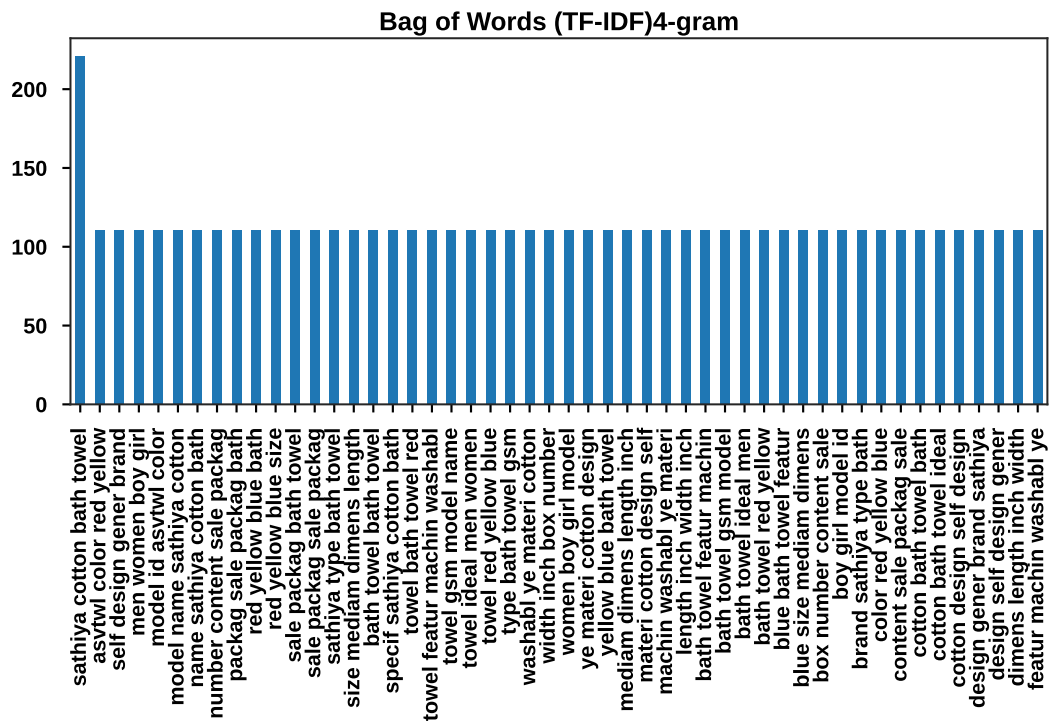


Pre-processing

place de marché

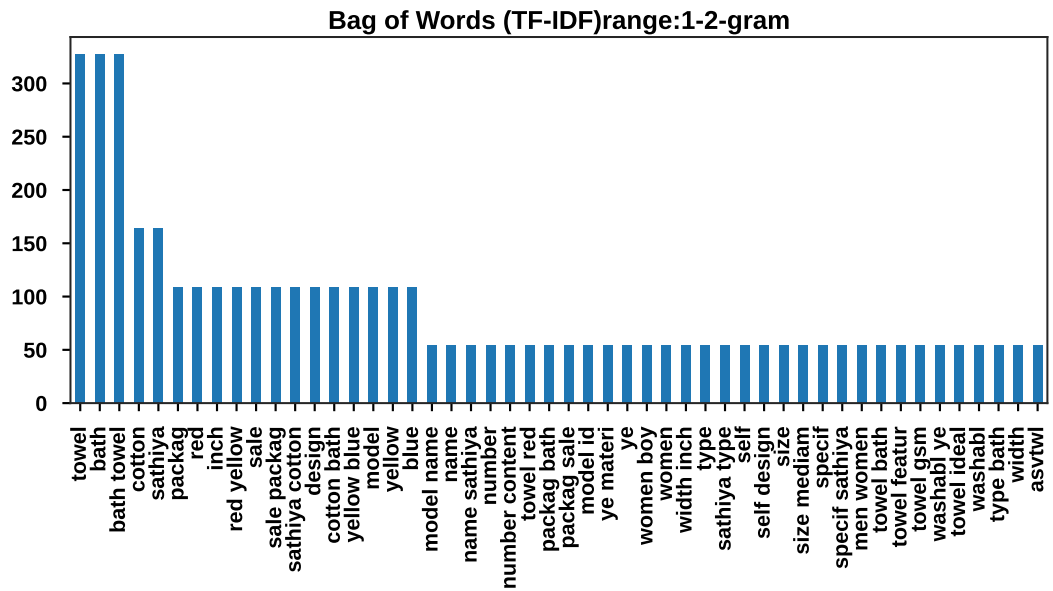
1. Preprocessing pipeline
   - Tokenization
   - StopWords
   - Lemmatization
   - Stemming
1. Bag of words
   - Singel-gram
   - Range-Gram
1. TF-IDF
   - **Singel-gram**
   - Range-Gram
1. T-SNE

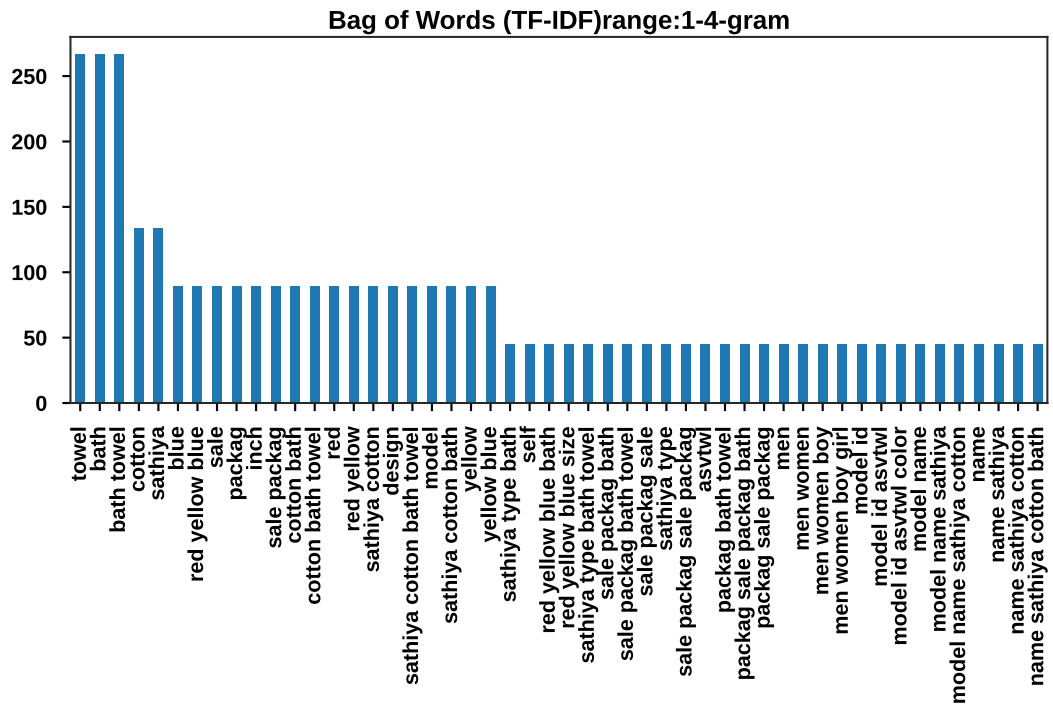




Pre-processing

1. Preprocessing pipeline
   - Tokenization
   - StopWords
   - Lemmatization
   - Stemming
1. Bag of words
   - Singel-gram
   - Range-Gram
1. TF-IDF
   - Singel-gram
   - **Range-Gram**
1. T-SNE

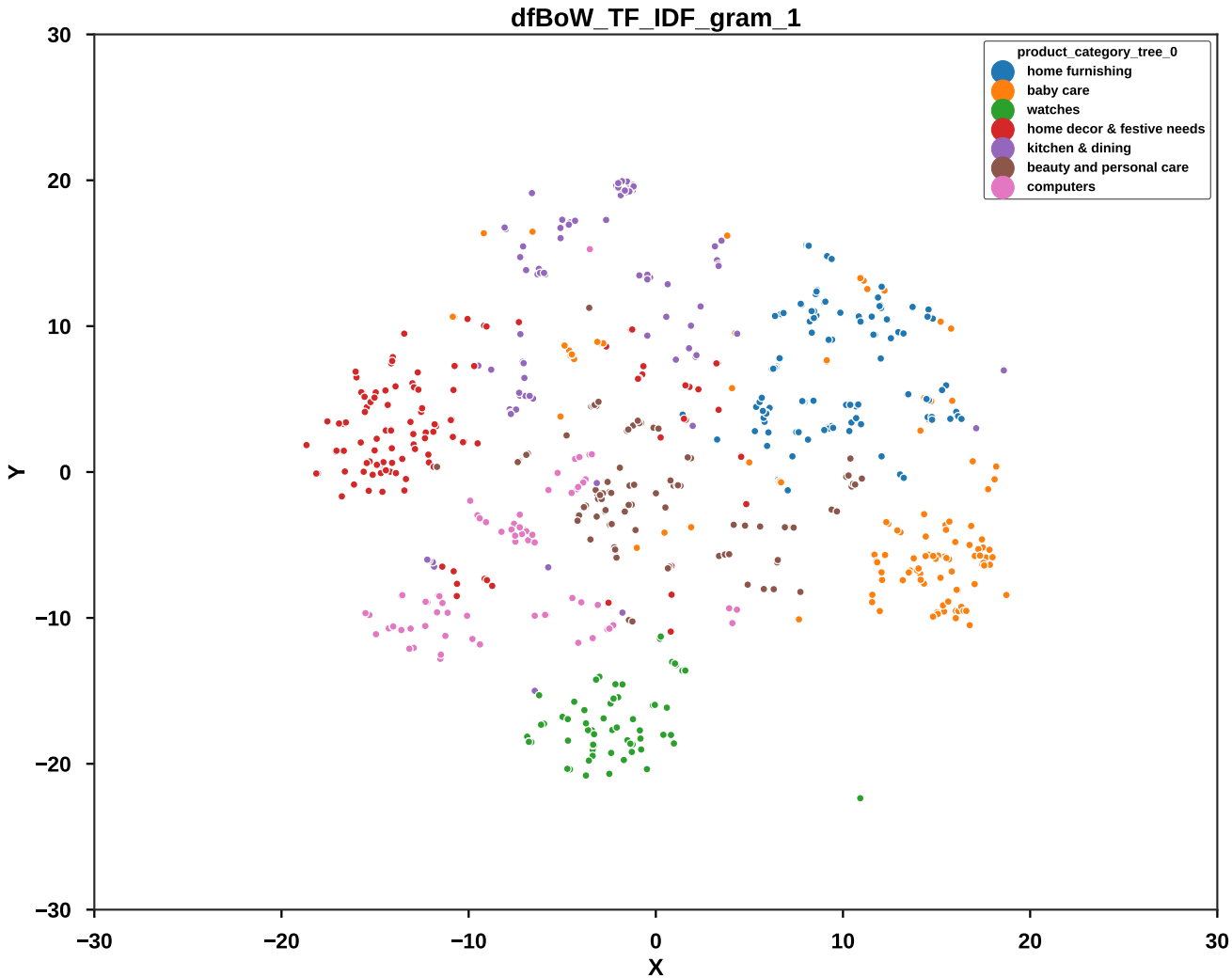




Pre-processing

place de marché

1. Preprocessing pipeline
   - Tokenization
   - StopWords
   - Lemmatization
   - Stemming
1. Bag of words
   - Singel-gram
   - Range-Gram
1. TF-IDF
   - Singel-gram
   - Range-Gram
1. **T-SNE**

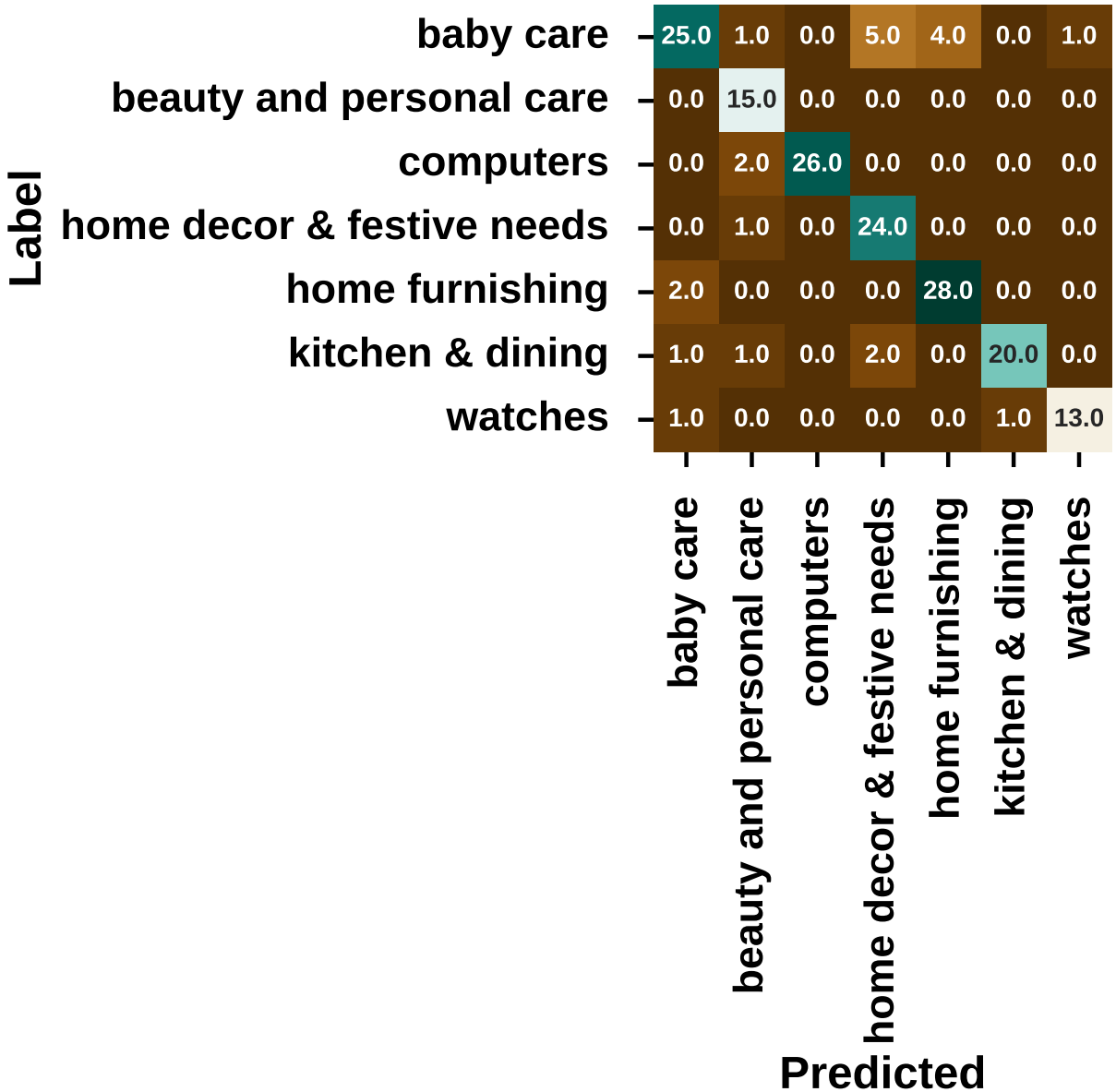


Modeling

place de marché

Accuracy : 87%

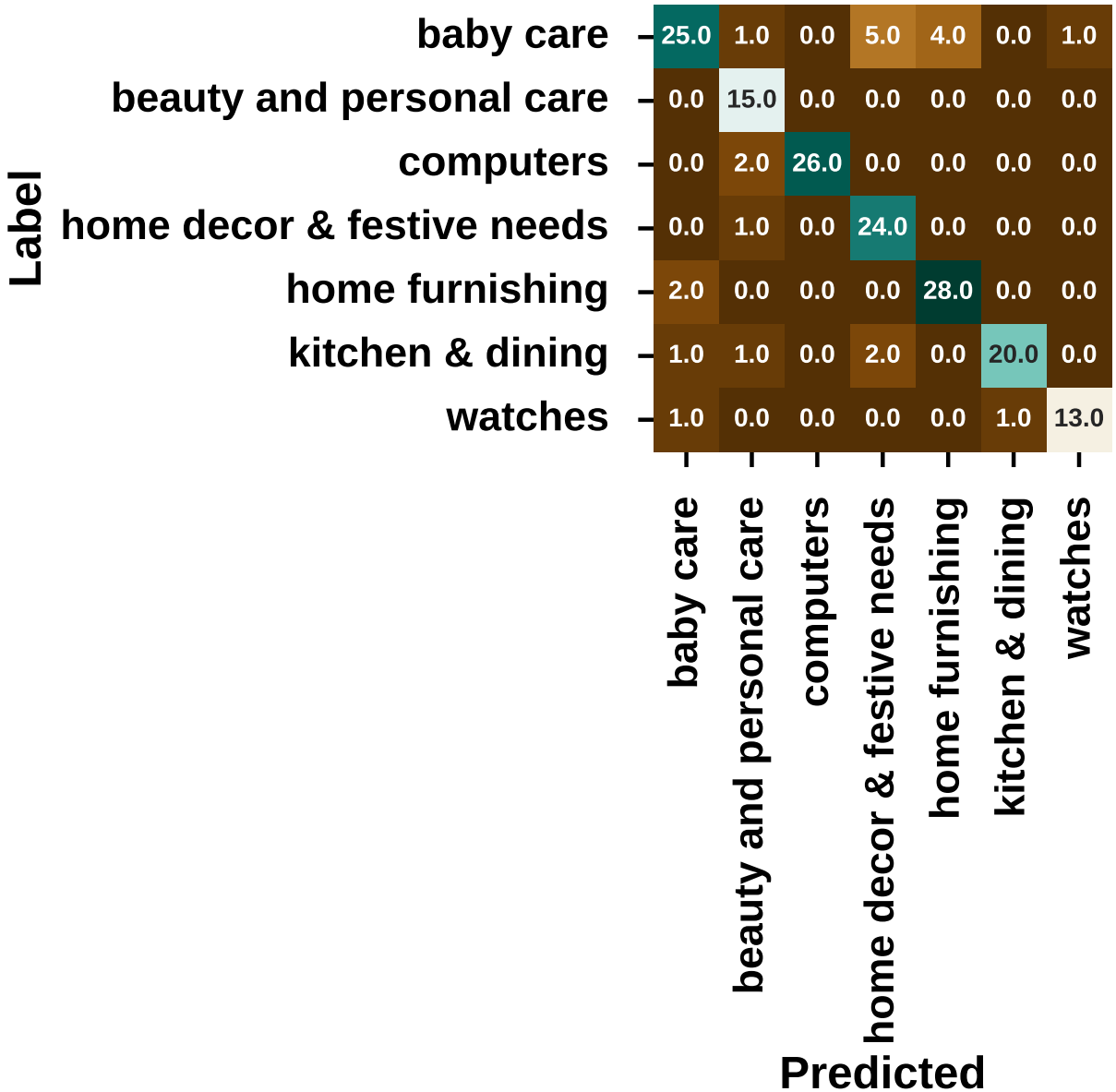
1. **Naive Bayes**
   - **Confusion matrix**
   - **Weight**
   - **Cloud of words**
2. LDA & NMF*



Modeling

place de marché

Accuracy : 87%



1. Naive Bayes
   - **Confusion matrix**
   - Weight
   - Cloud of words
2. LDA & NMF*

Modeling

place de marché

1. Naive Bayes
   - Confusion matrix
   - **Weight**
   - Cloud of words
2. LDA & NMF*

y=baby care top features		**y=beauty and personal care** top features		**y=computers** top features		**y=home decor & festive needs** top features		**y=home furnishing** top features		**y=kitchen & dining** top features		**y=watches** top features	
Weight?	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**
+14.717	baby	+12.234	combo	+15.162	usb	+17.023	showpiece	+8.054	cover	+12.807	kadhai	+16.464	watch
+7.798	cotton	+8.700	jewellery	+11.503	laptop	+11.300	cm	+7.227	blanket	+8.915	bottle	+11.869	analog
+7.519	photo	+8.689	vanity	+5.798	battery	+9.197	wall	+6.855	code	+8.292	bulb	+6.919	discounts
+7.175	frame	+7.568	hair	+5.440	light	+6.543	rice	+6.683	at	+7.982	mug	+6.530	men
+7.080	small	+6.724	flipkart	+5.303	from	+6.468	inch	+6.216	cushions	+7.721	ceramic	+6.411	dial
+6.934	offering	+6.530	com	+5.236	only	+5.753	hookah	+6.127	polyester	+6.623	pizza	+6.066	great
+6.793	girls	+6.361	lowest	+4.397	keyboard	+5.656	lights	+5.811	cushion	+5.895	cutter	+5.629	india
+6.142	neyth	+6.258	set	+4.226	edimax	+5.412	decor	+5.775	blankets	+5.651	ml	+4.572	sonata
+6.074	980	+5.925	lip	+4.130	replacement	+5.295	creative	+5.555	eyelet	+5.230	prithish	+4.210	women
+6.074	carter	+5.626	surgical	+4.017	warranty	+5.260	painting	+5.427	quilts	+5.094	safe	+4.203	boys
+5.969	girl	+5.570	eye	+4.017	bluetooth	+4.985	art	+5.298	single	+4.991	tawa	+4.012	display
+5.916	479	+5.553	traits	+3.958	pad	+4.603	shelf	+5.202	plain	+4.951	kitchen	+3.845	digital
+5.680	size	+5.372	kit	+3.936	flexible	+4.573	item	+5.183	comforters	+4.908	capacity	+3.528	at
+5.635	wallmantra	+5.345	oil	+3.849	zyxel	+4.390	beautiful	+5.066	velvet	+4.642	water	+3.251	strap
+5.561	ireeya	+5.149	and	+3.814	led	+4.178	brass	*... 753 more positive ...*		+4.597	steel	+2.498	for
+5.425	ideal	+5.141	blush	*... 1000 more positive ...*		+4.166	paper	*... 4273 more negative ...*		+4.577	mugs	+2.403	girls
+5.316	brush	+4.963	massage	*... 4026 more negative ...*		+4.125	with	-5.206	479	+4.253	lock	+2.334	hands
+5.288	coral	+4.939	body	-3.834	combo	*... 1341 more positive ...*		-5.312	baby	+4.249	juicer	+2.241	in
+5.243	portia	+4.879	oxyglow	-3.871	set	*... 3685 more negative ...*		-5.345	neyth	*... 969 more positive ...*		*... 411 more positive ...*	
+5.186	529	*... 1171 more positive ...*		-3.927	cm	-4.111	com	-5.503	980	*... 4057 more negative ...*		*... 4615 more negative ...*	
... 950 more positive ...		*... 3855 more negative ...*		-5.390	at	-4.184	usb	-5.503	carter	-4.281	inch	-2.489	cm
... 4076 more negative ...		-5.911	at	-5.468	online	-4.407	vanity	-6.428	for	-4.611	showpiece	-2.503	with

Weights

Modeling

place de marché

1. Naive Bayes
   - Confusion matrix
   - Weight
   - **Cloud of words**
2. LDA & NMF*

Cloud of words

Modeling

place de marché

1. Naive Bayes
   - Confusion matrix
   - Weight
   - Cloud of words
2. **LDA & NMF***

```
Topic 0:
com flipkart product ship day
Topic 1:
com flipkart ship product replac
Topic 2:
flipkart com ship product rs
Topic 3:
com flipkart ship product genuin
Topic 4:
flipkart com ship product deliveri
```

LDA

```
Topic 0:
com flipkart product
Topic 1:
box materi set watch
Topic 2:
box materi set watch
Topic 3:
box materi set watch
Topic 4:
box materi set watch
```

NMF

Index

Pre-processing

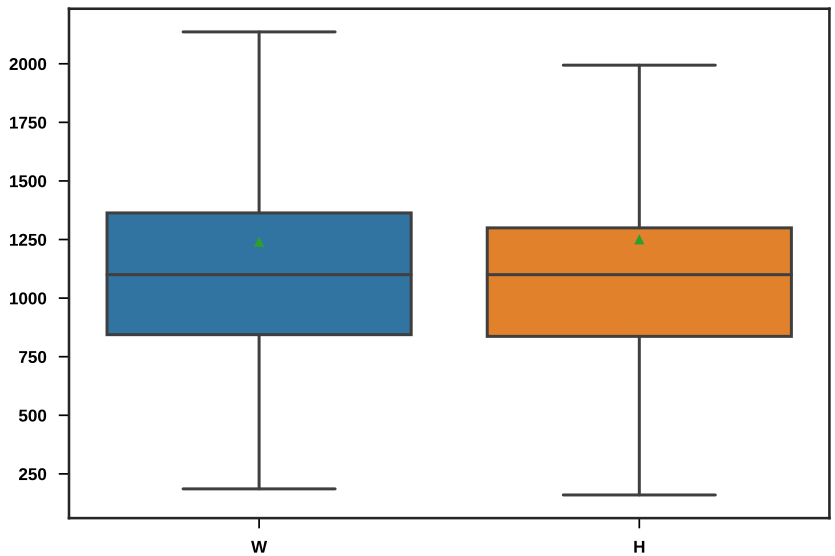
place de marché

1. **GRAY**
2. Scaling
3. Gaussian filtering
4. CLAHE

Pre-processing

place de marché

1. GRAY
2. **Scaling**
3. Gaussian filtering
4. CLAHE



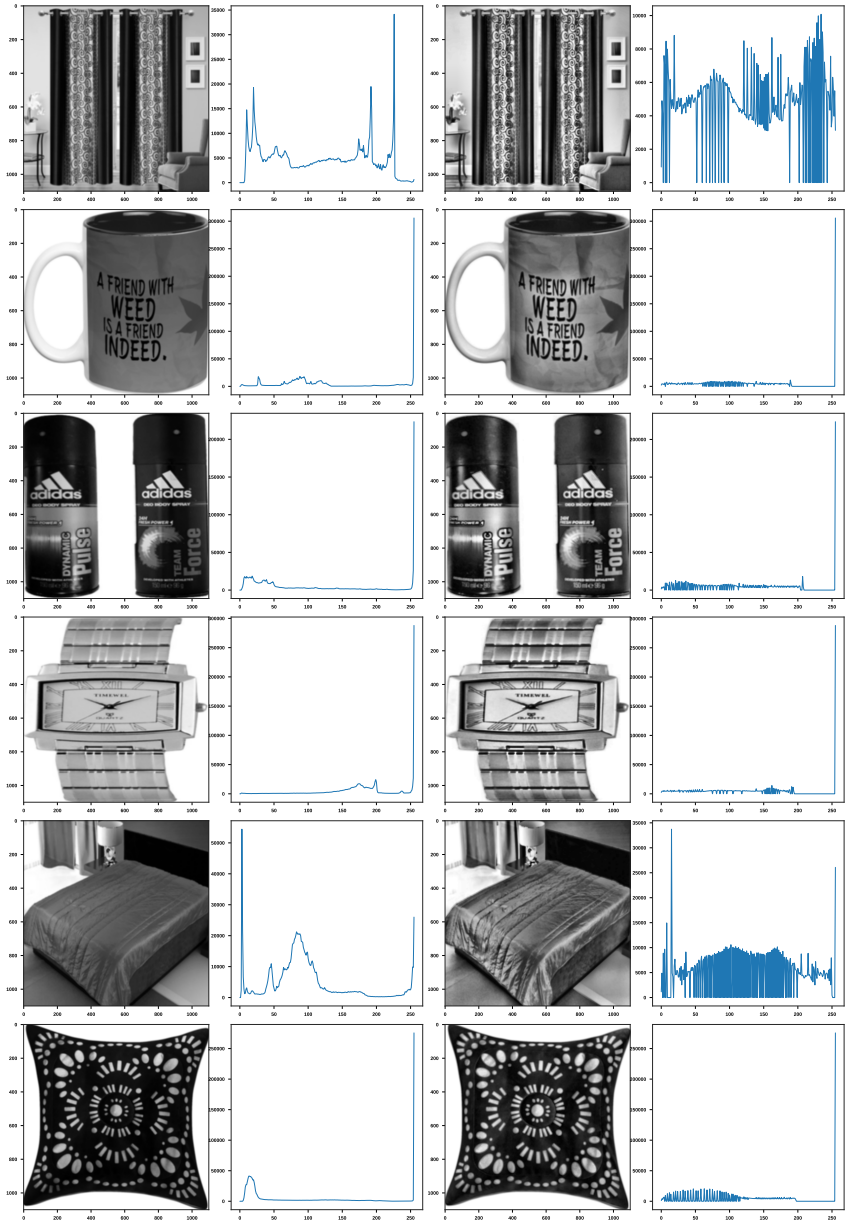
Pre-processing

1. GRAY
2. Scaling
3. **Gaussian filtering**
4. CLAHE

Pre-processing

place de marché

1. GRAY
2. Scaling
3. Gaussian filtering
4. **CLAHE**

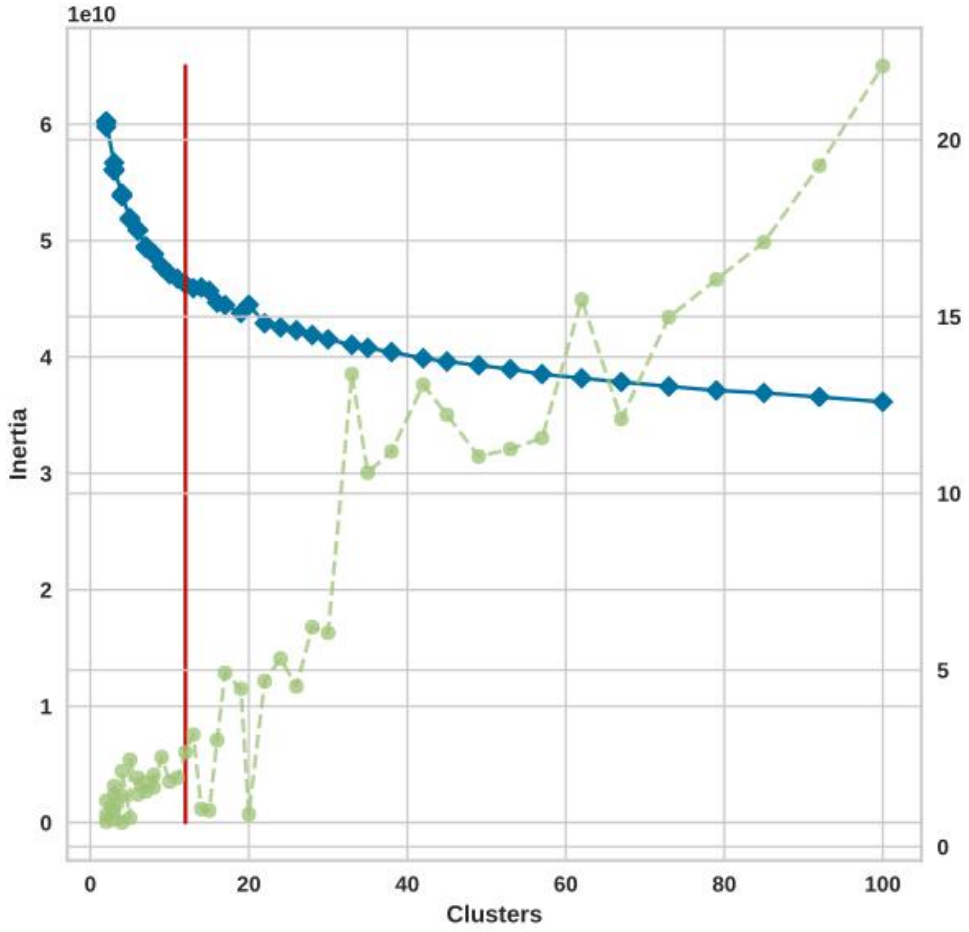


Feature extration

place de marché

1. **KMeans**
2. BoVW
   - T-SNE
3. Naive Bayes

ORB (386913 Features)



MiniBatchKMeans (k=12)

Feature extration

place de marché

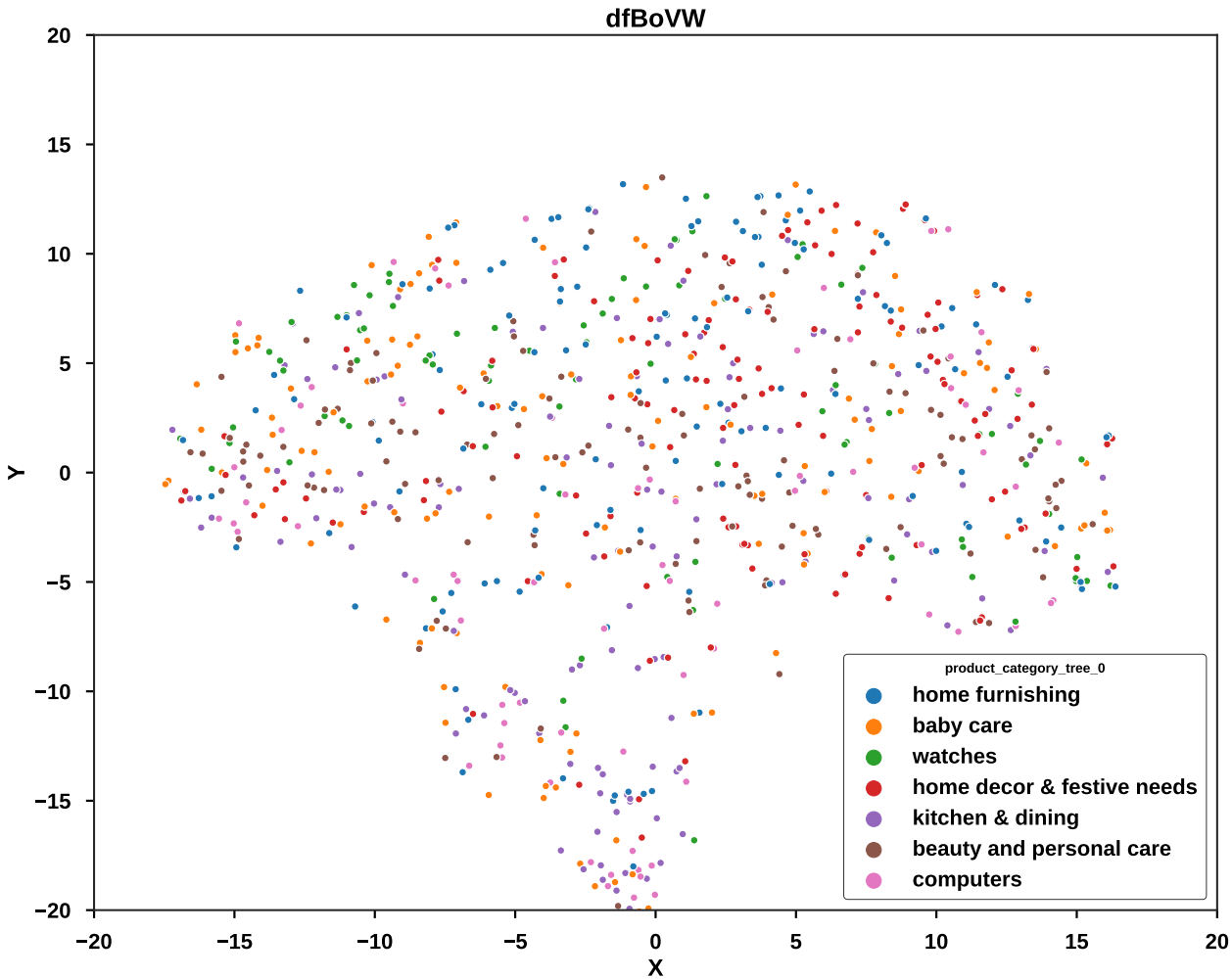
1. KMeans
2. **BoVW**
   - **T-SNE**
3. Naive Bayes

	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0
0.0	66	33	37	19	27	24	41	50	59	14	68	62
1.0	56	24	20	25	23	72	38	74	65	25	38	40
2.0	18	40	33	19	51	32	60	52	48	35	48	64
3.0	37	50	15	105	15	79	17	45	78	19	21	19
4.0	33	26	34	31	31	55	45	73	48	21	66	37
...	...	...	...	...	...	...	...	...	...	...	...	...
858.0	11	38	16	106	45	99	34	35	22	26	46	22
859.0	51	27	16	89	29	73	44	57	45	27	28	14
860.0	36	35	28	91	63	61	28	31	23	32	28	22
861.0	27	48	35	105	29	45	51	38	15	26	46	35
862.0	32	57	18	67	40	53	38	41	31	37	28	17

Feature extration

place de marché

1. KMeans
2. BoVW
   - **T-SNE**
3. Naive Bayes

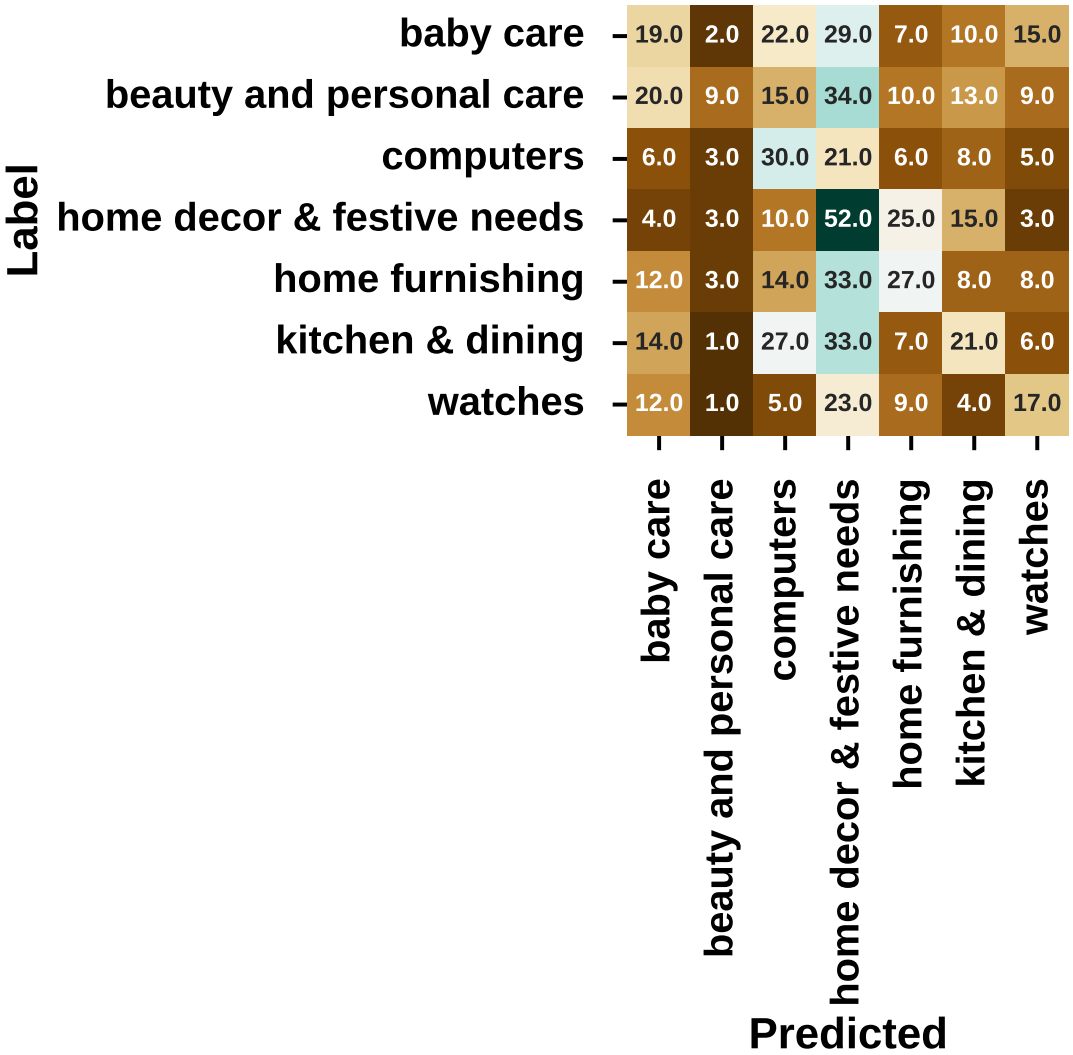


Feature extration

place de marché

Accuracy : 27%

1. KMeans
2. BoVW
   - T-SNE
3. **Naive Bayes**



Confusion matrix

Index

VGG16

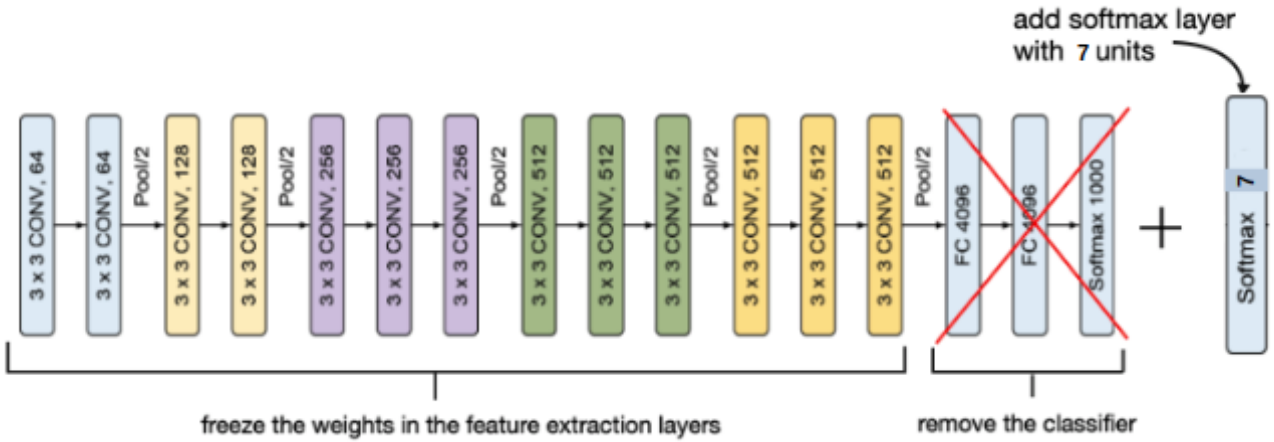
place de marché

1. **Model**
2. Results
   - Validation
   - Confusion matrix

- Freeze the feature extraction part

- Reemove the classifier part

- Add our new classifier softmax layer 7 hidden units



Total params: 14,890,311

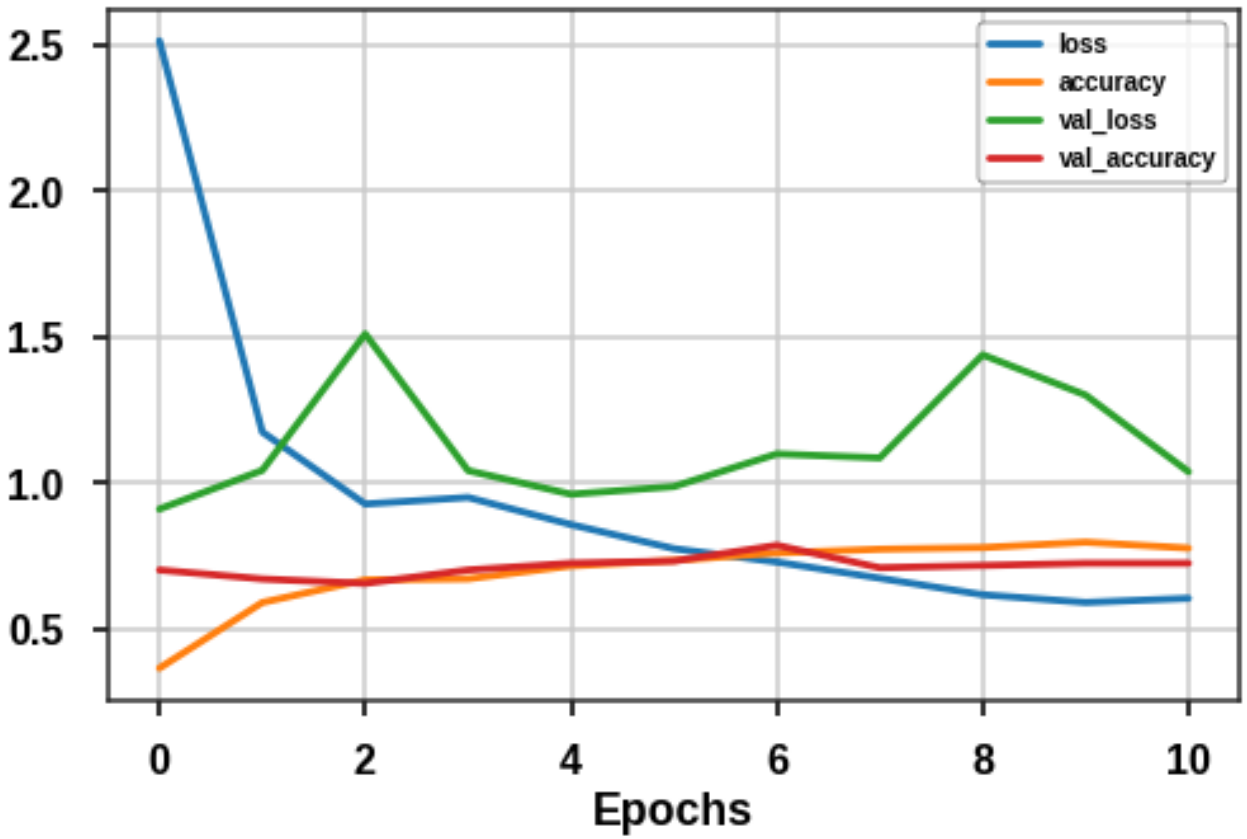
Trainable params: 175,623

Non-trainable params: 14,714,688

VGG16

place de marché

1. Model
2. **Results**
   - **Validation**
   - **Confusion matrix**

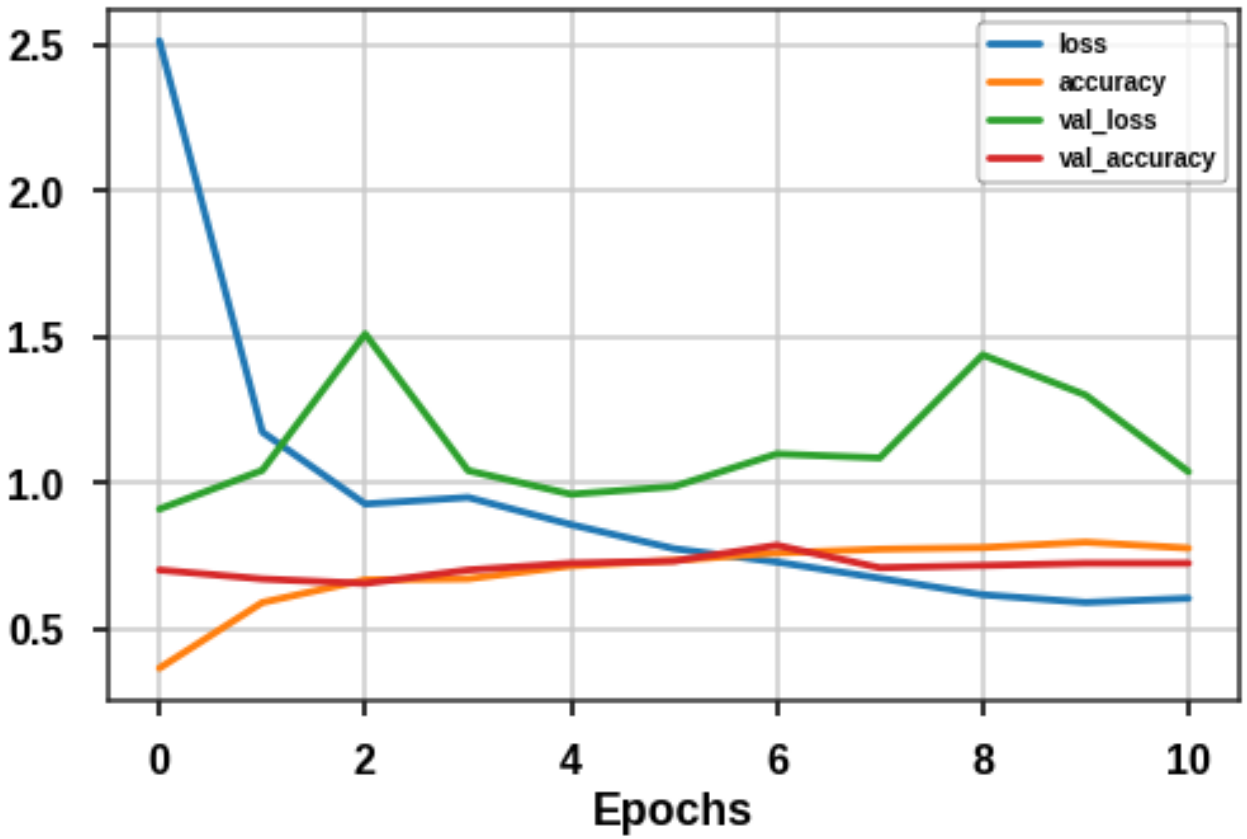


Validation accuracy : 78%

VGG16

place de marché

1. Model
2. Results
   - **Validation**
   - Confusion matrix

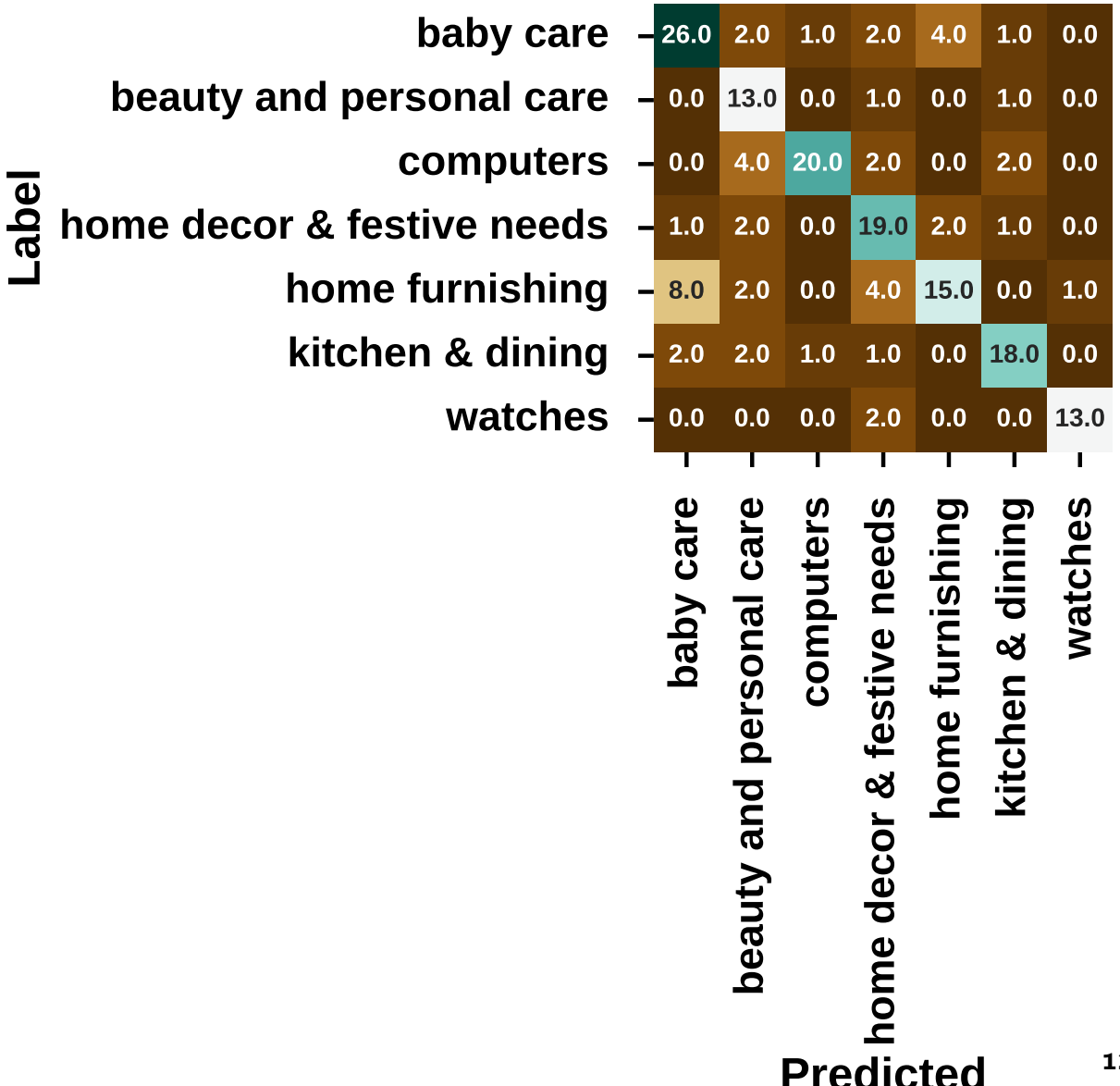


Validation accuracy : 78%

VGG16

place de marché

Accuracy : 72%



1. Model
2. Results
   - Validation
   - **Confusion matrix**

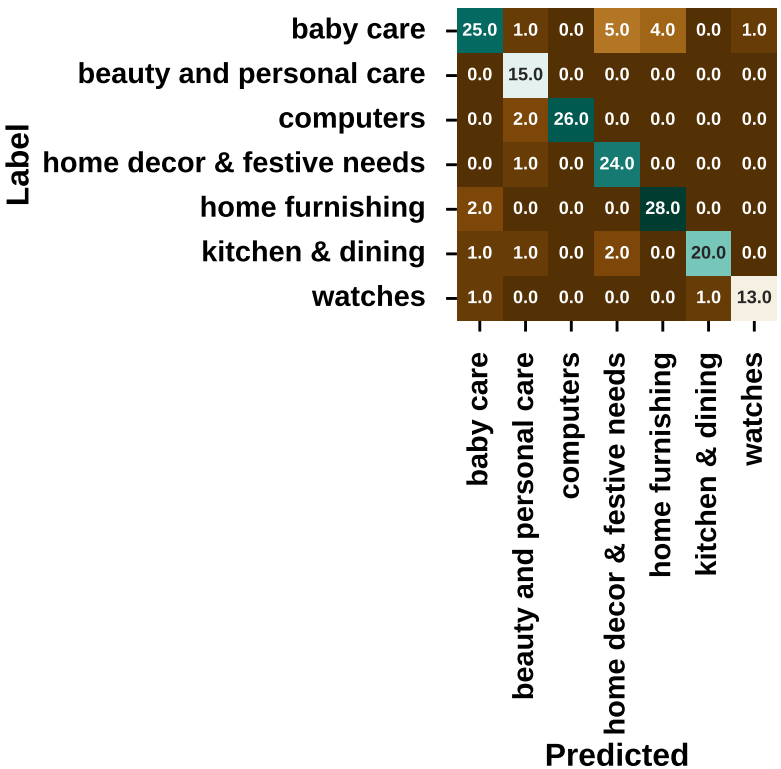
Index

Conclusions

place de marché

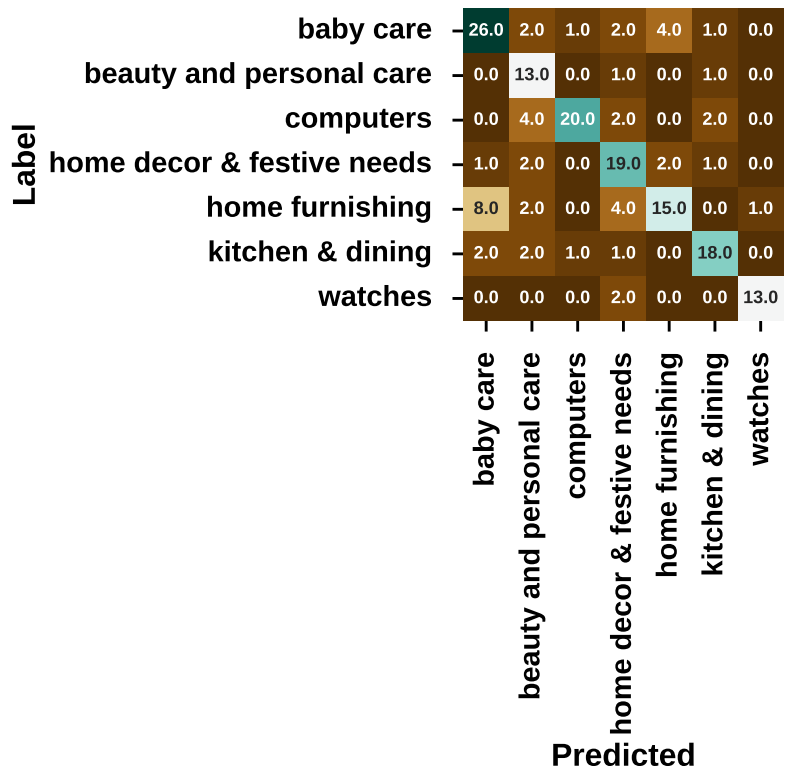
1 Dans l'ordre coissant, les meilleurs resultats sont les suivant :

Accuracy : 87%



NLP

Accuracy : 72%



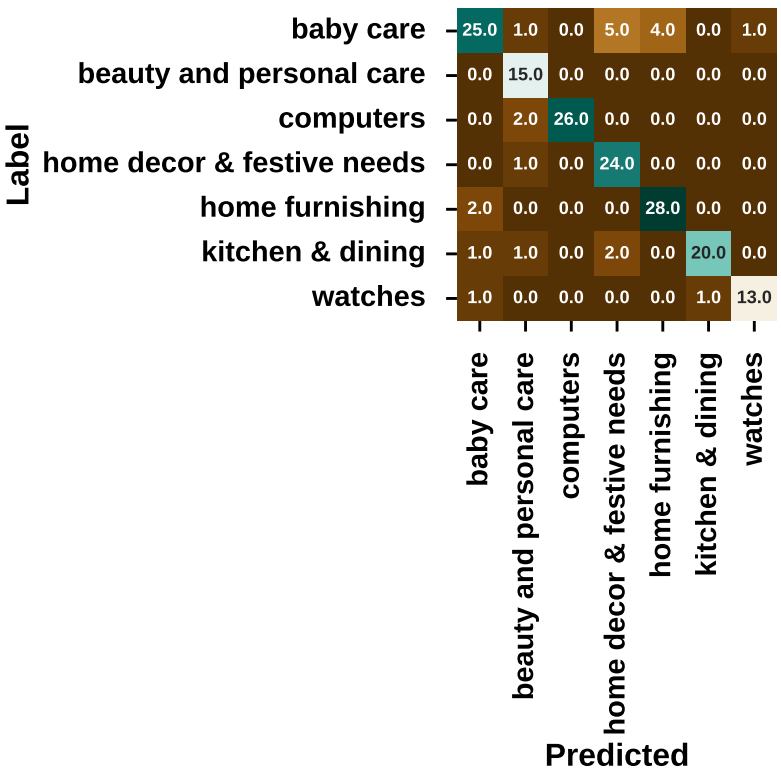
DL : VGG16

Conclusions

place de marché

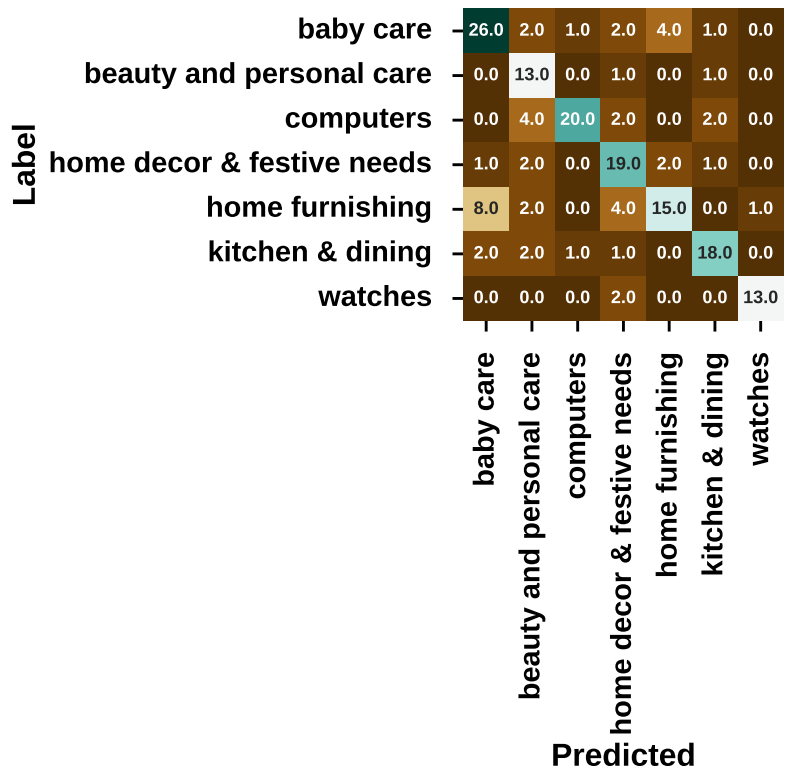
1. Dans l'ordre coissant, les meilleurs resultats sont les suivant :
   - **NLP (87%)**, Deep Learning (VGG16, 72%), ~~Maching learning (OpenCV, 27%)~~

Accuracy : 87%



NLP

Accuracy : 72%



DL : VGG16