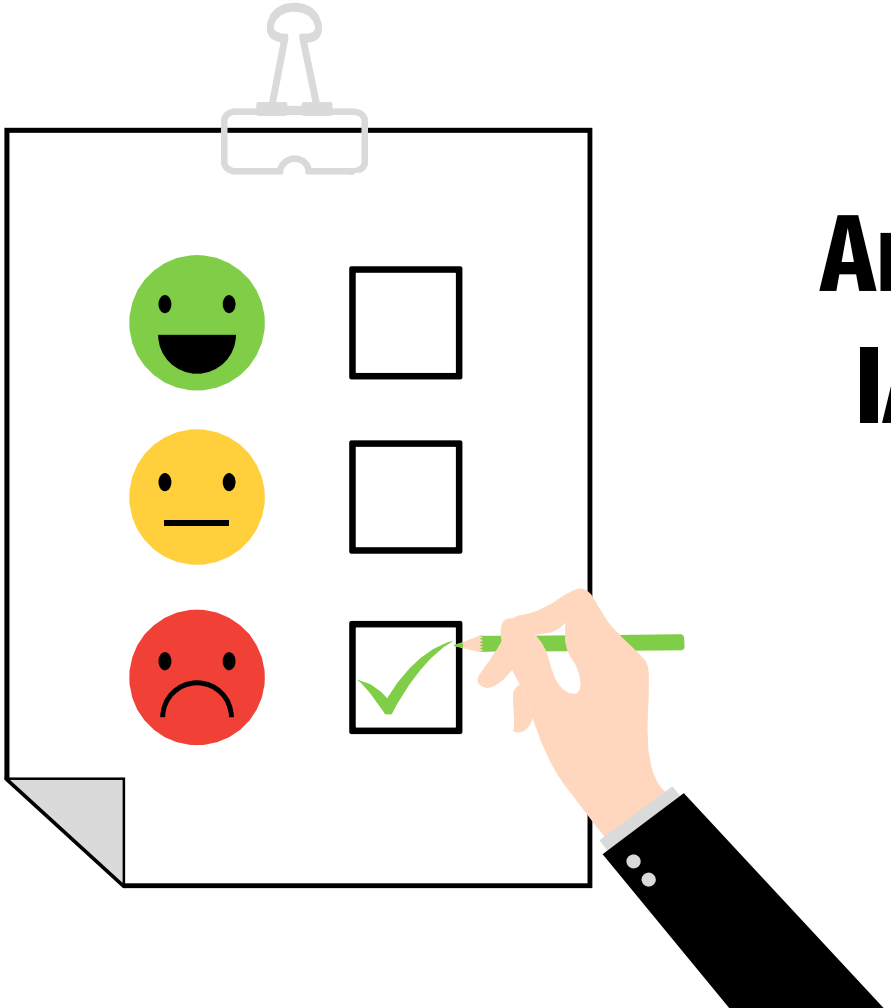


Améliorez le produit IA de votre start-up

Hourdin Charlène - Août 2022

Agenda

Présentation du projet

Collecte des données via l'API

Détecter les sujets d'insatisfaction

Classification d'images

Appel à projet

01 Le projet

Améliorer la plateforme d'Avis restau avec une nouvelle fonctionnalité de collaboration

02 Les objectifs

Etudier la faisabilité pour :

- Détecter les sujets d'insatisfaction
- Labelliser automatiquement les photos postées
- Agrémenter la base de données

03 La mission

- **Analyser les commentaires négatif** postée par les clients
- Collecter un échantillon de **200 données via l'API Yelp**
- **Analyser les photos** postée par les clients

04 La méthode

Fournir une page web présentant les résultats de l'analyse

PRÉSENTATION DU JEU DE DONNÉES

Présentation des jeux de données

Description des données

I cancelled a reservation here BEFORE the reserved time and was still charged $40 to my credit card!! This policy makes no sense. I understand if I was a...

peas in the guacamole. and the birria is offensively bad. there are a few things worth eating but so many flops

NEW YORK

50 000 avis

Filtré sur la catégorie
Restaurant

5000 images

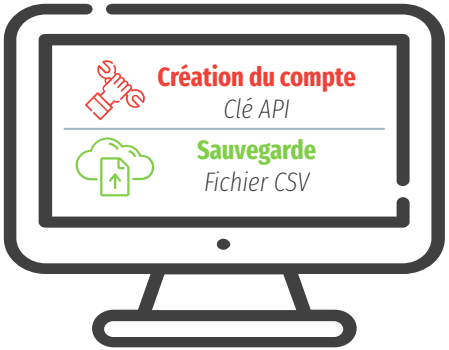
2000 train
2000 Validation
1000 test

200 avis

Restaurant New york

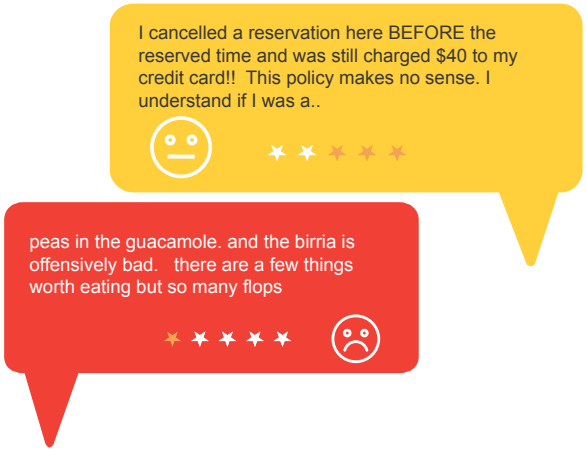
COLLECTE DES DONNÉES VIA L'API

Collecte des données



Filtre les données

Restaurant New york



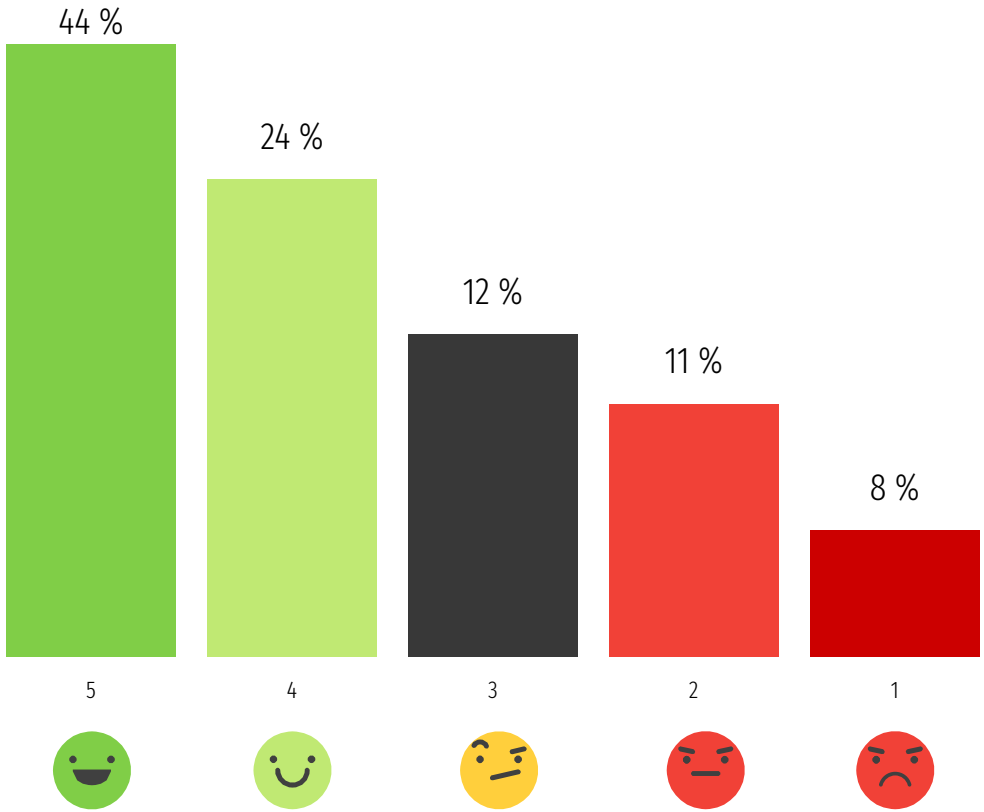
Requêtes

5000 / jours
50 commentaires

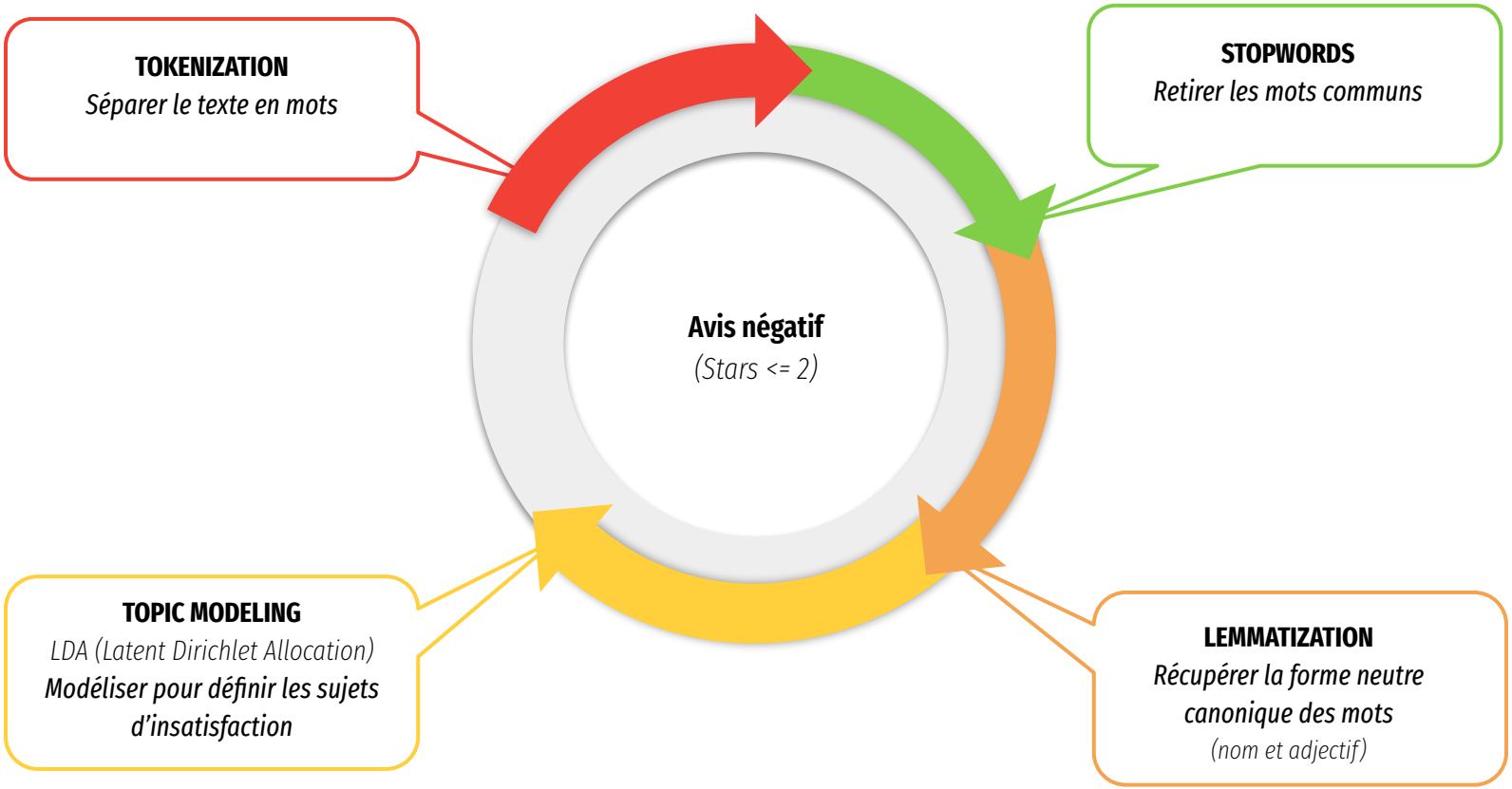
API yelp

DÉTECTER LES SUJETS D'INSATISFACTION

Répartition des notes



Démarche méthodologique



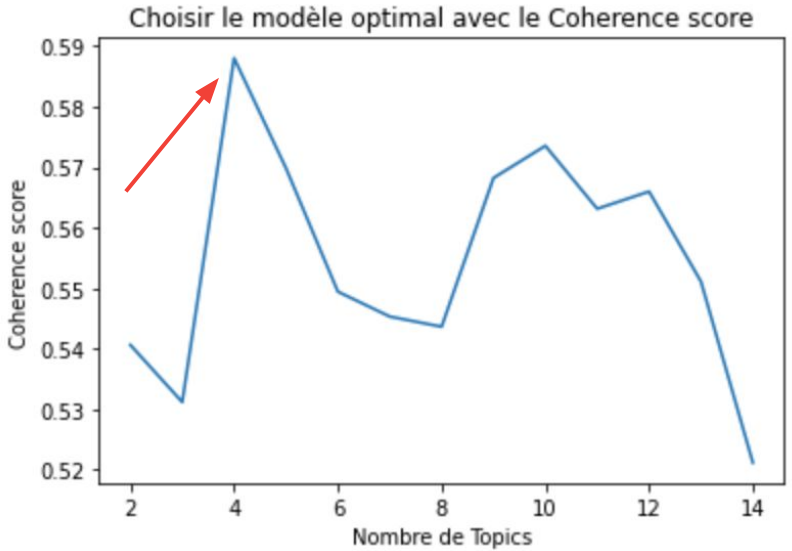
LDA (Latent Dirichlet Allocation)

Algorithme non supervisé

Le modèle tente de découvrir la proportion de rubriques partagées par des documents au sein d'un corpus de texte

Nombre de sujets

Le nombre de sujets est spécifié par l'utilisateur



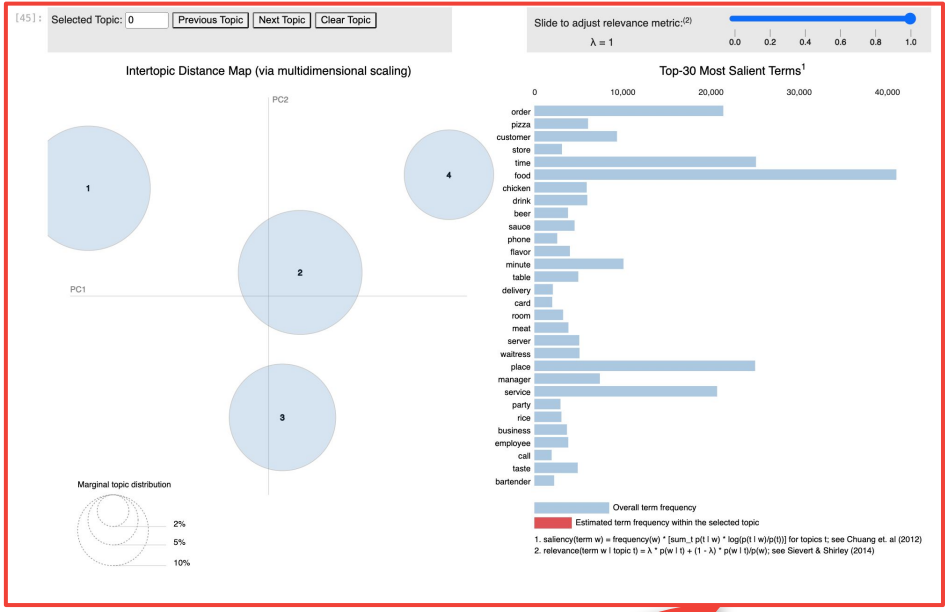
Distribution probabiliste

Les sujets sont appris par le modèle sous forme de distribution de probabilité sur les mots rencontrés

Réduction de dimension

Réduction de la taille du vocabulaire en un nombre **K** de rubriques spécifié par l'utilisateur

Extraction des topics



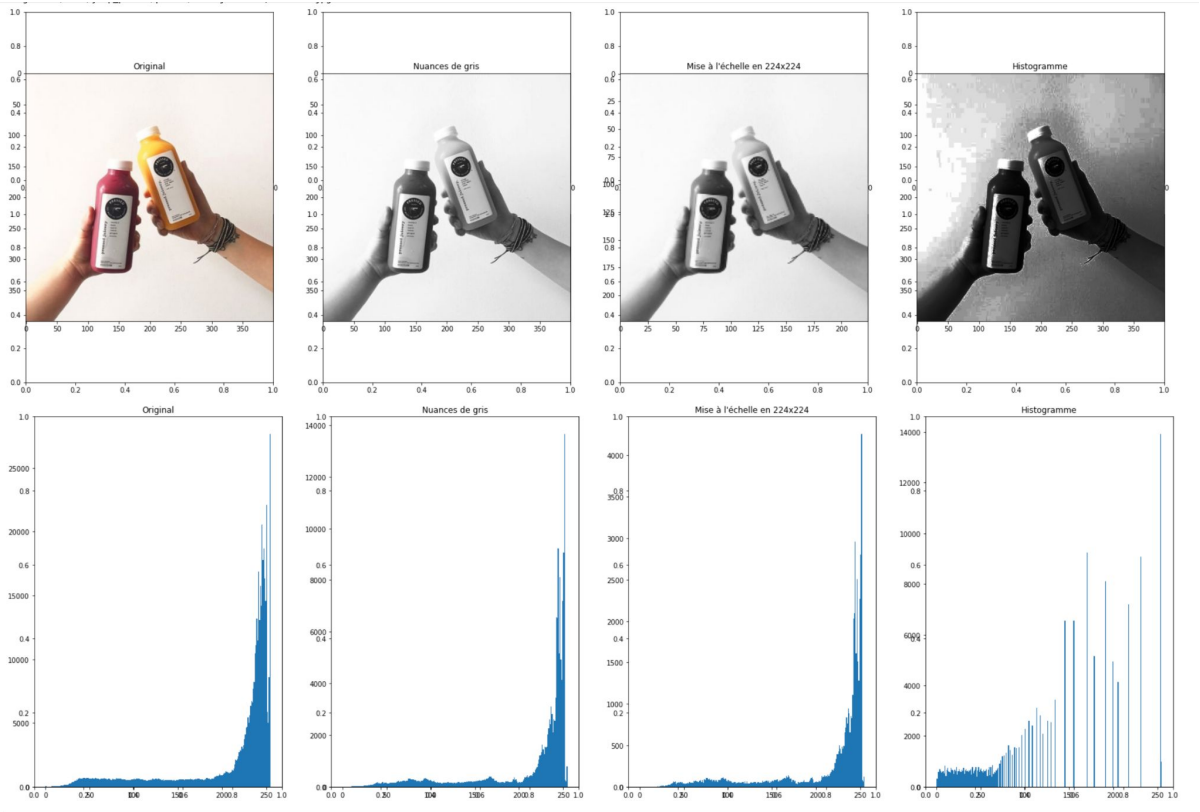
4 sujets d'insatisfaction :

- Le service (temps et personnel)
- Le lieu
- La nourriture
- Le rapport qualité prix médiocre


voilà

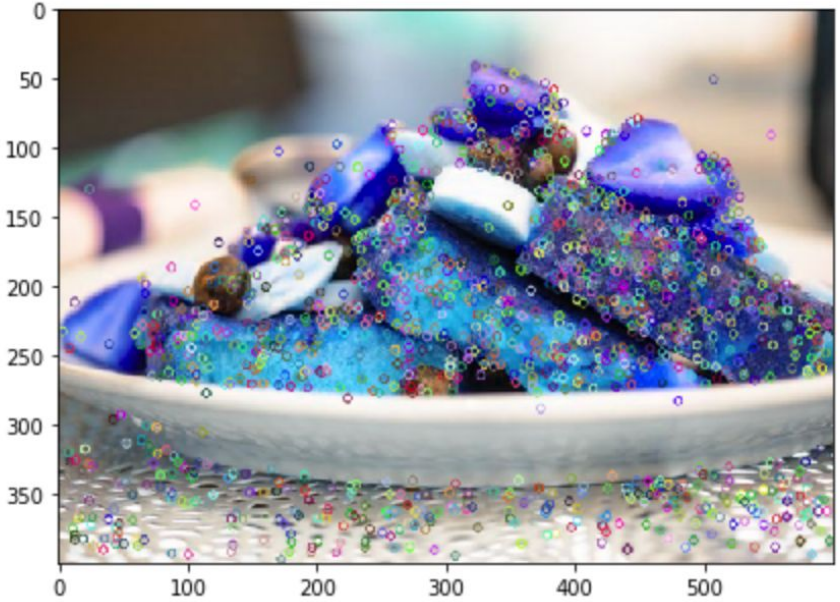
Classification des images

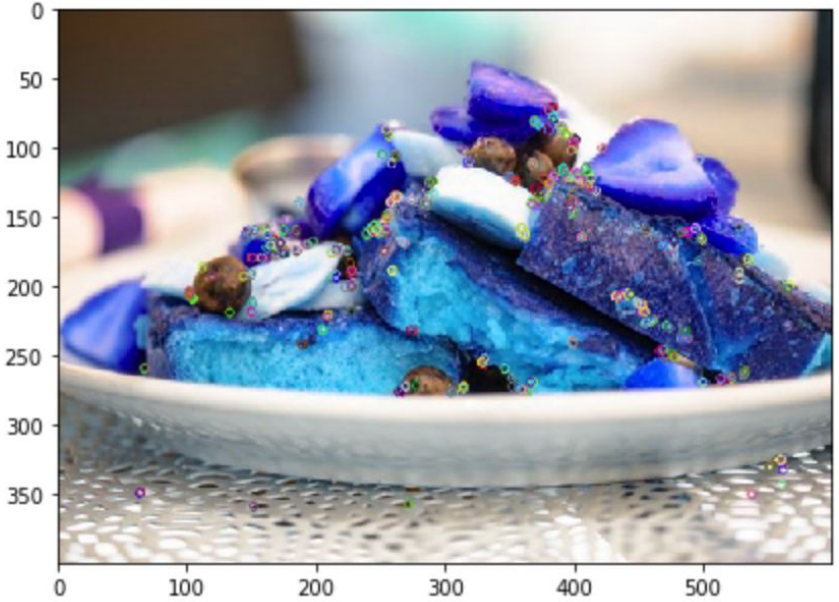
Traitement des images



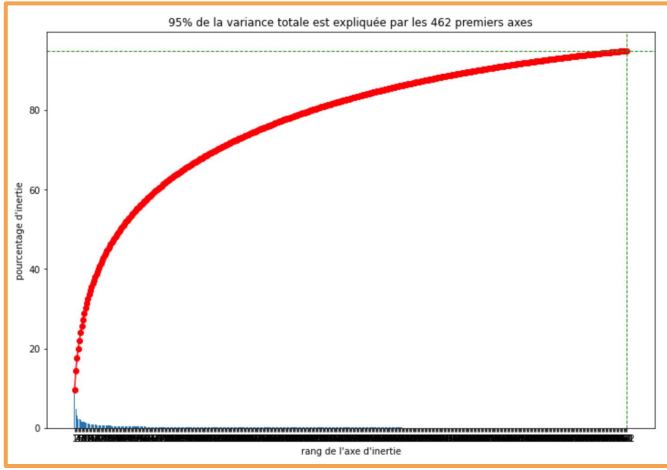
Features extraction SIFT et ORB

```
la moyenne des descripteurs Sift: 26.74167
la moyenne des descripteurs ORB: 136.64275
```

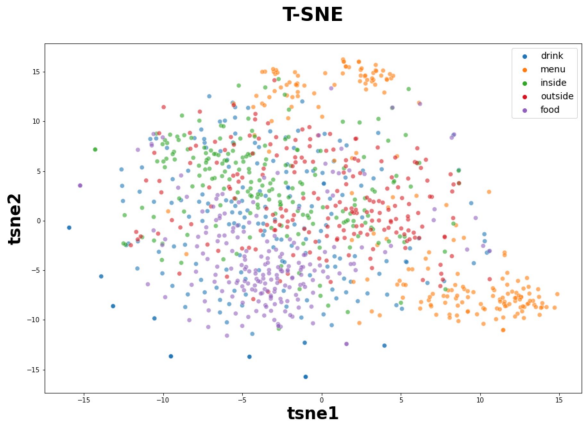


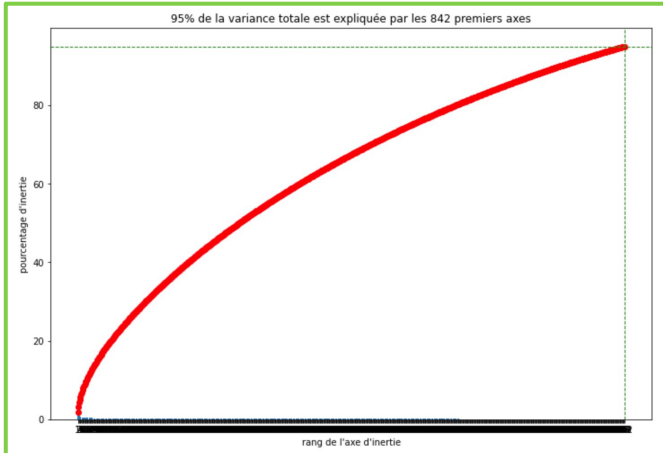


ACP

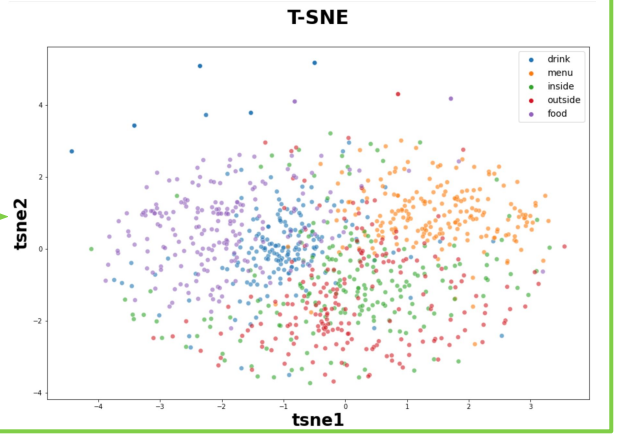


Sift



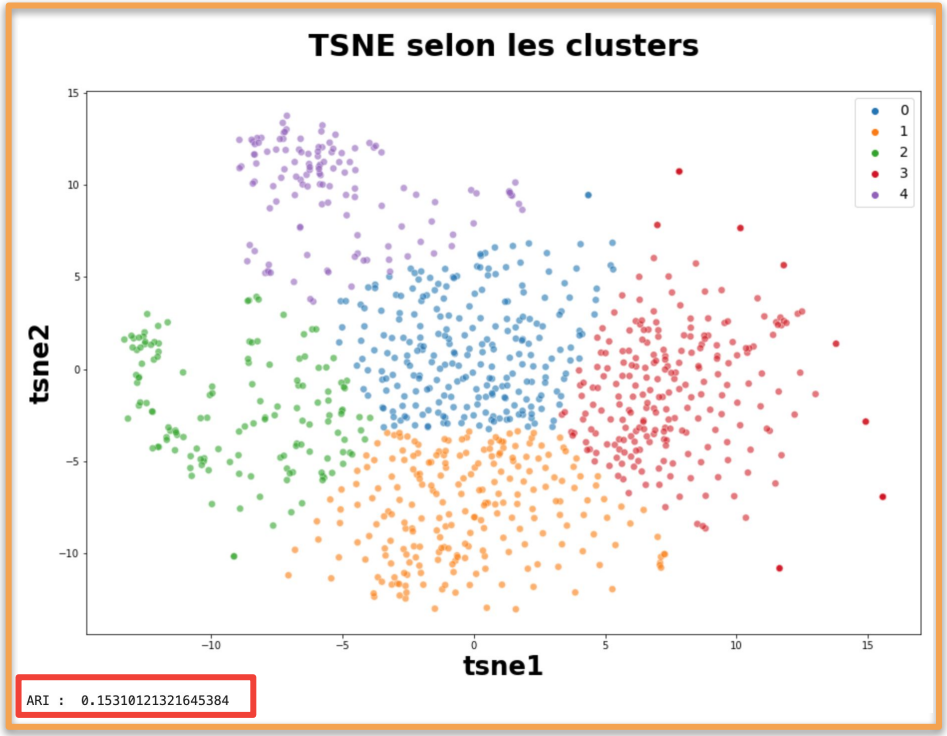


VGG16

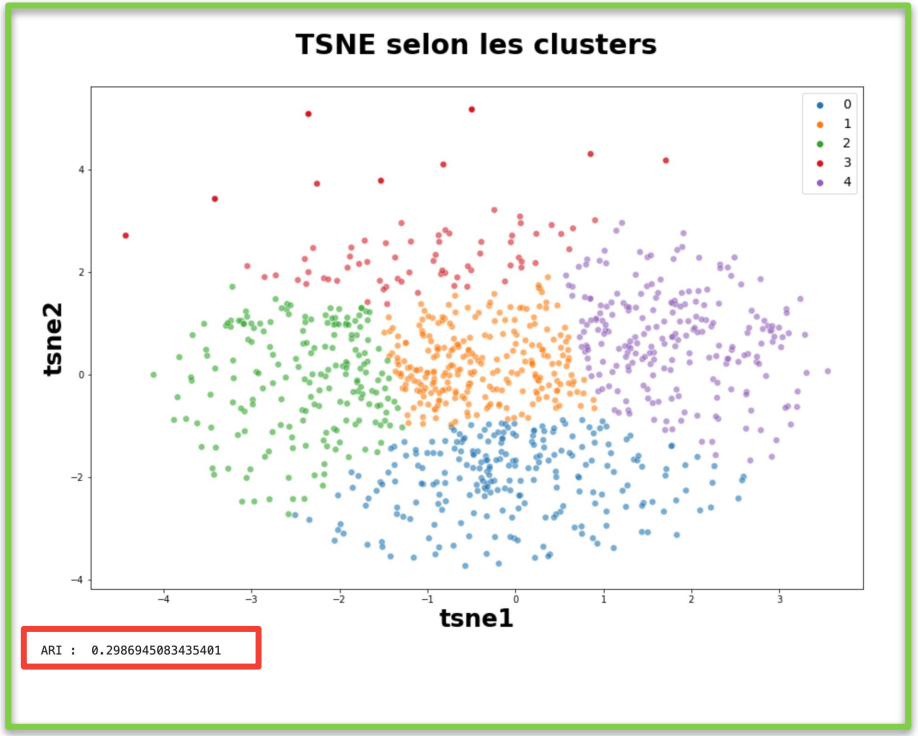


Comparaison ARI - SIFT et VGG16

SIFT

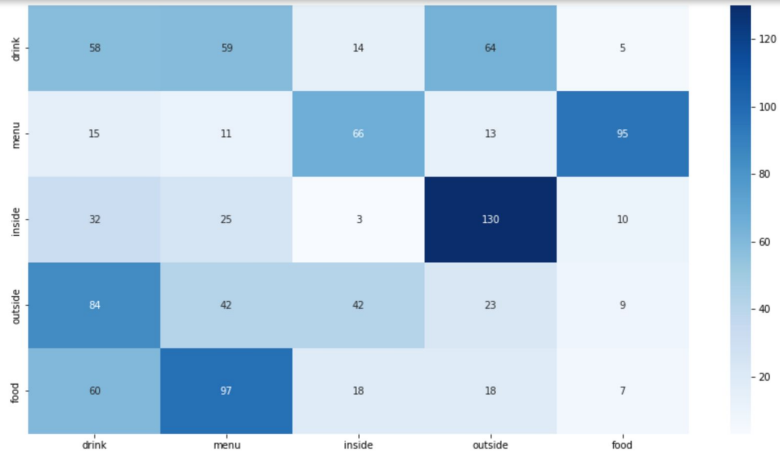


VGG16



Comparaison des résultats du clustering

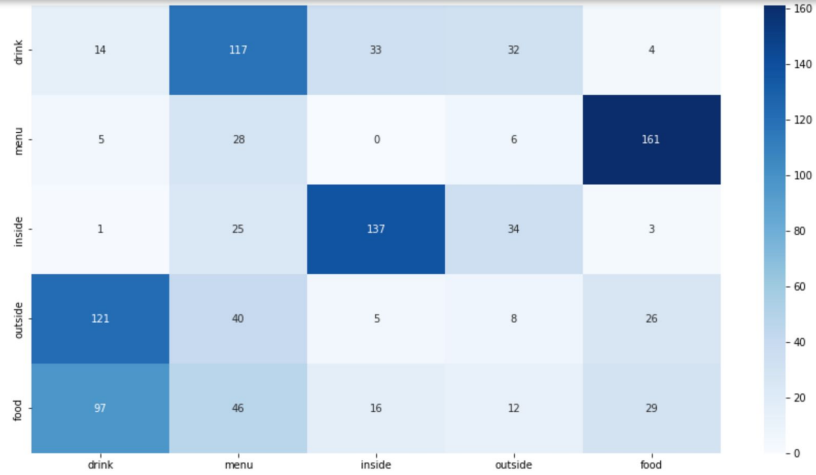
SIFT



Catégorie	N°
Drink	0
Menu	1
Food	2
Outside	3
Inside	4

	precision	recall	f1-score	support
0	0.23	0.29	0.26	200
1	0.05	0.06	0.05	200
2	0.02	0.01	0.02	200
3	0.09	0.12	0.10	200
4	0.06	0.04	0.04	200
accuracy			0.10	1000
macro avg	0.09	0.10	0.09	1000
weighted avg	0.09	0.10	0.09	1000

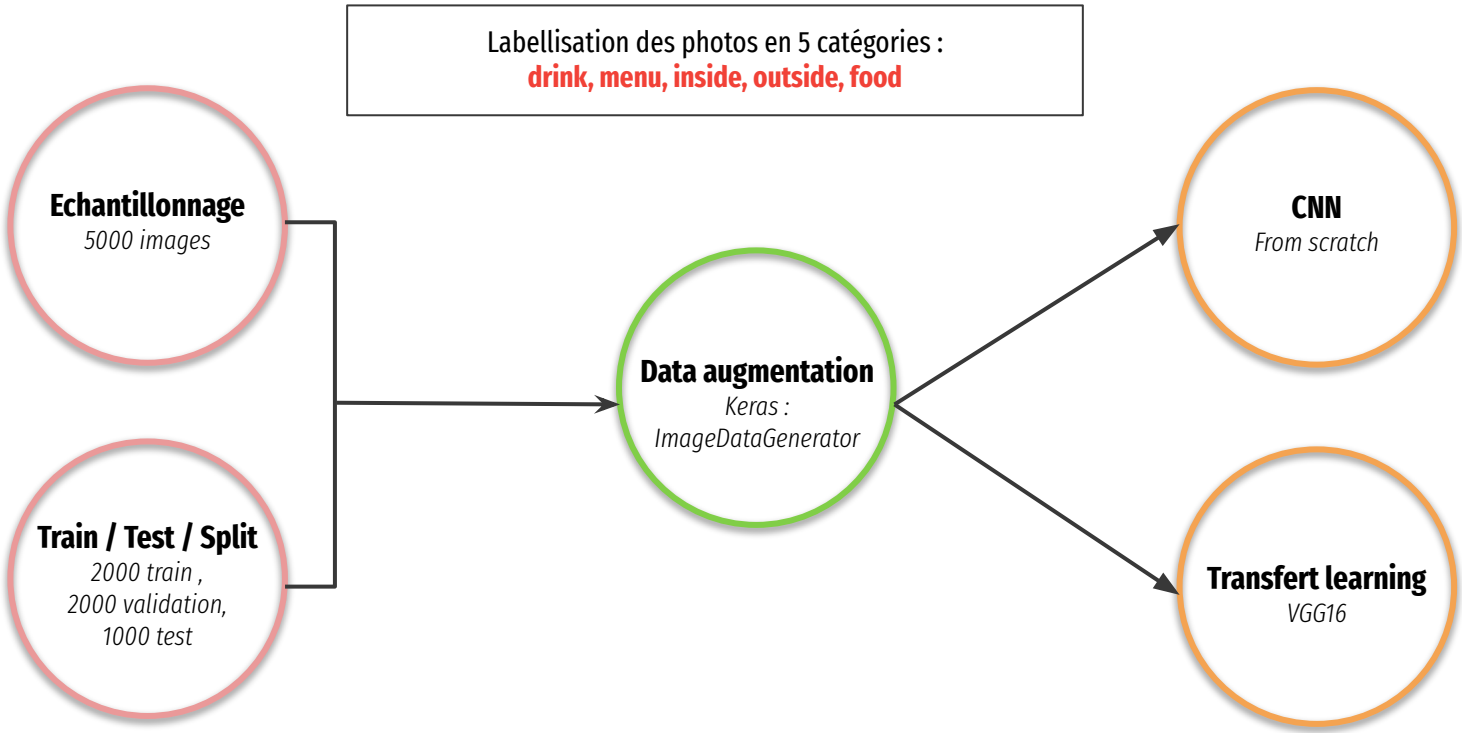
VGG16



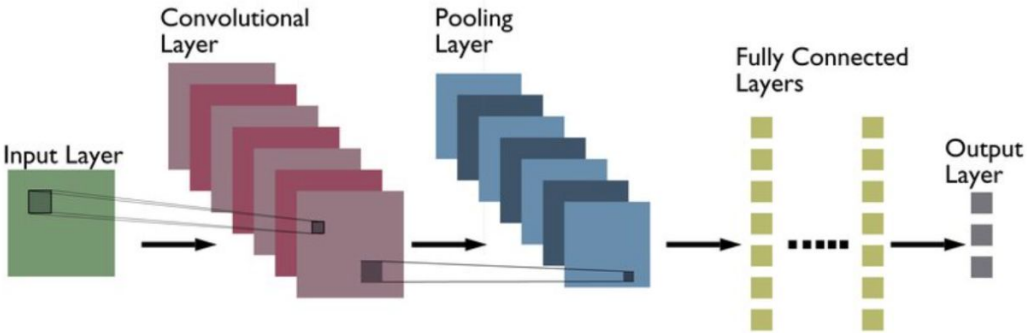
Catégorie	N°
Drink	0
Menu	1
Food	2
Outside	3
Inside	4

	precision	recall	f1-score	support
0	0.06	0.07	0.06	200
1	0.11	0.14	0.12	200
2	0.72	0.69	0.70	200
3	0.09	0.04	0.05	200
4	0.13	0.14	0.14	200
accuracy			0.22	1000
macro avg	0.22	0.22	0.22	1000
weighted avg	0.22	0.22	0.22	1000

Démarche de traitement des images



Apprentissage : CNN



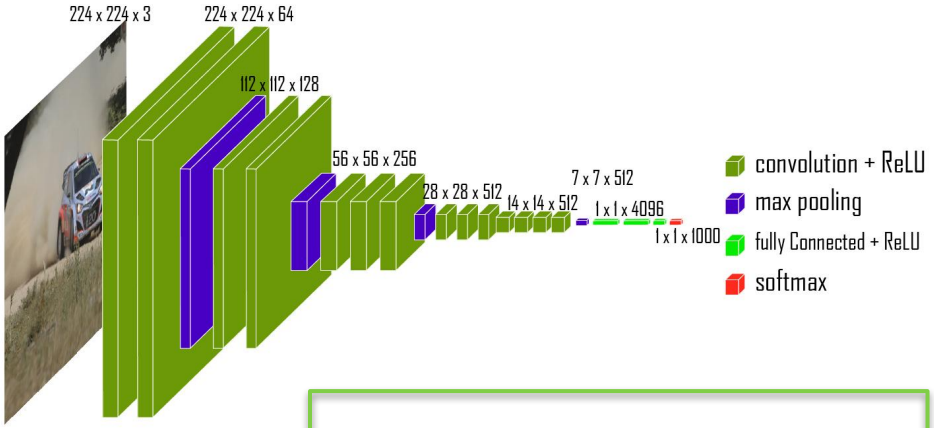
- ★ **tâches de classification, de détection et segmentation d'images**
- ★ **méthode incontournable pour les prédictions impliquant n'importe quelle image en entrée.**
- ★ **Précision des résultats très élevées**

Layer (type)	Output Shape	Param #
conv2d_9 (Conv2D)	(None, 224, 224, 32)	896
max_pooling2d_9 (MaxPooling 2D)	(None, 112, 112, 32)	0
dropout_9 (Dropout)	(None, 112, 112, 32)	0
conv2d_10 (Conv2D)	(None, 112, 112, 64)	18496
max_pooling2d_10 (MaxPoolin g2D)	(None, 56, 56, 64)	0
dropout_10 (Dropout)	(None, 56, 56, 64)	0
conv2d_11 (Conv2D)	(None, 56, 56, 128)	73856
max_pooling2d_11 (MaxPoolin g2D)	(None, 28, 28, 128)	0
dropout_11 (Dropout)	(None, 28, 28, 128)	0
flatten_3 (Flatten)	(None, 100352)	0
dense_3 (Dense)	(None, 5)	501765

Total params: 595,013
Trainable params: 595,013
Non-trainable params: 0

None

Apprentissage : Transfer learning *(VGG16)*



- ★ **16 couches d'apprentissage**
- ★ **139 millions de paramètres**
- ★ **Classification en 1000 catégories**

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 224, 224, 3)]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten_1 (Flatten)	(None, 25088)	0
dense_1 (Dense)	(None, 128)	3211392
dense_2 (Dense)	(None, 5)	645

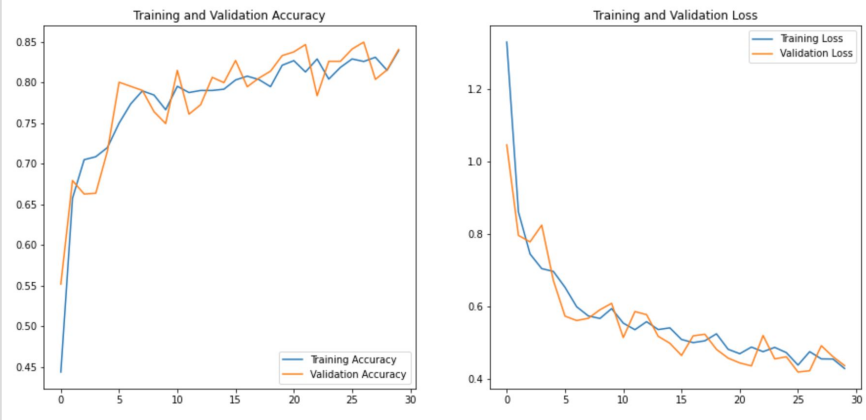
Total params: 17,926,725
Trainable params: 3,212,037
Non-trainable params: 14,714,688

Comparaison des modèles de classification CNN et VGG16

CNN

La précision estimée est de 56.400 %

VGG16



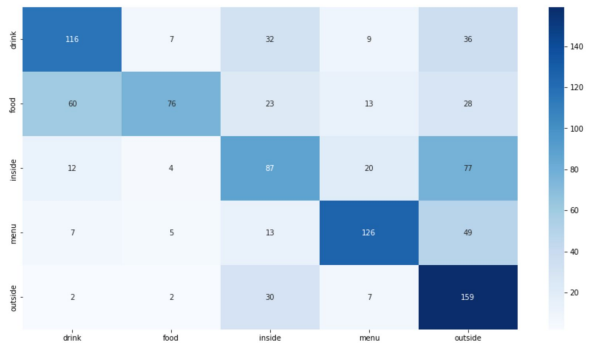
La précision estimée est de 82.200 %

Comparaison finale

SIFT

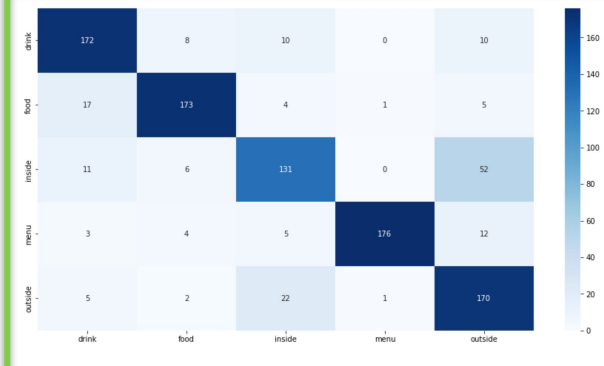
Catégorie	N°		precision	recall	f1-score	support
Drink	0	0	0.25	0.32	0.28	200
Menu	1	1	0.79	0.54	0.64	200
Food	2	2	0.55	0.67	0.60	200
Outside	3	3	0.18	0.13	0.15	200
Inside	4	4	0.24	0.27	0.26	200
		accuracy			0.39	1000
		macro avg	0.40	0.39	0.39	1000
		weighted avg	0.40	0.39	0.39	1000

CNN



Catégorie	N°		precision	recall	f1-score	support
Drink	0	0	0.59	0.58	0.58	200
Menu	1	1	0.81	0.38	0.52	200
Food	2	2	0.47	0.43	0.45	200
Outside	3	3	0.72	0.63	0.67	200
Inside	4	4	0.46	0.80	0.58	200
		accuracy			0.56	1000
		macro avg	0.61	0.56	0.56	1000
		weighted avg	0.61	0.56	0.56	1000

VGG16



Catégorie	N°		precision	recall	f1-score	support
Drink	0	0	0.83	0.86	0.84	200
Menu	1	1	0.90	0.86	0.88	200
Food	2	2	0.76	0.66	0.70	200
Outside	3	3	0.99	0.88	0.93	200
Inside	4	4	0.68	0.85	0.76	200
		accuracy			0.82	1000
		macro avg	0.83	0.82	0.82	1000
		weighted avg	0.83	0.82	0.82	1000


voilà

Conclusion

	Resultats	Axe d'amélioration
NLP	➤ **Interprétation de l'analyse des sujets d'insatisfaction pas évidente.** ➤ **Récupération de 200 avis sur la ville de New-York**	➤ **faire une analyse de sentiment afin de détecter si un commentaire est plutôt positif ou négatif.** ➤ **Filtrer les avis par langues.** ➤ **Essaie avec d'autres méthodes** *(LSA, PLSA, lda2Vec, ...).*
Images	➤ **Bonne précision sur la classification avec les modèles CNN et VGG16.** ➤ **Manque de ressource pour l'entraînement** *(gpu).*	➤ **Finetune le modèle VGG16** (en utilisant un classifieur XGBoost). ➤ **Améliorer les modèles en les complexifiant** *(ajout de couches, ...).*

MERCI !

Avez-vous des questions ?