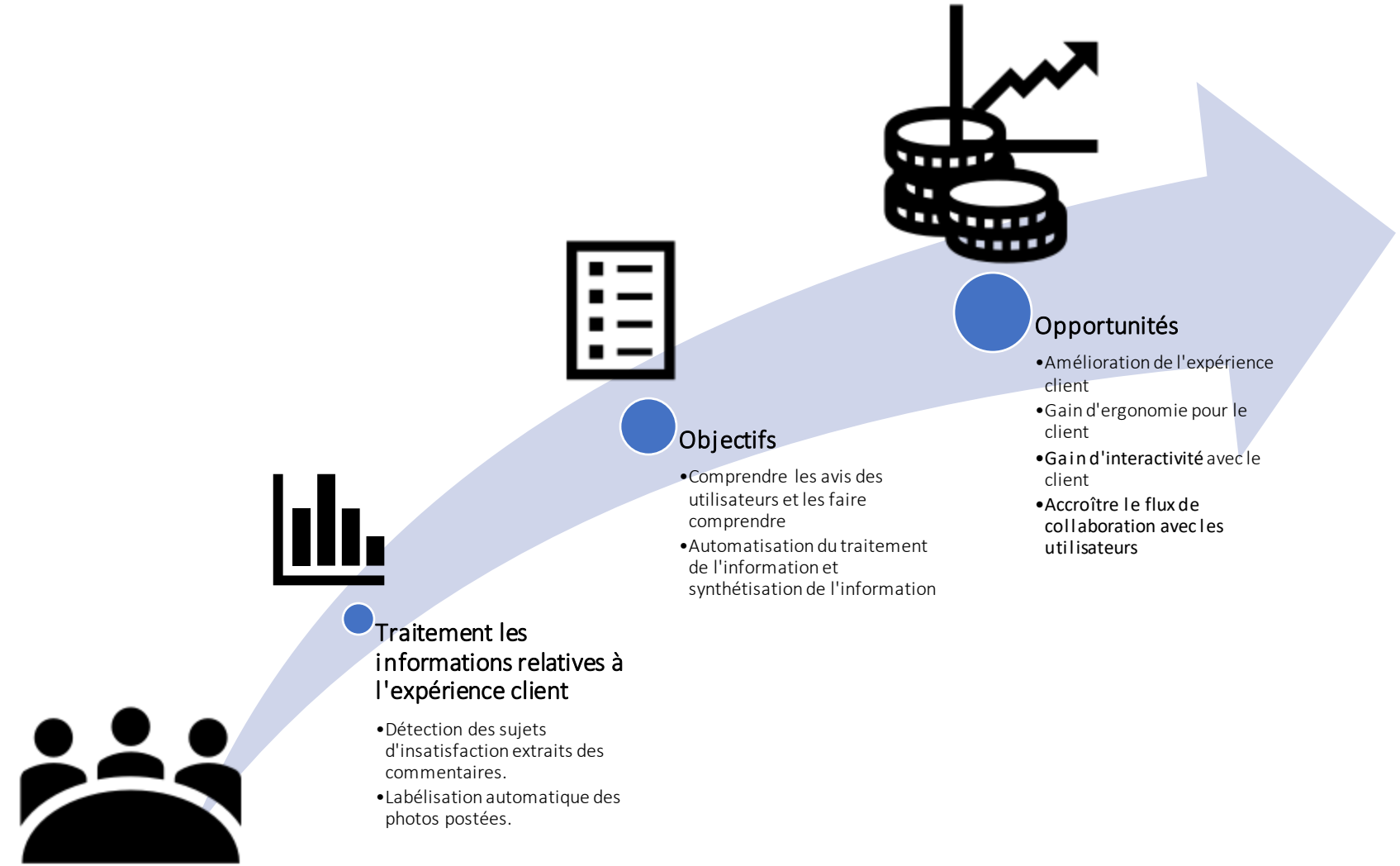
Détection de sujets d'insatisfaction et labélisation automatique de photos

Start-up AVIS RESTAU

Bailly DIOUNOU, Ingénieur IA – 25/06/2021

Contexte & périmètre du projet

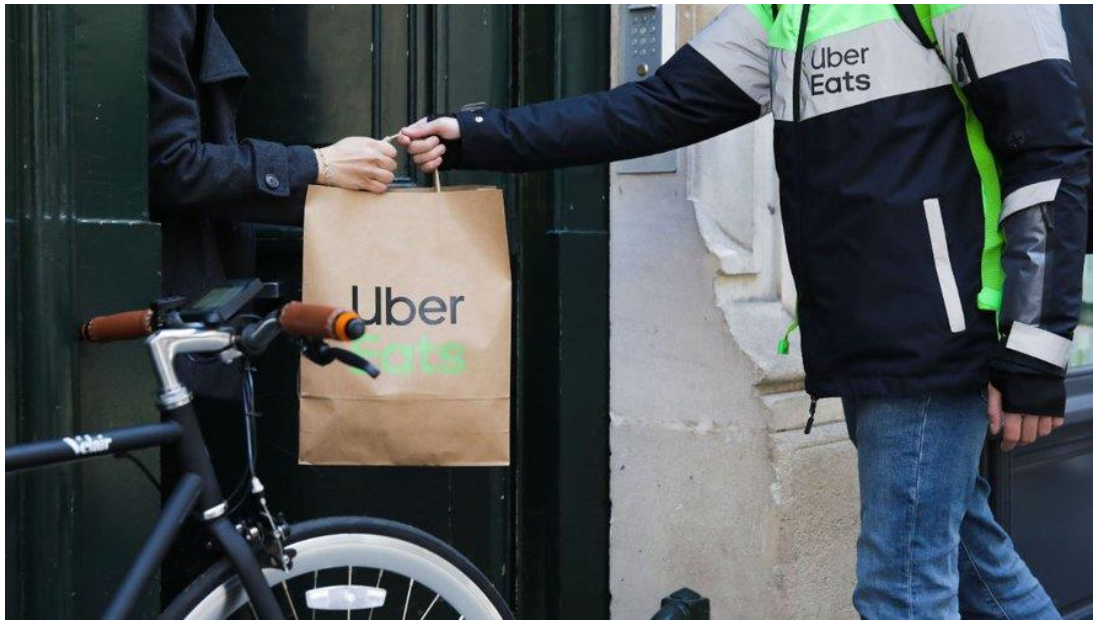


Présentation de la problématique métier

Interprétation & pistes de recherche

Problématique métier

Contexte sanitaire

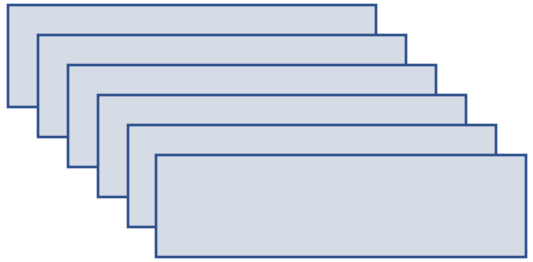


Problématique métier

L'affaire " The Shed at Dulwich"

Restaurants

Clients



Problématique métier

L'affaire " The Shed at Dulwich"

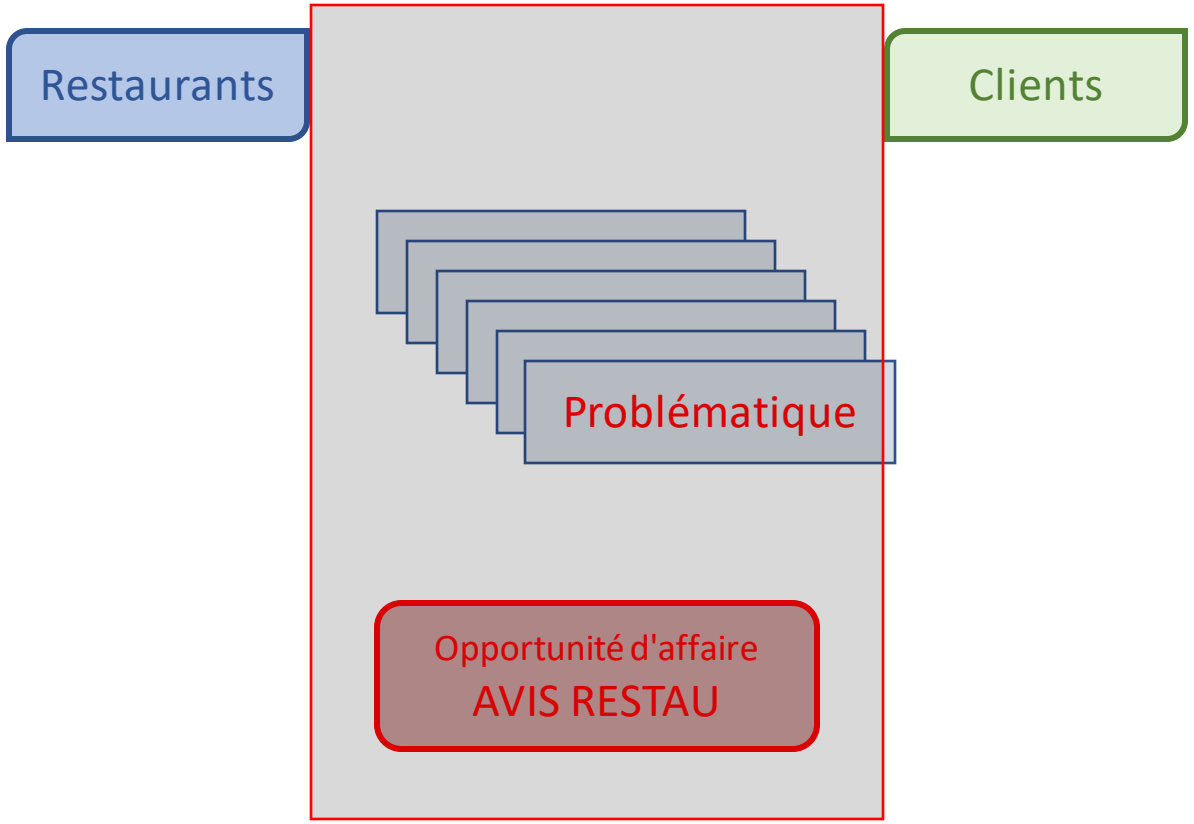
Restaurants

Clients

Problématique

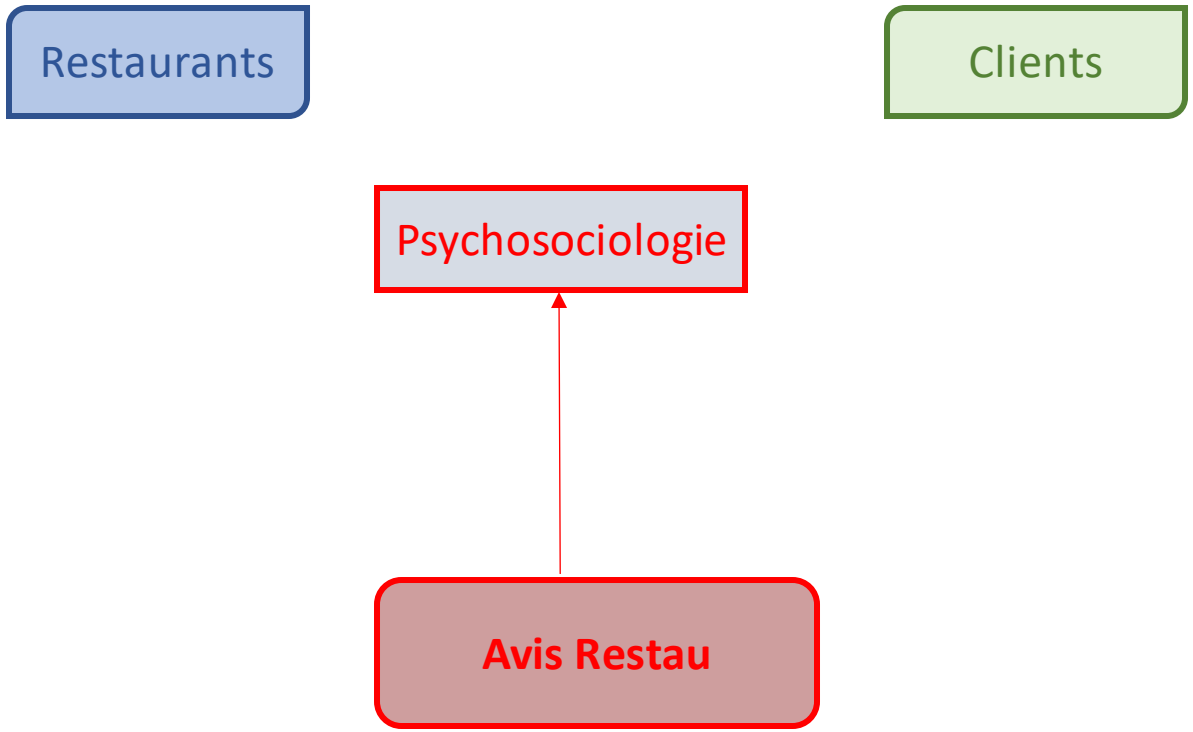
Problématique métier

L'affaire " The Shed at Dulwich"



Problématique métier

Psychosociologie de la consommation

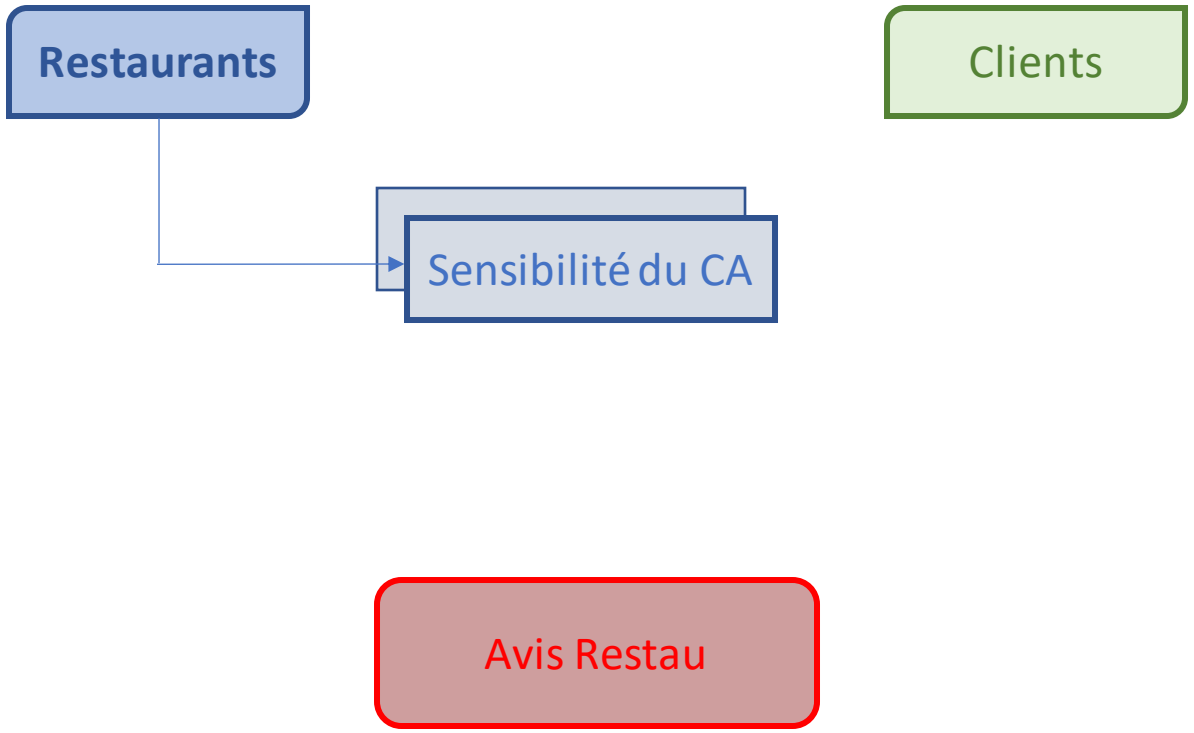


Les usagers postent-ils vraiment *que* lorsqu'ils sont (1)très mécontents, (2) très mécontents ou très contents ?

Quelles raisons principales : qualité, accueil/contact, rapport qualité-prix ?

Problématique métier

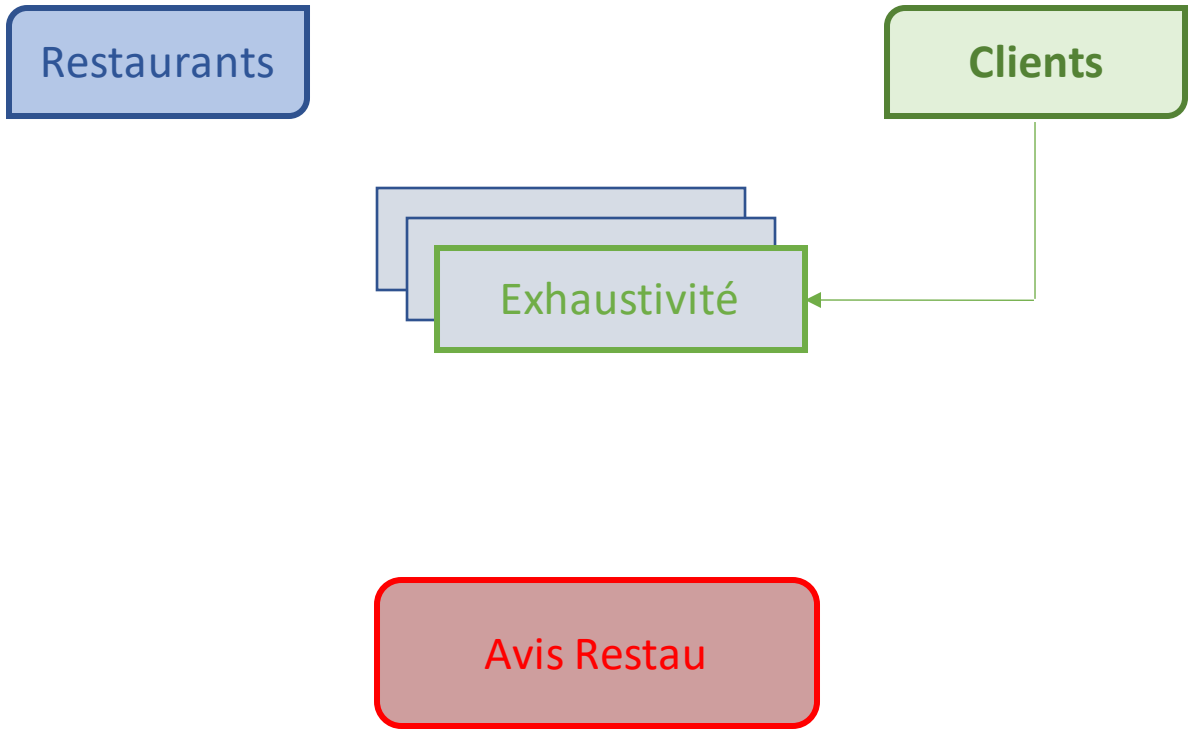
Sensibilité du chiffre d'affaire aux avis clients



Influence des avis négatifs accumulés sur la décision d'achat

Problématique métier

Exhaustivité de l'investigation avant l'achat

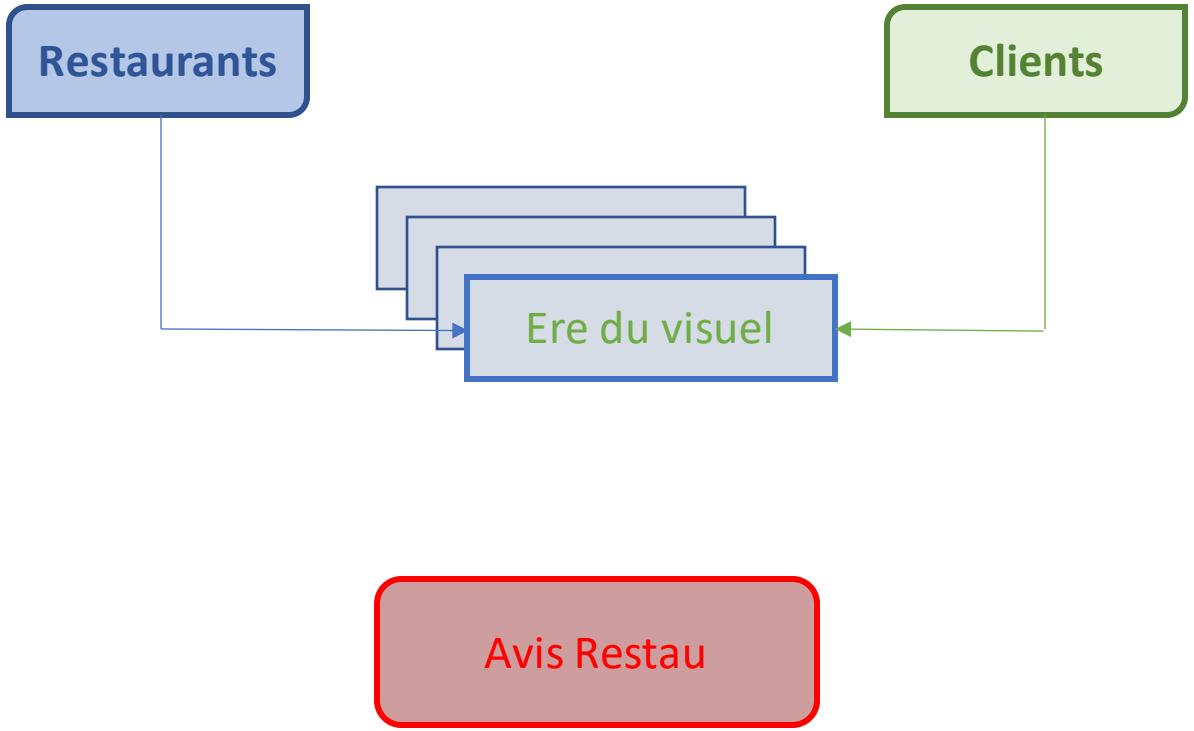


Pour mon futur pot de thèse: vais-je lire tous les commentaires de ce traiteur? Pour tous les traiteurs de la commune? du département ?

Attention ! Ton introduction de manuscrit de thèse ne portera pas sur les raisons de l'augmentation saisonnière du kilo de sorbet juste avant l'été...

Problématique métier

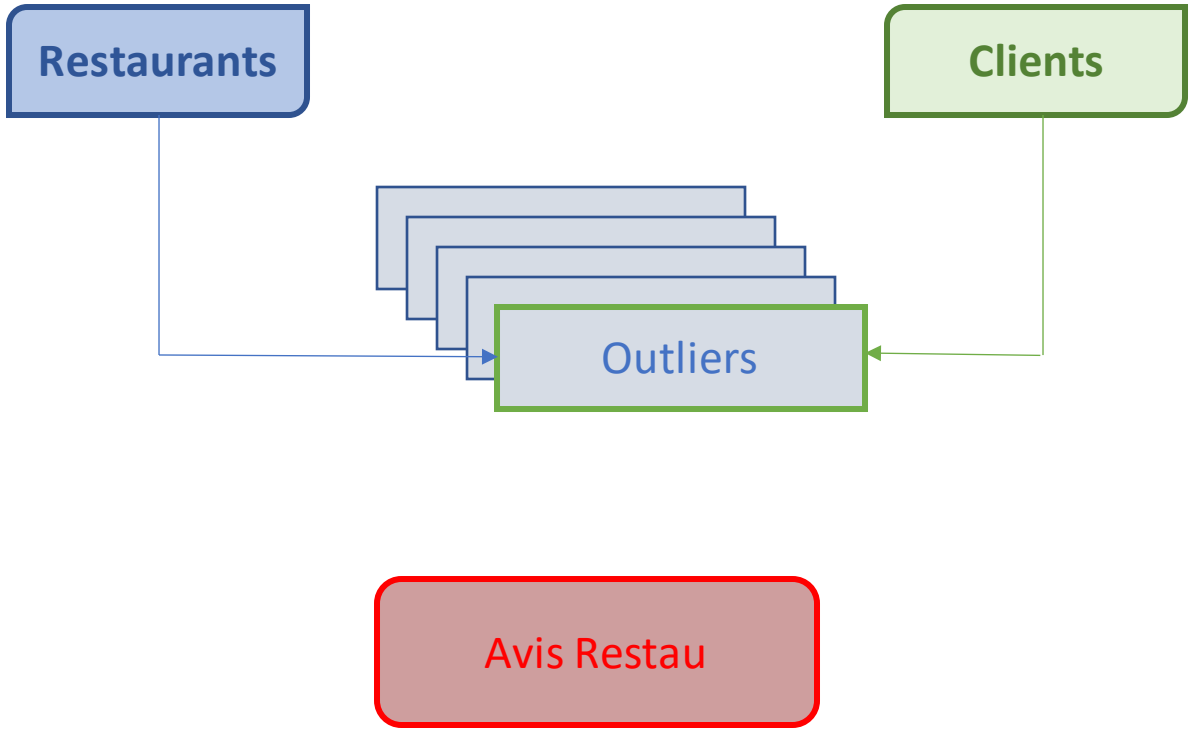
Ere du visuel et de l'image



Simple curiosité des sens

ou

Complément d'aide à la décision d'achat ?



Client mécontent,

Concurrent malveillant

ou... Simple rageux ?

Description des jeux de données

Description du jeu de données

Commentaires clients

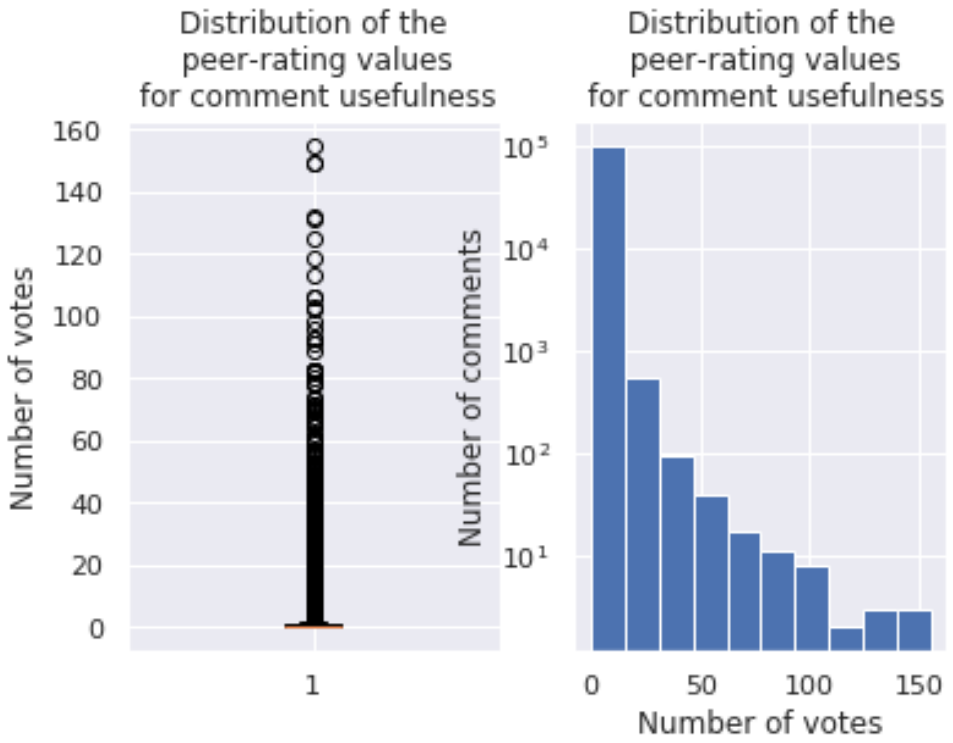
- 8M de commentaires
- 2M de clients
- 209K restaurants

Description du jeu de données

Commentaires clients

- **8M de commentaires**
  - ***useful***, *funny, cool*

  Peer-rating of "quality"
- 2M de clients
- 209K restaurants

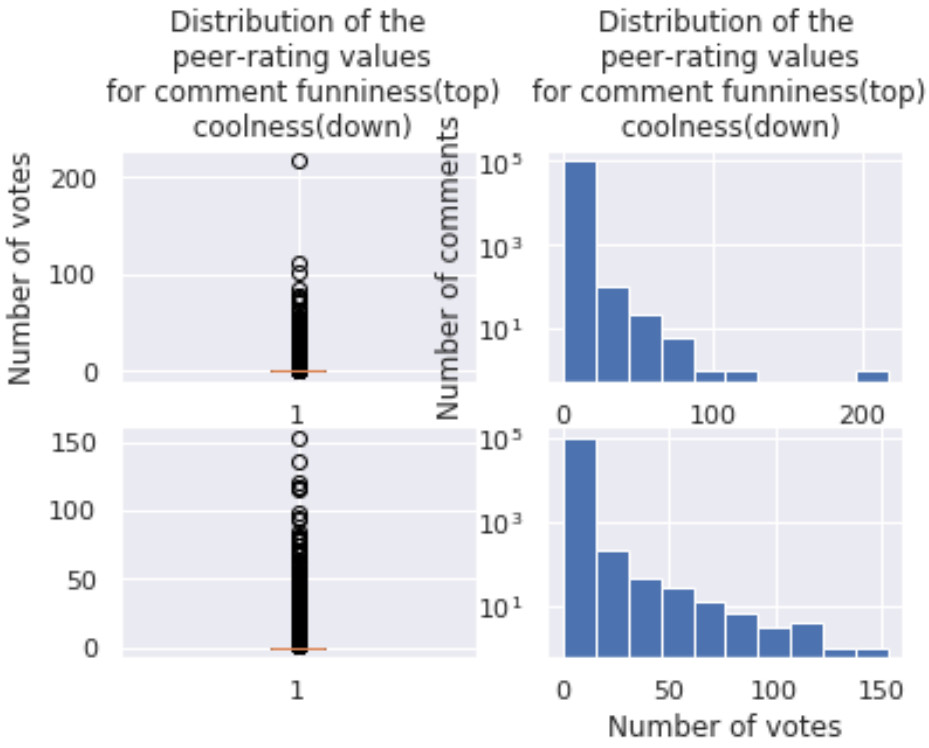


Description du jeu de données

Commentaires clients

- **8M de commentaires**
  - *useful,* ***funny, cool***

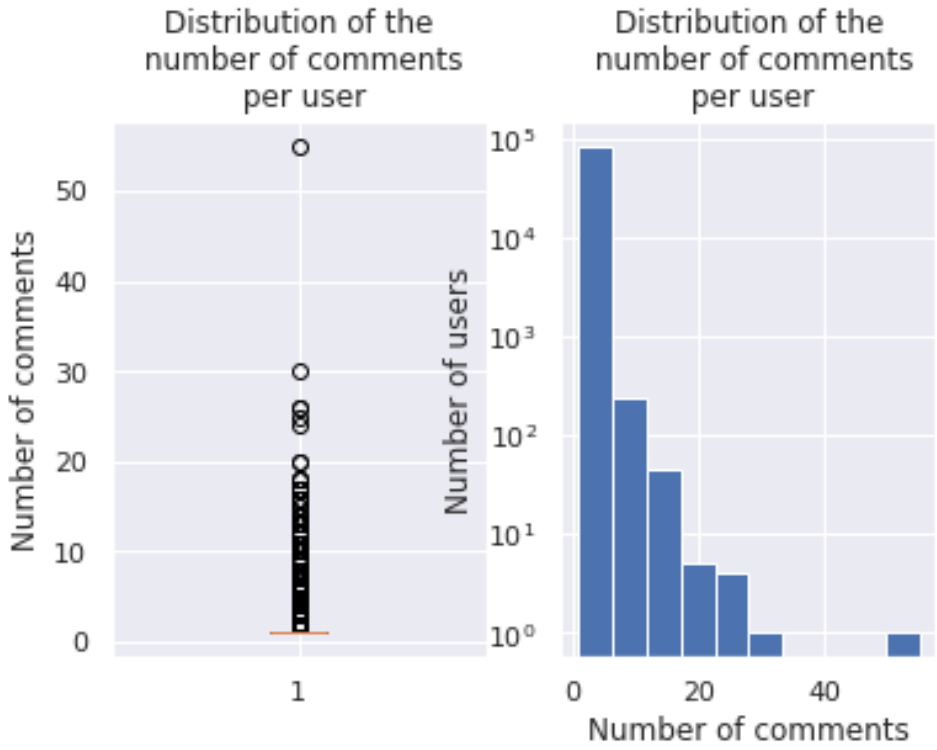
  Peer-sentiment analysis
- 2M de clients
- 209K restaurants



Description du jeu de données

Données clients

- 8M de commentaires
- **2M de clients**
  - Rapport ~ 1/10^3 : effectif mode/autres modalités -> **Peu d'utilisateurs commentent plus d'une fois!**
- 209K restaurants



Description du jeu de données

Données clients

- 8M de commentaires
- **2M de clients**
  - Peu d'utilisateurs commentent plus d'une fois.
  - **Commentaires plutôt positifs !**
- 209K restaurants

Description du jeu de données

Données clients

- 8M de commentaires
- 2M de clients
- **209K restaurants**
  - **Com/restau: (moy, std) = (38, 127)**

    -> Faible nombre de com/restau

Description du jeu de données

Données clients

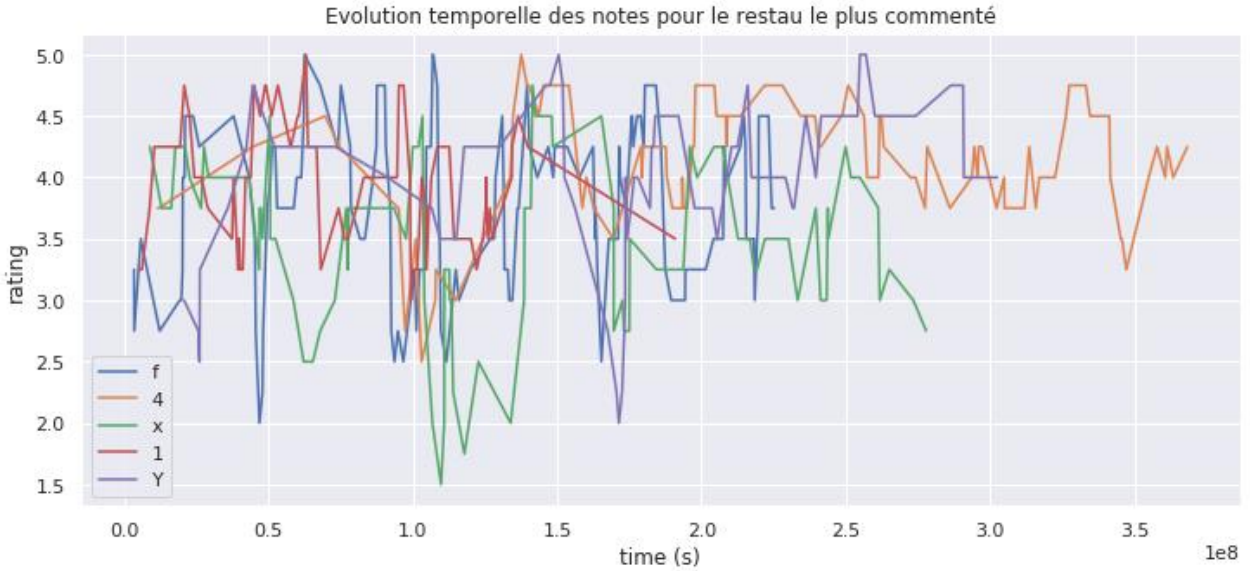
- 8M de commentaires
- 2M de clients
- **209K restaurants**
  - Com/restau: (moy, std) = (38, 127)

    -> Faible nombre de com/restau
  - **Les plus commentés le sont positivement!**

Description du jeu de données

Données clients

- 8M de commentaires
- 2M de clients
- **209K restaurants**
  - Com/restau: (moy, std) = (38, 127)
    -> Faible nombre de com/restau
  - Les plus commentés le sont positivement!
  - **Analyse prédictive: sensibilité du CA**
    - Quasi-stationnarité native (fréquence!)



Description du jeu de données

Commentaires clients

- **8M de commentaires**
  - *useful, funny, cool*

  Peer-group data
- 2M de clients
- 209K restaurants

Consumer peer-group data crowdsourcing:
Où se situe-t-on sur le DIKW ?

Vulnérabilité, *Menaces & Risques*

- Taille et complétude médiocre des jeux de données Avis Restau
  -> Risque sur l'objectivité et la pertinence de l'information, et l'intégrité du service
- Ere du deep fake : créations de profils fictifs Ton "Shed at Dulwich" from home !!!
  -> Risque sur l'intégrité de l'information et du service

Description du jeu de données

Commentaires clients

- **8M de commentaires**
  - *useful, funny, cool*

  Peer-group data
- 2M de clients
- 209K restaurants

Consumer peer-group data crowdsourcing:
Où se situe-t-on sur le DIKW ?

Vulnérabilité, Menaces & Risques

Peer-group data à la trappe ?

Description du jeu de données

Commentaires clients

- **8M de commentaires**
  - *useful, funny, cool*

  Peer-group data
- 2M de clients
- 209K restaurants

Consumer peer-group data crowdsourcing:
Où se situe-t-on sur le DIKW ?

Mesures \ Solutions

- DIC to DIC ?

Description du jeu de données

Commentaires clients

- **8M de commentaires**
  - *useful, funny, cool*

  Peer-group data
- 2M de clients
- 209K restaurants

Consumer peer-group data crowdsourcing:
Où se situe-t-on sur le DIKW ?

Mesures \ Solutions

- DIK to DIC ?

Renforcement de l'authenticité

- Authentification
- Non-répudiation

Traitement des données "crowdsourced"

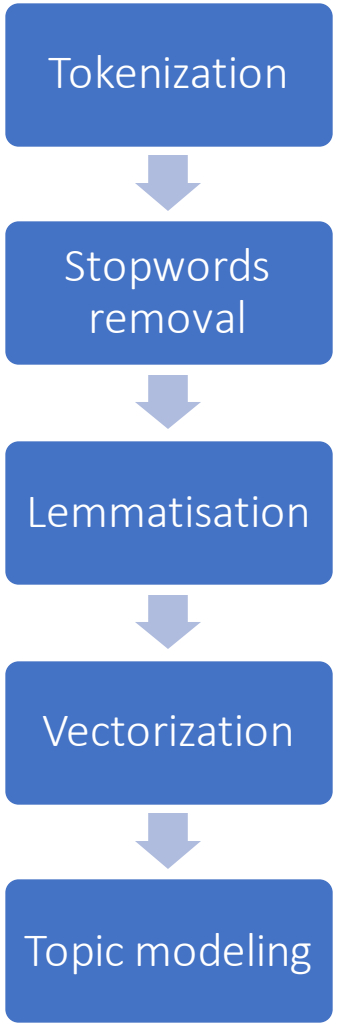
Pipelines pour commentaires et photos

Traitements des données

Commentaires

Filtres pré-traitement

- Sélection des commentaires négatifs (après binarisation de la note de satisfaction – 0 à 5)
- Sélection des commentaires anglais

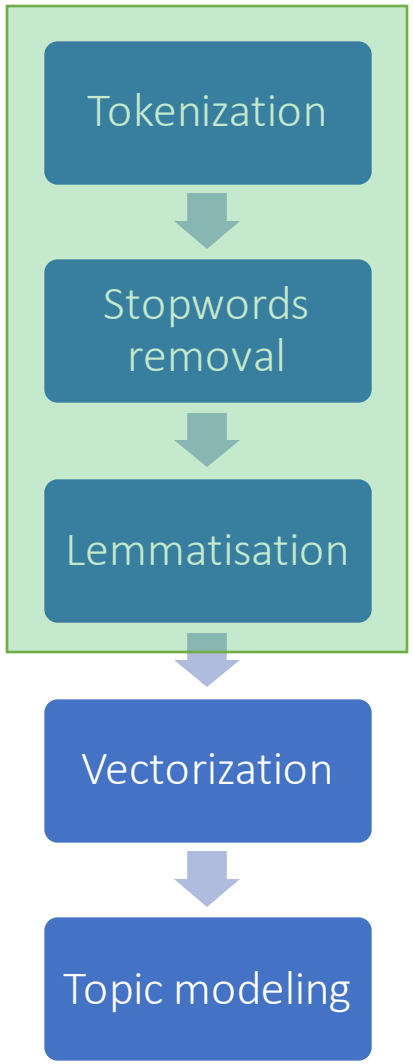


Traitements des données

Commentaires

Pipeline

- **Nettoyage & Normalisation avec NLTK**

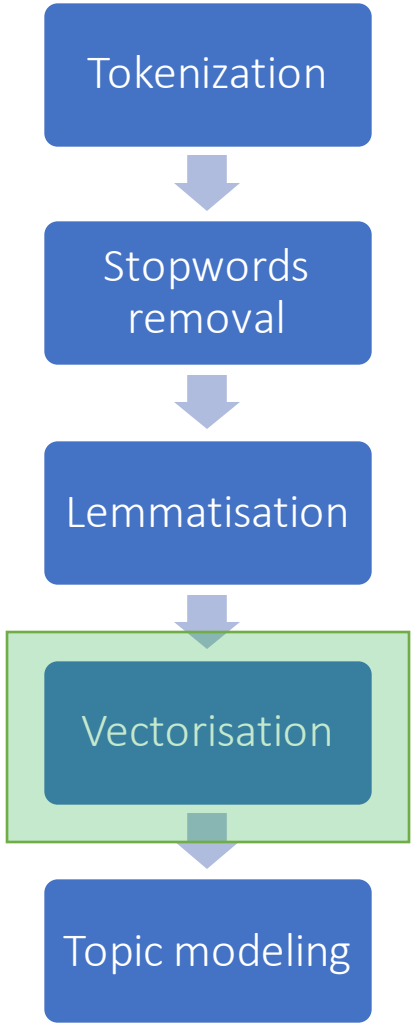


Traitements des données

Commentaires

Pipeline

- Nettoyage & Normalisation, NLTK
- **Vectorisation, Tf-idf**
  - CountVectorizer (max_features = 1000)

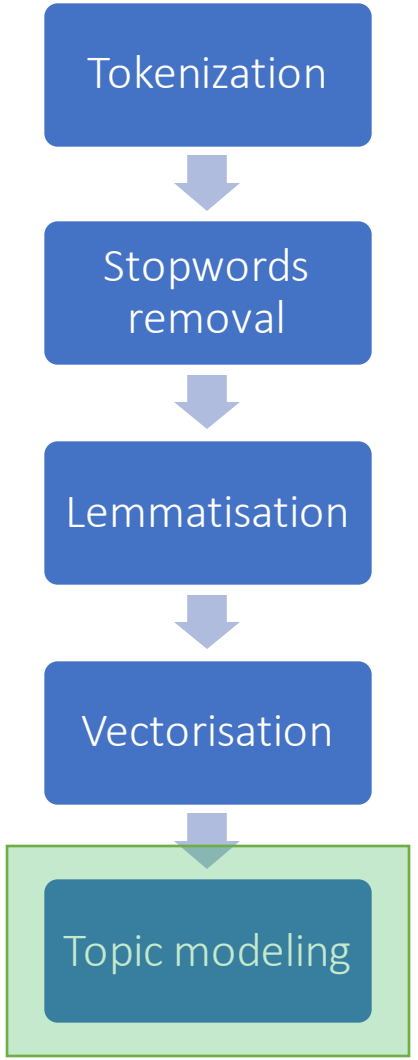


Traitements des données

Commentaires

Pipeline

- Nettoyage & Normalisation, NLTK
- Vectorisation, Tf-idf
- **Topic modeling, NMF pré-entrainé**
  - Nombre de topic K optimisé par coherence score (spacy similarity)



Traitements des données

Commentaires

Pipeline

- Nettoyage & Normalisation, NLTK
- Vectorisation, Tf-idf
- **Topic modeling, NMF**
  - Nombre de topic K optimisé par coherence score (spacy similarity)
  - Exemple:

    K dans 2:10 -> **K_opt = 8 (**coh=0.87)

	Topic # 01 Poor order process	Topic # 02 Poor food	Topic # 03 Poor service	Topic # 04 Poor delay	Topic # 05 Poor pizza	Topic # 06 Poor car accomodation	Topic # 07 Poor accomodation	Topic # 08
0	back	food	service	order	pizza	car	room	place
1	said	chicken	customer	minute	ordered	wash	hotel	worst
2	told	good	horrible	wait	slice	oil	stay	like
3	call	ordered	rude	time	cheese	drive	desk	ever
4	never	restaurant	worst	drink	sauce	change	night	people
5	time	eat	terrible	waited	tasted	dealership	check	star
6	called	quality	ever	hour	taste	vehicle	bed	bad
7	day	cold	bad	table	delivery	get	front	even
8	get	like	store	drive	like	tire	bathroom	really
9	company	taste	poor	waiting	wing	take	stayed	give

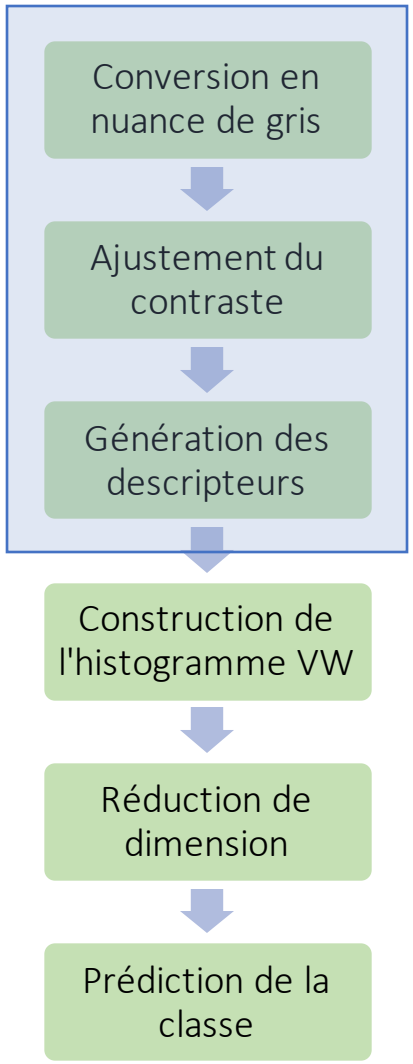
Traitements des données

Photos

Pipeline

- **Normalisation et génération des descripteurs de l'image, SIFT**
  - Echantillon de 3000 images
  - Nuances de Gris <- SIFT
  - Ajustement de contraste
  - Max des = 500

  --> **1.5M des**

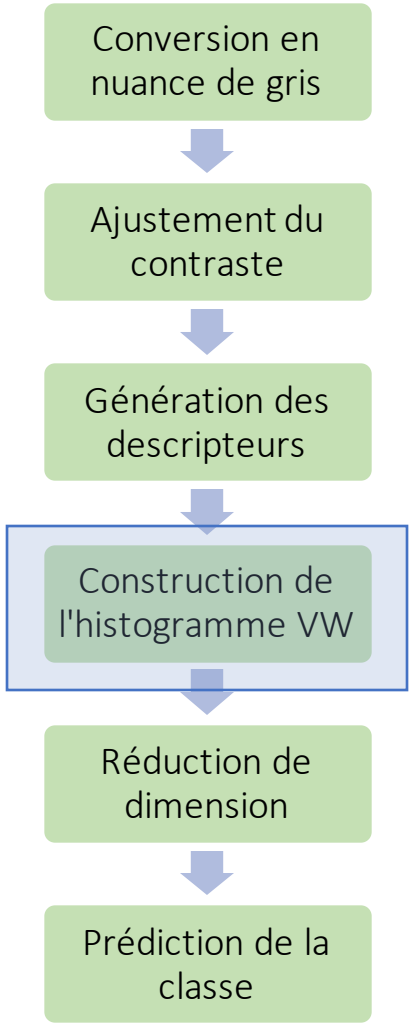


Traitements des données

Photos

Pipeline

- Normalisation et génération des descripteurs de l'image, SIFT
- **Construction de l'histogramme des mots visuels, KMeans pré-entrainé**
    - K in [N*#Etiq ; **sqrt(Nb total des)**]

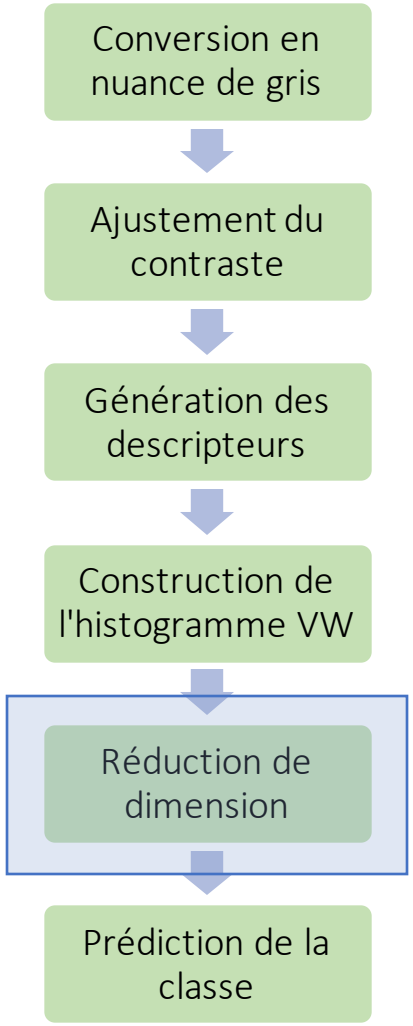


Traitements des données

Photos

Pipeline

- Normalisation et génération des descripteurs de l'image, SIFT
- Construction de l'histogramme des mots visuels, KMeans pré-entrainé
- **Réduction de dimension, PCA pré-entrainé**
    - 60% var. expl.
    - 1212 -> 141 feat

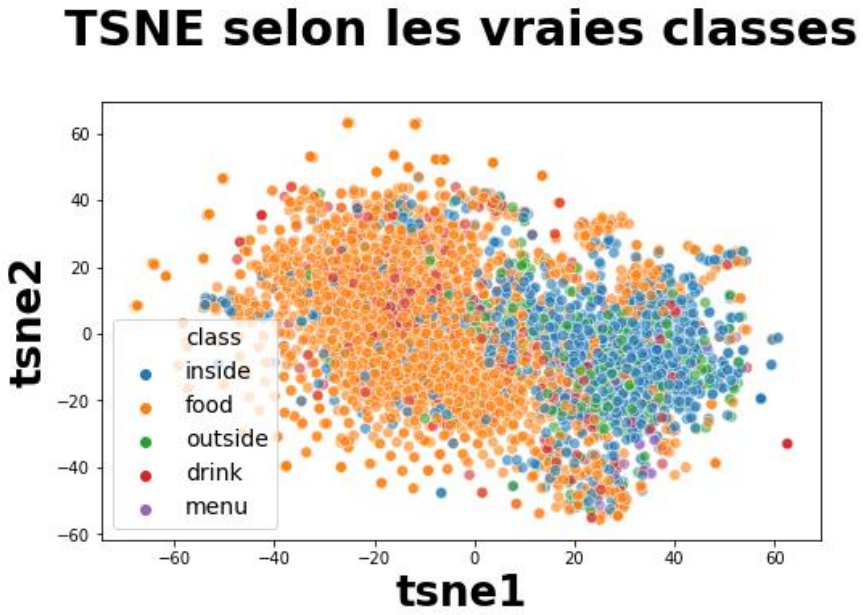


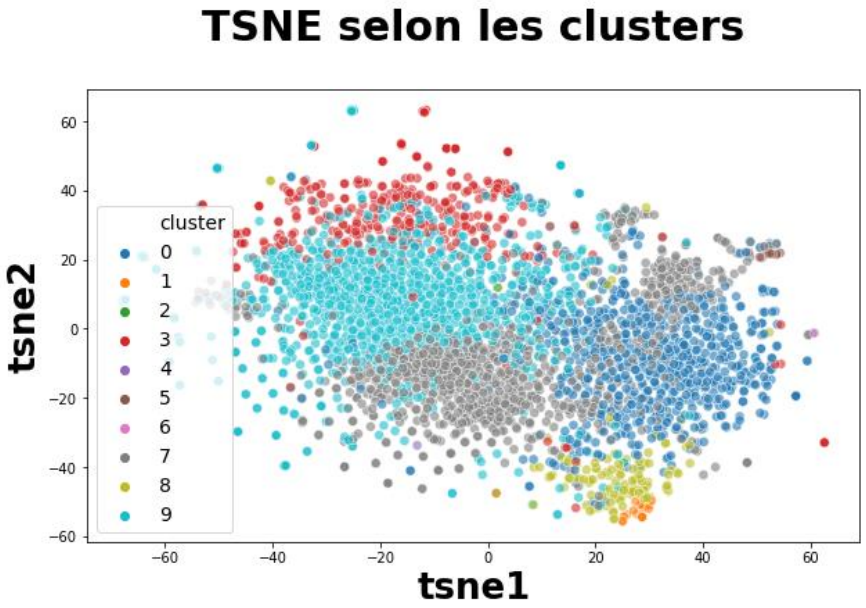
Traitements des données

Photos

Pipeline

- Normalisation et génération des descripteurs de l'image, SIFT
- Construction de l'histogramme des mots visuels, KMeans pré-entrainé
- **Réduction de dimension, PCA pré-entrainé**
    - 60% var. expl.
    - 1212 -> 141 feat
    - Clustering Kmeans sur les images dans l'espace de sortie PCA



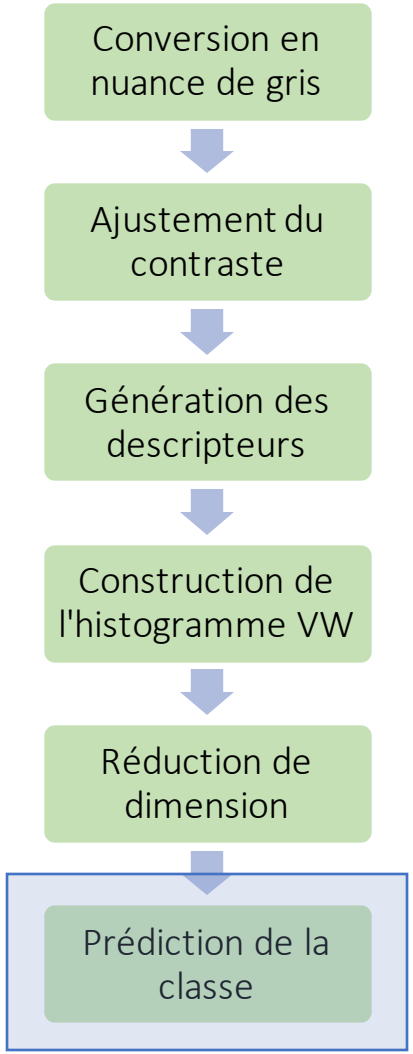


Traitements des données

Photos

Pipeline

- Normalisation et génération des descripteurs de l'image, SIFT
- Construction de l'histogramme des mots visuels, KMeans pré-entrainé
- Réduction de dimension, PCA pré-entrainé
- **Prédiction de la classe, LinearSVC pré-entrainé**
    - •Classification Clusters PCA vs Etiquettes
    - •plusieurs classifieurs testés (LR, LDA,CART KNN, RF, SVM)
    - •gridsearch/classifieur
    - •linear_svc {recall:0.499, f1_score:0.507}



Visualisation graphique de la solution

Tableau de bord web

Conclusion sur la faisabilité

- Detection de sentiments
  - Topics définis, avec distance inter-topic bien claire
  - Bonne scalabilté envisagée, montée en échelle conseillée !
- Classification d'image
  - Classification peu performante (F1_score ~ 50%)
  - Test d'une méthode plus efficace avant décision (Ex: CNNs)

Merci pour votre attention

Disponible pour des questions/réponses