Classification automatique des articles

Étude de faisabilité

État des lieux

Multiples vendeurs
+
Catégorisation manuelle

Catégorisation
peu fiable et
donc inutilisable

Mission

Étude de faisabilité d'un moteur de classification automatique des biens de consommations à partir de la description du bien

Outils à tester

2 approches Bag-of-words dont Tf-idf
Word2Vec / Glove / FastText
BERT
USE (Universal Sentence Encoder)
SIFT / ORB / SURF
CNN Transfer Learning

Données à disposition

Un fichier de données sur 1050 articles accompagné de 1050 photos

Pour l'étude : utilisation des catégories indiquées manuellement par les vendeurs

Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion

Anna Andre Paris Combos
Printland Coffee Mugs
Lilliput Solid Baby Boy's Blue Cargo Shorts
Wow! Dresses
Bhavya Velvet Sofa Cover (Grey Pack of 10)
JayEnterprise Hand Juicers

1050 produits
Répartis dans de 2 à 7 niveaux de catégories

Niveau 2 : 62 sous-catégories, répartition des produits non équilibrée

Niveau 1 : 7 catégories, répartition des produits équilibrée

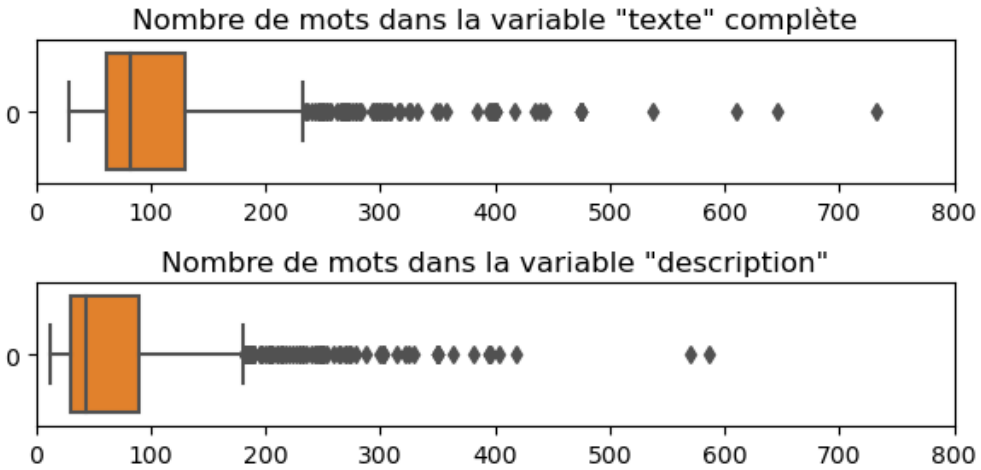
Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion

Features textuelles

- Nom
- Spécifications

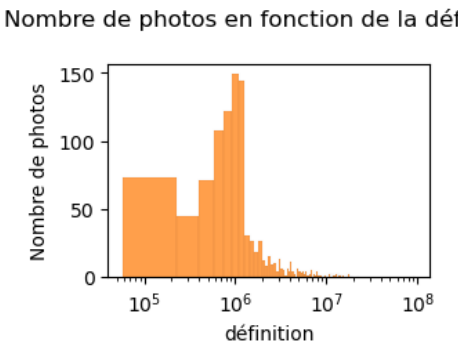
1015 entrées différentes
Une seule commune à presque tous les articles

- Description



Features images

- Photographie



Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion

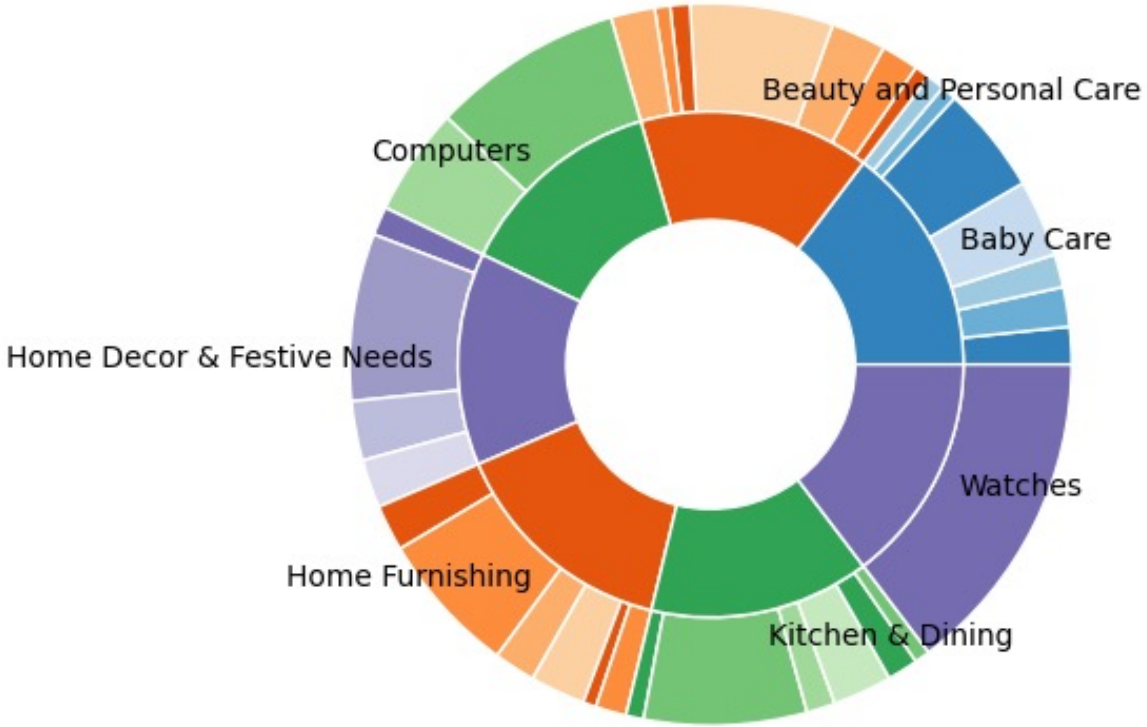
Étapes - Général	Lignes restantes
Suppression des produits avec une valeur NaN à 'product_category_tree', 'specifications', 'nom' ou 'description'	1049

Étapes - Target	Lignes restantes
Extraction de chaque niveau de catégorie	1049
Réimputation de certaines catégories de niveau 2 sous-représentées	1049
Suppression des catégories de niveau 2 de moins de 6 individus	1015
Création de 2 targets : target_wide pour catégorie de niveau 1 et target pour catégorie de niveau 2 et leurs pendants numériques target_wide_num et target_num réalisés avec LabelEncoder	1015

→ Suppression de l'unique poussette, de l'unique cocotte minute…

Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion

- Catégories à classifier
  - 1er niveau sur l'ensemble des données : 7 classes équilibrées
  - 2ème niveau sur l'ensemble des données : 33 classes
  - 2ème niveau pour les catégories de plus de 20 individus : 15 classes
  - 2ème niveau pour les catégories de plus de 45 individus : 8 classes



Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion

*Mu : Mots uniques

Étapes – Features textuelles

Récupération des textes inclus dans les spécifications, sans les items d'en-têtes

Concaténation de ces textes avec product_name et description → feature texte

Feature description non retravaillée pour l'instant

Tokenization - Lemmatization

Feature texte – 3183 Mu*

Feature description – 2866 Mu*

Retrait des mots rares (<5)

Feature texte – 1404 Mu*

Feature description – 1078 Mu*

Retrait des doublons selon target_wide

Feature texte – 1387 Mu*

Feature description – 1061 Mu*

Tokenization - Stemming

Feature texte – 2892 Mu*

Feature description – 2620 Mu*

Retrait des mots rares (<5)

Feature texte – 1371 Mu*

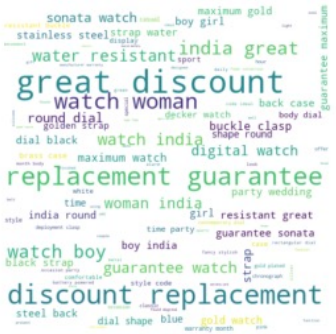
Feature description – 1069 Mu*

Retrait des doublons selon target_wide

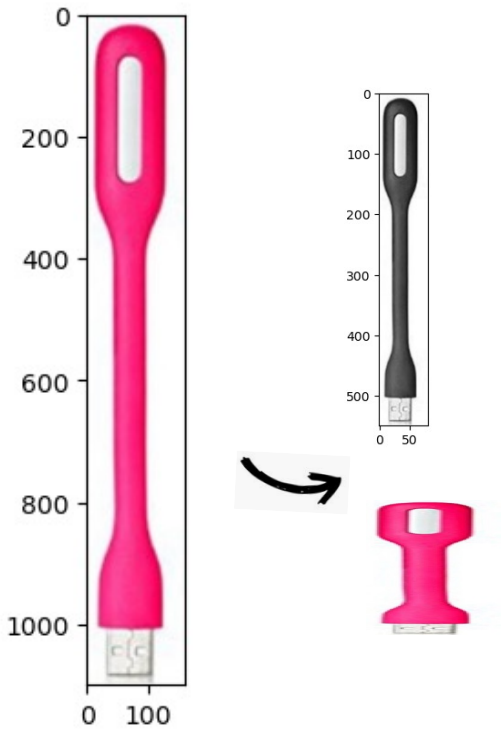
Feature texte – 1355 Mu*

Feature description – 1053 Mu*

Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion

Description_lem	Texte_lem	Description_stem	Texte_stem
Home Furnishing	Home Furnishing	Home Furnishing	Home Furnishing
			
Watches	Watches	Watches	Watches
			
Beauty and Personal Care	Beauty and Personal Care	Beauty and Personal Care	Beauty and Personal Care
			

Étapes – Features images	Lignes restantes
Suppression de l'article tapis de souris Apple pour lequel SIFT ne voit pas de keypoint avec les paramètres par défaut	1014
Passage des images en niveaux de gris et réduction de dimension des images de définition > 120 000 pixels → feature + equalization + filtre médian : image	1014
Prétraitement des images pour le VGG16 → feature image_VGG16	1014



Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion

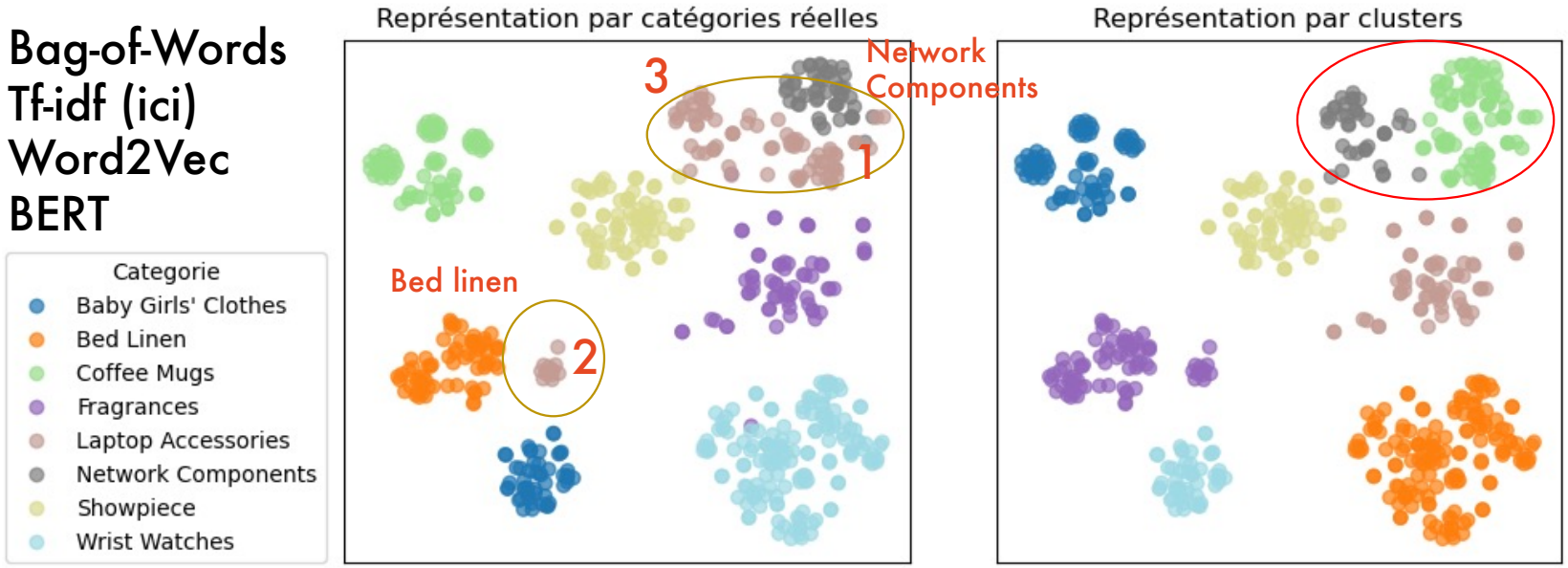
- En résumé :
  - 4 jeux de features textuelles
  - 1 feature image formatée selon la méthode utilisée
  - 2 targets de niveaux différents
    - Target_wide – niveau 1
    - Target – niveau 2
  - 3 sélections de données
    - Totalité des individus : pour target et target_wide
    - Individus des catégories de plus de 20 individus : pour target
    - Individus des catégories de plus de 48 individus : pour target
  - 1 séparation train/test pour les méthodes avec fit / transform

Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion

Méthode	Features	Séparation train/test	Kmeans + histog.+ PCA	T-sne + Kmeans	Matrice de confusion
BoW counter	txt	✓		✓	
BoW counter PCA	txt	✓		PCA + ✓	
BoW Tf-idf	txt	✓		✓	
Word-2-Vec	txt	✓		✓	
BERT	txt			✓	
USE	txt			✓	
Naive Bayes	txt	✓			✓
SIFT	img		✓	✓	
ORB	img		✓	✓	
CNN TL	img	✓			✓
Mix : SIFT+Tf-idf	img + txt			✓	
MixPCA : SIFT+Tf-idf	Img + txt		✓	✓	

Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion

Bag-of-Words
Tf-idf (ici)
Word2Vec
BERT



1 Laptop Accessories
Classe prédite 2

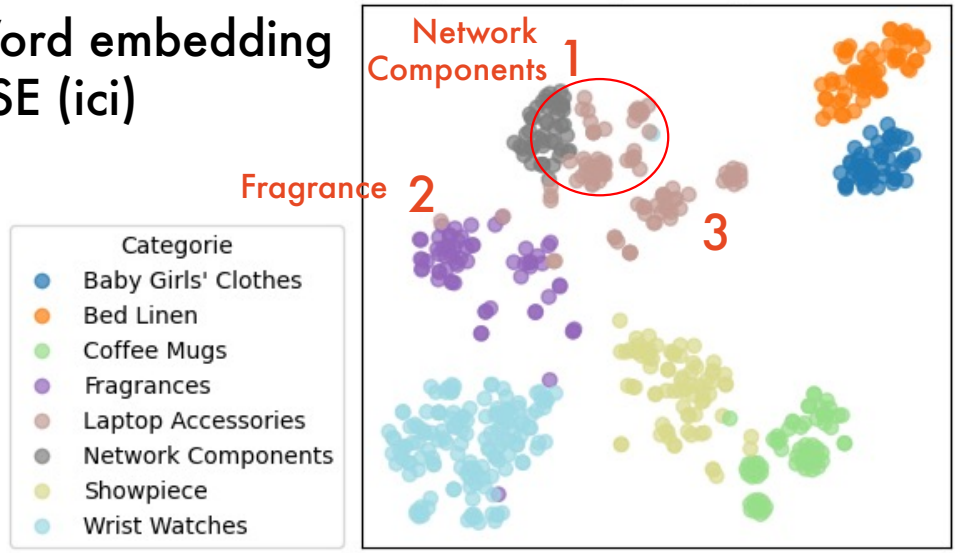
2 Laptop Accessories
Classe prédite 3

3 Laptop Accessories
Classe prédite 5

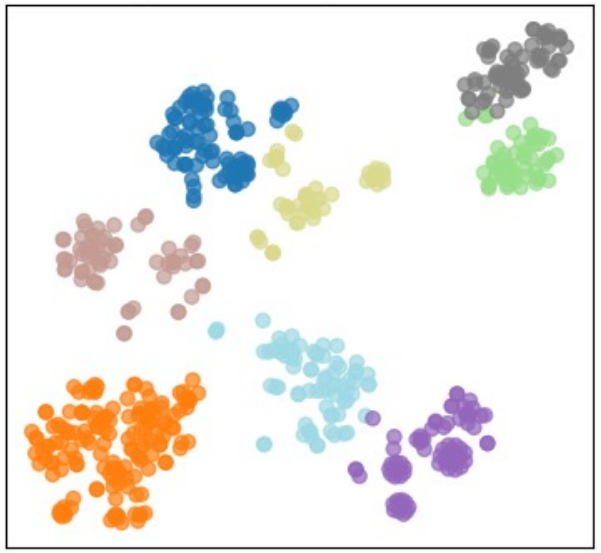
Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion

Word embedding USE (ici)

Représentation par catégories réelles

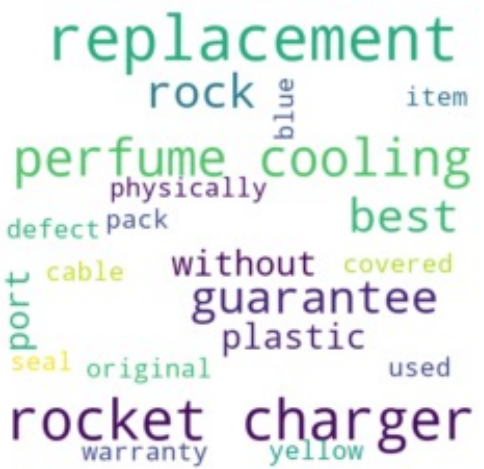


Représentation par clusters



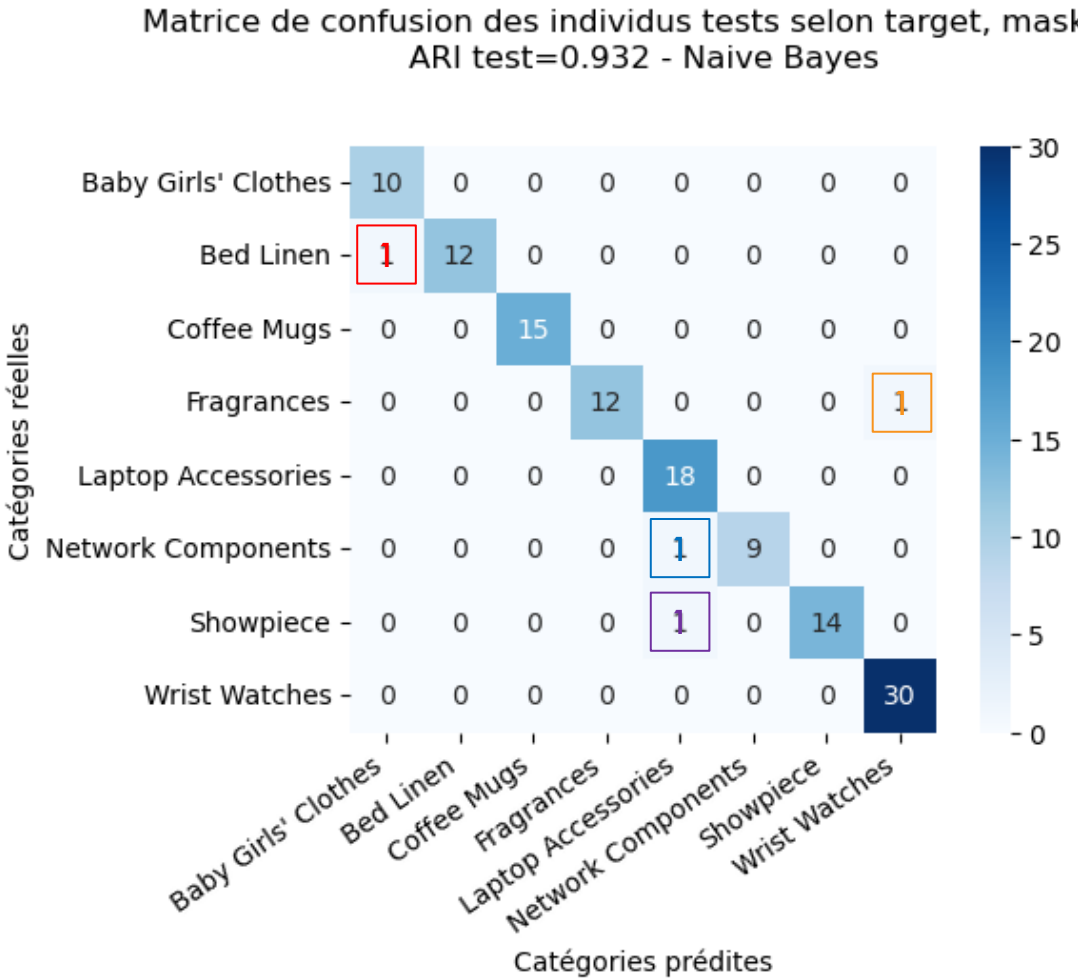
1 Laptop Accessories
Classe prédite 0

2 Laptop Accessories
Classe prédite 4



3 Laptop Accessories
Classe prédite 6

- Classification Naive Bayes (peu de données d'entrainement nécessaires)



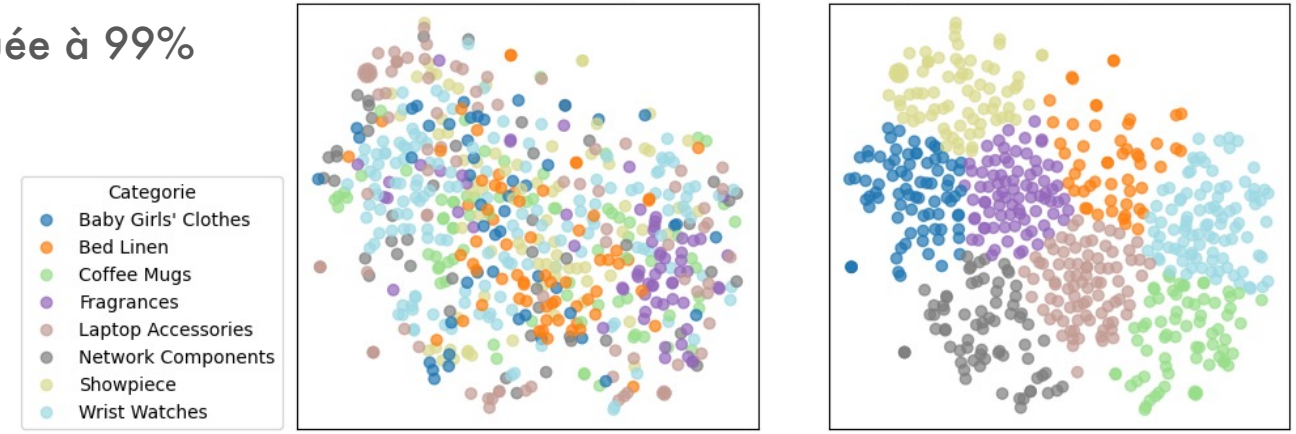
Erreurs :

1 wireless modem
1 ethnic ravishing golden showpiece statue
1 men sport online
1 printed button

- SIFT : test de différentes valeurs de variance expliquée pour réduction PCA

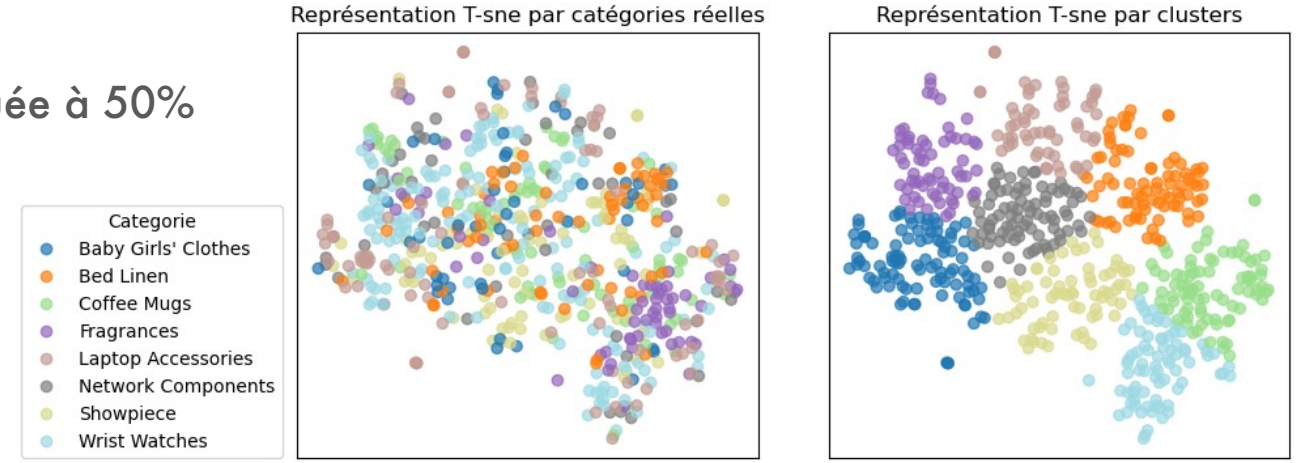
Variance expliquée à 99%

ARI = 0,0655
mask_48
target



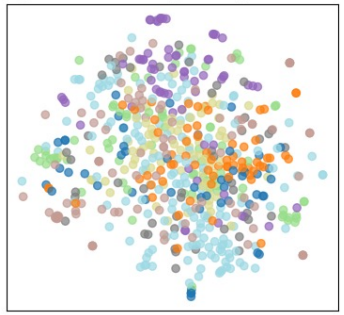
Variance expliquée à 50%

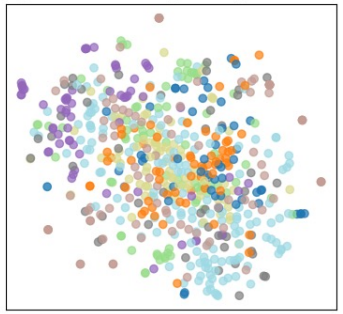
ARI = 0,0507
mask_48
target

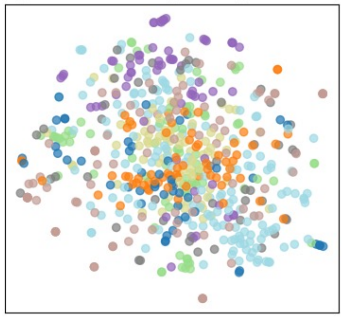


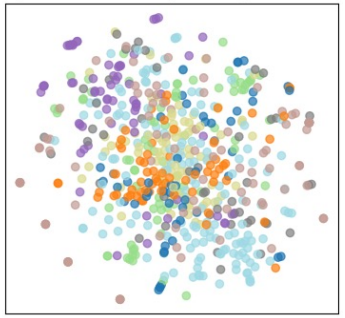
Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion

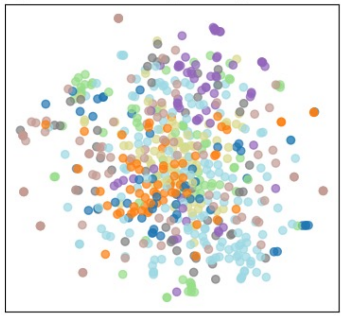
- SIFT : test de différentes valeurs de perplexité du diagramme T-Sne

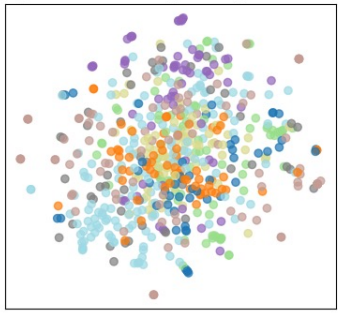

P=25

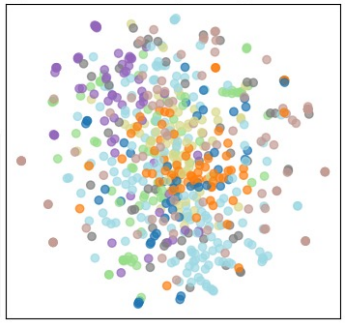

P=30

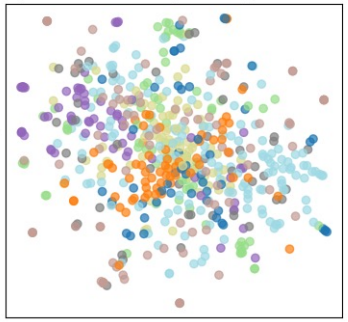

P=35

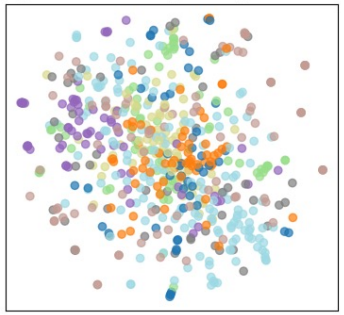

P=40


P=45


P=50

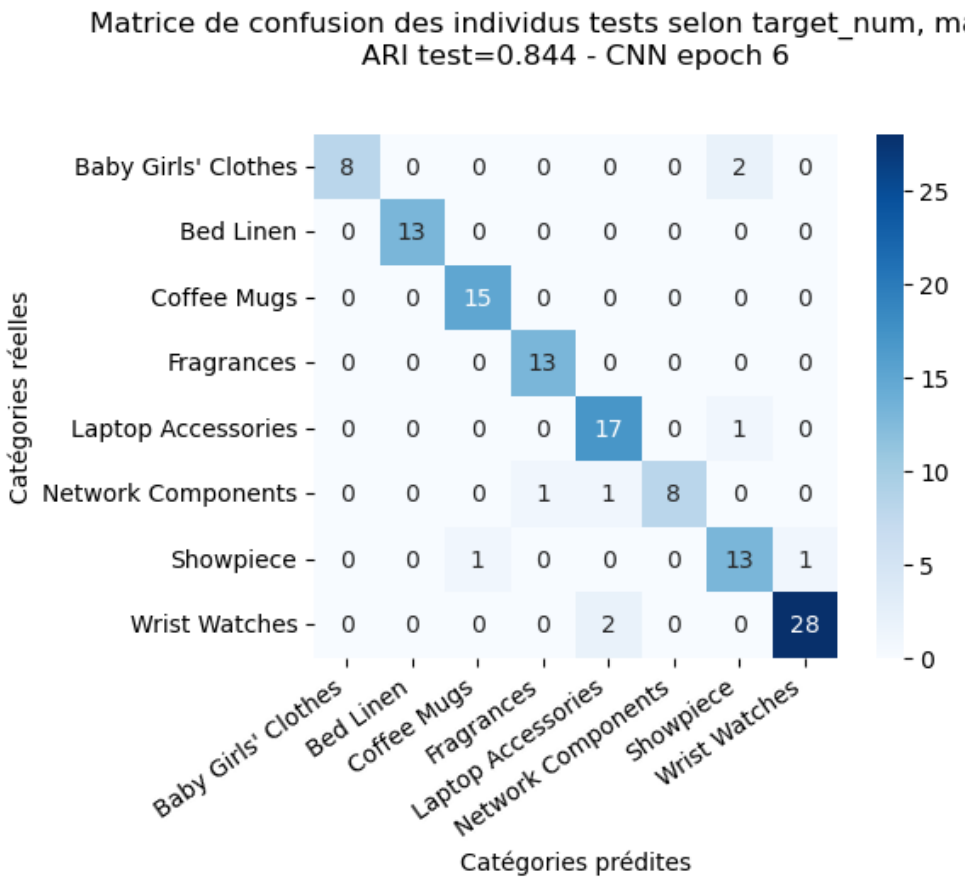

P=55


P=60


P=65

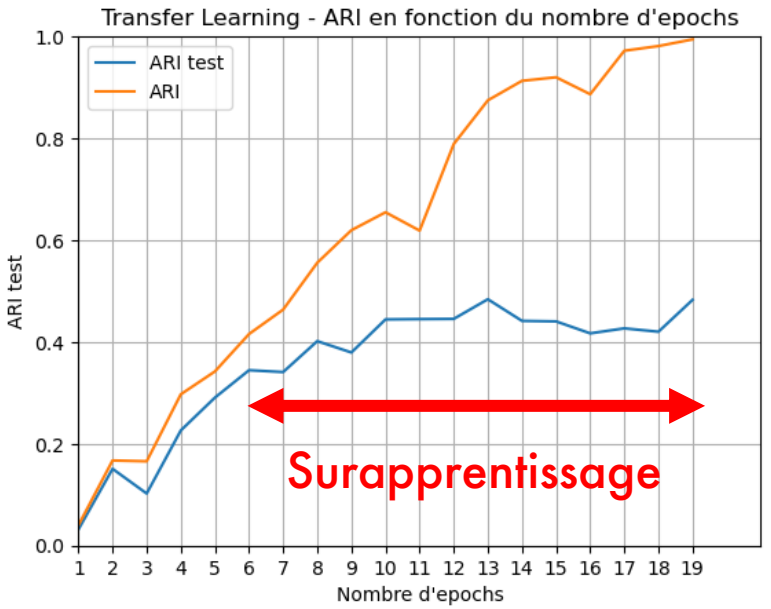
Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion

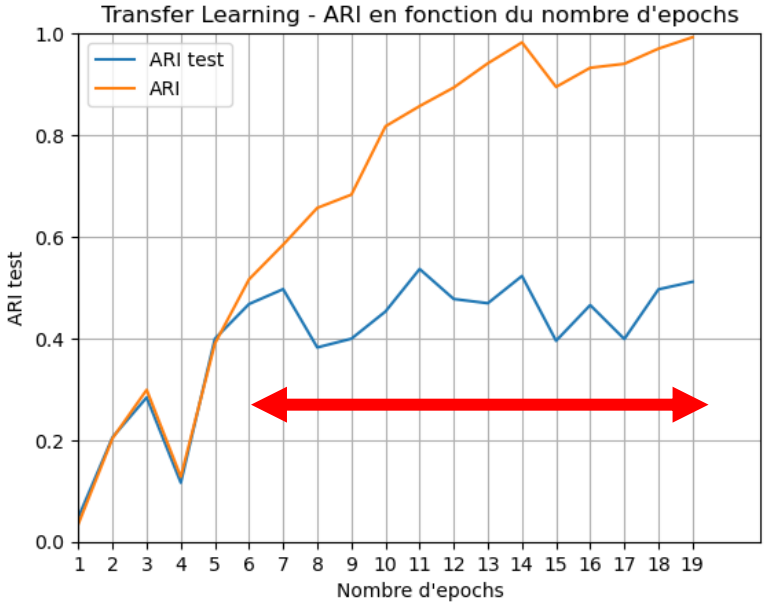
- CNN Transfer Learning – include_top = False

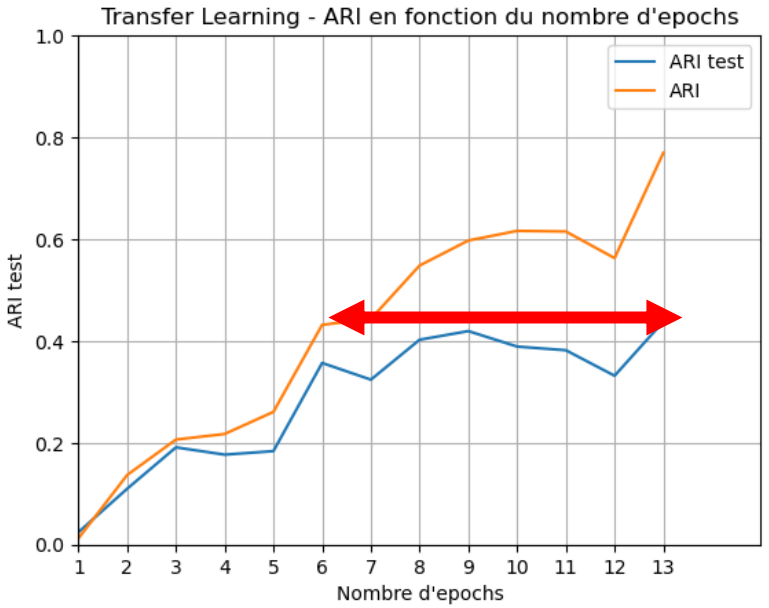


Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion

- CNN Transfer Learning – entrainement de la dernière couche – target – mask_tot
  - Besoin de davantage de données pour l'entrainement
  - Très long (15 mn/epoch avec la totalité de ce jeu de données sur 33 catégories)



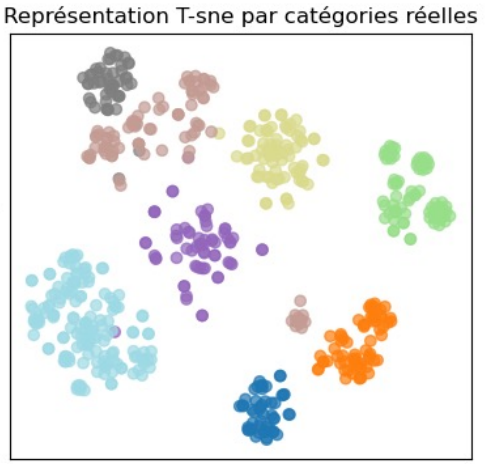


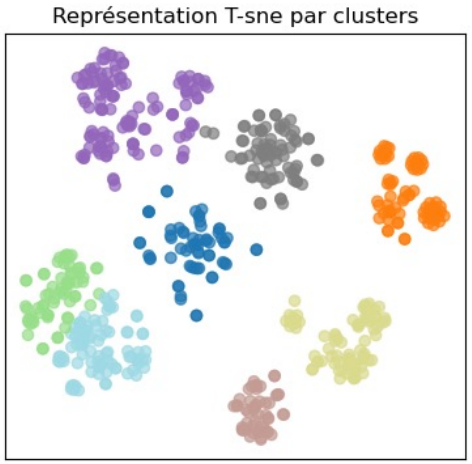


Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion

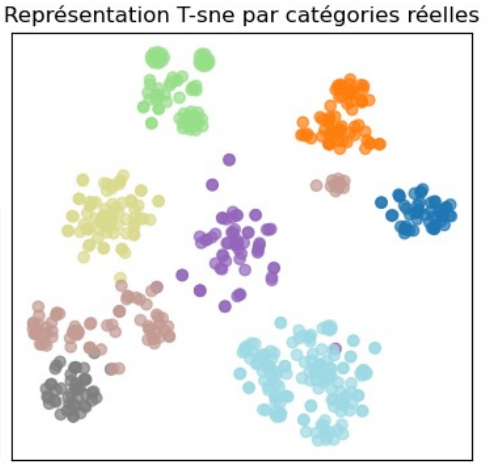
- Bag of Words + Bag of Visual Words (SIFT + Tf-idf)

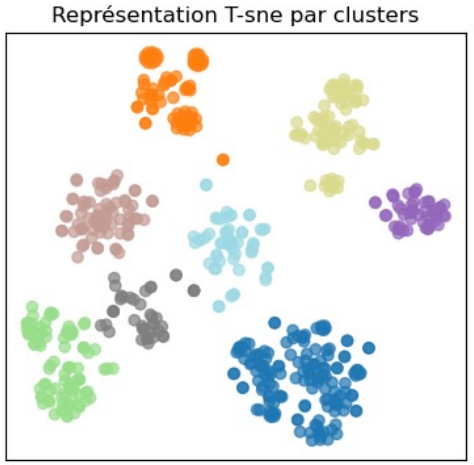
Sans réduction dimensionnelle par PCA
ARI = 0,75
mask_48
target





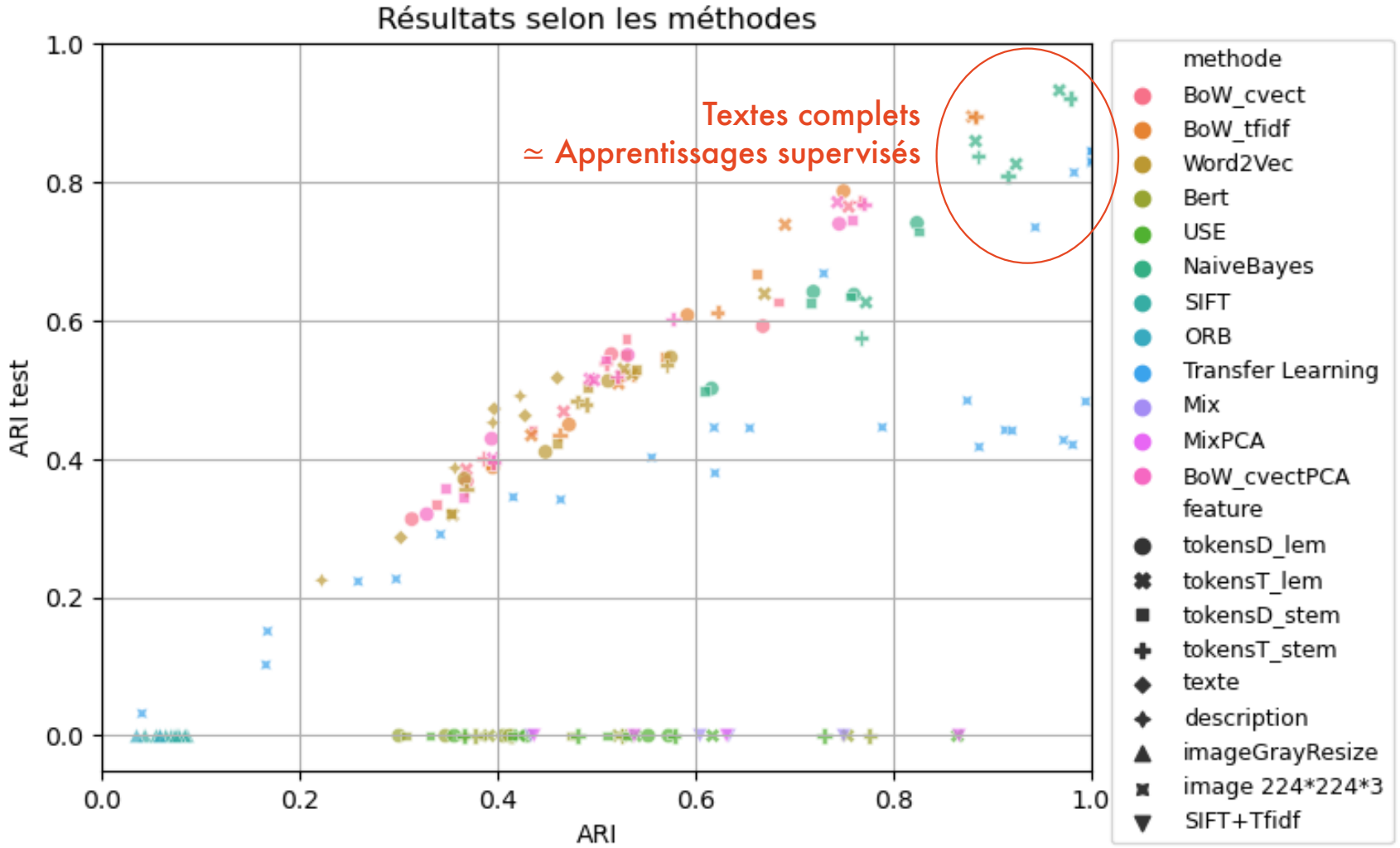
Avec réduction dimensionnelle par PCA
ARI = 0,87
mask_48
target





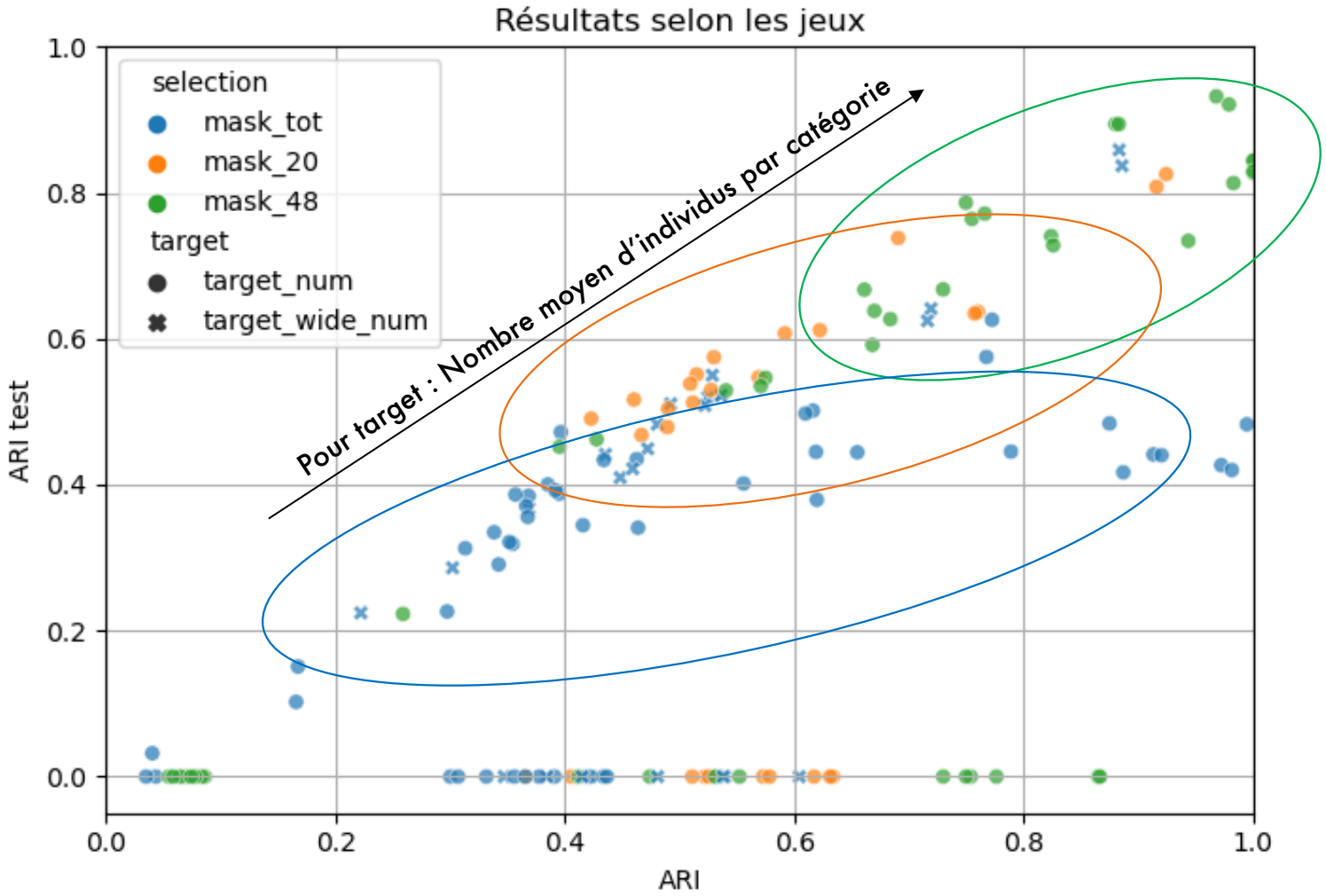
pour info : ARI SIFT = 0,09 ARI Tf-idf = 0,88

Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion



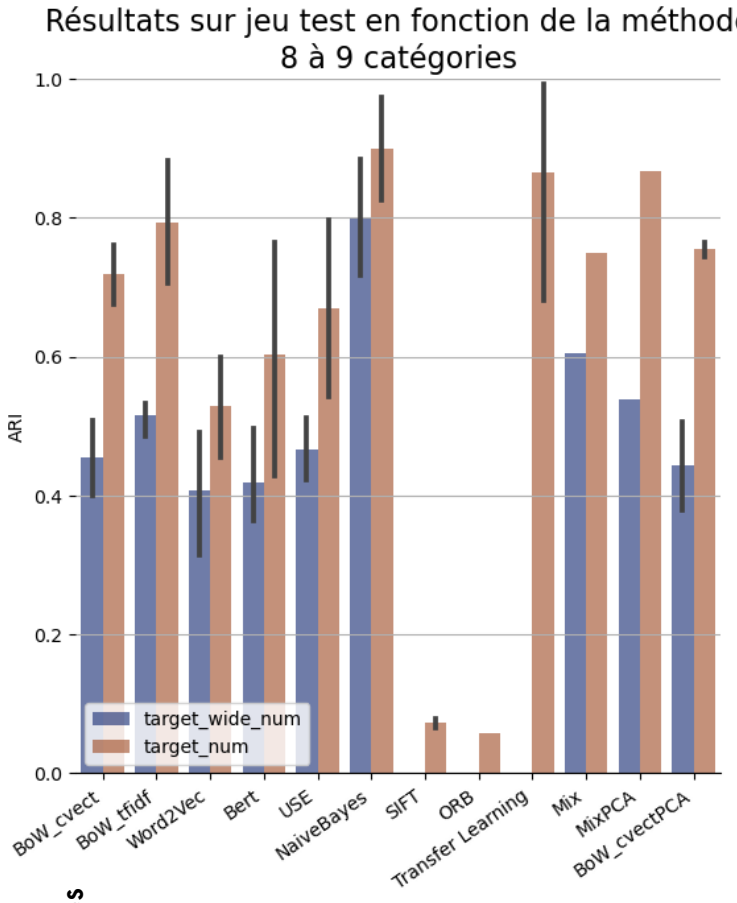
- Meilleurs résultats sur les textes complets
- Le choix stemming ou lemmatization joue peu
- Meilleurs résultats avec l'apprentissage supervisé

Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion



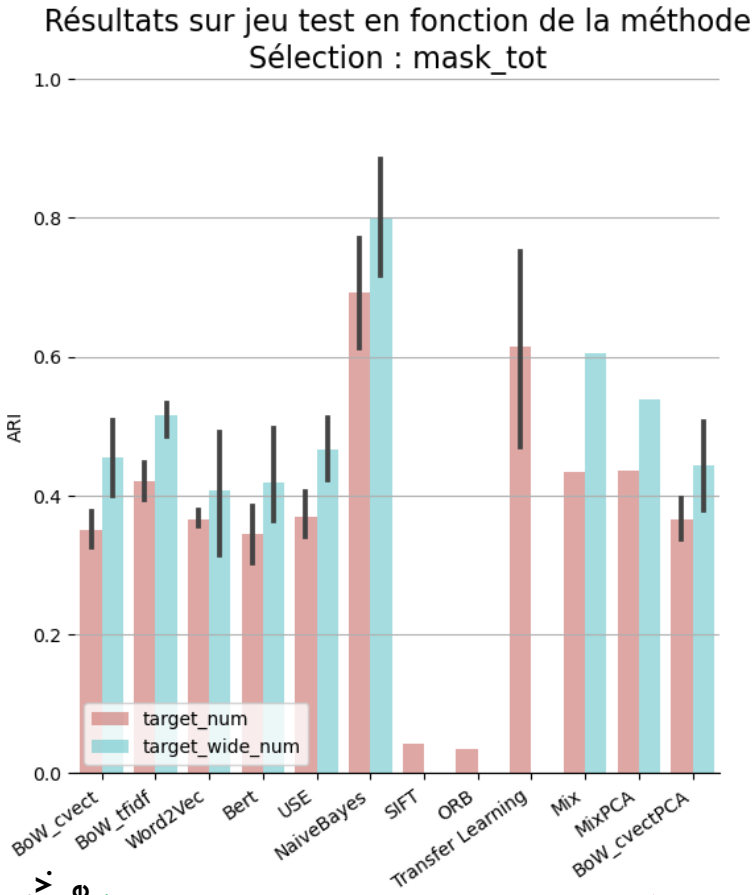
- Meilleurs résultats lorsque la sélection est plus réduite :
  - Car moins de catégories?
  - Car les catégories sont plus précises et plus distinctes?
  - Car il y a plus d'individus par catégories?

Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion



Précision des catégories

- 7 catégories pour target_wide
+ 8 catégories pour target
Nombres de catégories ± identiques



Nombre d'indiv. par catégorie

- 33 catégories pour target
+ 7 catégories pour target_wide
Nombres d'individus identiques

Nombre de catégories

Mission
Présentation du jeu de données
: la target
: les features
Nettoyage / préparation
: target
: features textuelles
: features images
: résumé
Classification
: features textuelles
: features image
: features textuelles + images
: résultats
Conclusion

Faisabilité d'un moteur de classification automatique des biens de consommation : Validée!

- Préconisations pour la mise en place de l'automatisme :
  - Redéfinition de catégories précises (une poussette et de la layette ne doivent pas se retrouver ensemble), et distinguables (« toilettage de bébé » et « peau et bain de bébé » doivent être regroupés)
  - Agrandir la base de données pour cette seconde étude (une cinquantaine d'articles par catégorie semble être suffisant pour la partie textuelle)
  - Utiliser toutes les données textuelles disponibles (description + nom + spécifications)
  - Ne pas utiliser seules les données images (un lit fait en photo dans une chambre doit-il être classé en mobilier ou en linge de lit? Impossible de le savoir avec la seule photo)
- Autres suggestions d'amélioration :
  - Réduire la taille des images de haute définition au chargement initial fait par le vendeur
  - Proposer des items pour les spécifications selon les catégories qui auront été déterminées

Place aux questions!

Et merci pour votre attention!