Data scientist
Projet 6

Classification automatique des biens de consommation

Dolores Valide
Mentor : Adrien Germond

Sommaire

- Problématique
- Présentation des données
- Processus de classification
- Résultats pour l’étude textuelle
- Résultats pour l’étude des images
- Conclusions

Problématique

L'entreprise Place de Marché souhaite lancer une marketplace e-commerce.
Sur leur site, des vendeurs proposent des articles à des acheteurs en postant une photo et une description.

Mission:

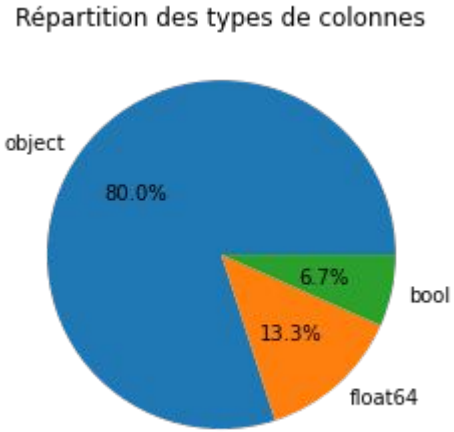
Étudier la faisabilité d'un moteur de classification d'articles, basé sur une image ou une description, pour l'automatisation de l'attribution de la catégorie de l'article.

Objectif :

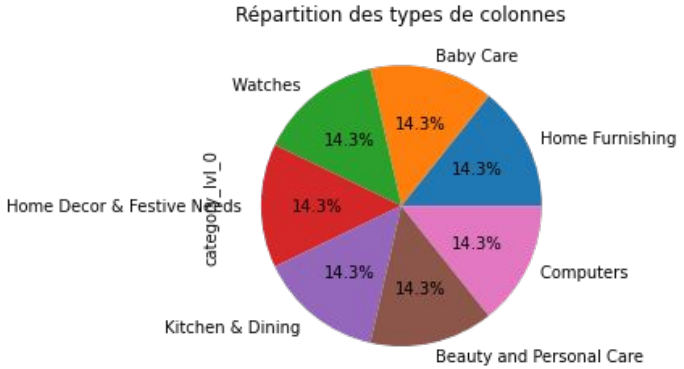
Simplifier l'expérience utilisateur, améliorer la fiabilité, et automatiser les traitements en vue d'un développement.

De quelles données disposons-nous?

Base de données de 15 variables et 1050 données avec 32% de valeurs manquantes.



Un dataset aux catégories équilibrés



Comment procédons-nous pour classer des données images et textes ?

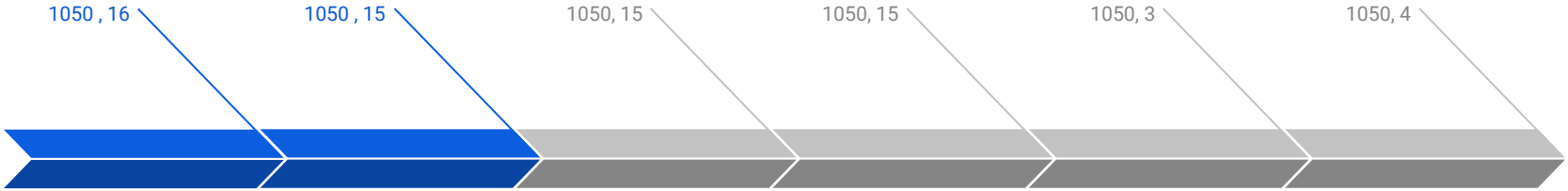
	Pré traitement	Features extraction et description *construction d'un vecteur numérique*	Réduction de dimension	Clustering	Visualisation	Evaluation
Données textuelles	Récupération des tokens et nettoyage, création d'un vocabulaire	➢ Bags of Words : **Count-vectorizer , TF-IDF** ➢ Words Embedding : **Word2Vec, BERT, USE**	ACP	Algorithme de classification Kmeans	TSNE à l'aide de l'ACP	Calcul de l'Indice de Rand Ajusté
Données graphiques (image)	Récupération des images Réduction de la taille image	➢ Bags of visual word :**SIFT** ➢ Embedding : **CNN**				

Etude de la classification par les descriptions

- Méthodologie
- Approche Bag of Words
- Approche vectorielle

Comment les données textuelles ont-elles été préparées ?

Dataset initial : 1050 lignes, 15 colonnes



Catégories de produits

Il y a plusieurs catégories et sous catégories . on ne garde que la première

Tokenisation

Retrait des caractères spéciaux

Stop word

Suppression des stop word, de la ponctuation, on garde les mots de plus de 2 lettres

Lower et lemmatisation

Passage en minuscule et récupération de la base des mots

Suppression des colonnes inutiles

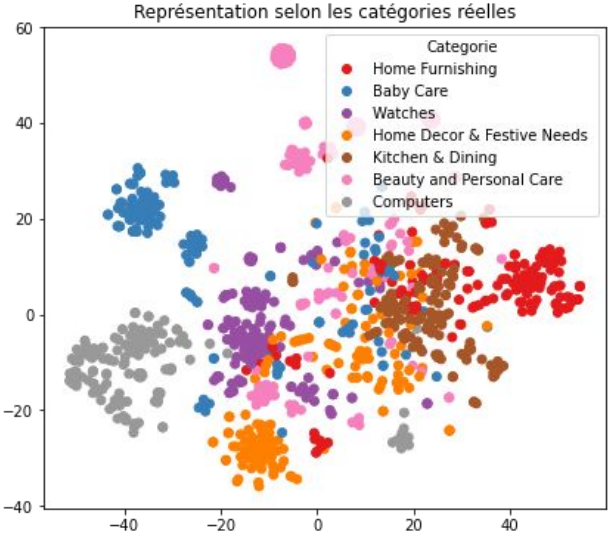
On garde les descriptions, les images et les catégories

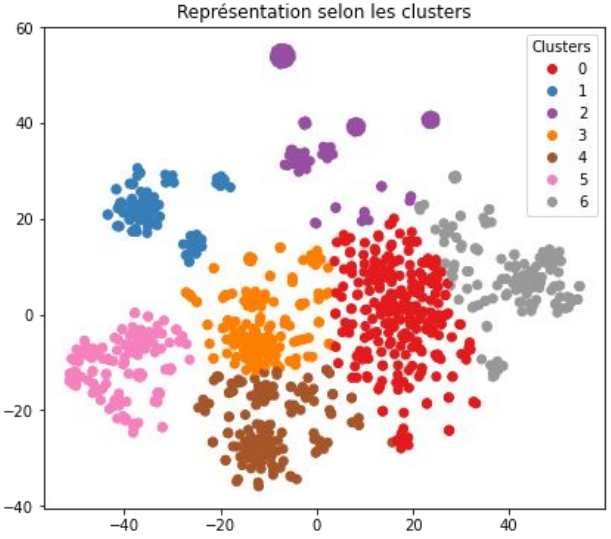
Encodage des catégories

création de labels

Classification avec Count-Vectorizer

Score ARI : 0.36

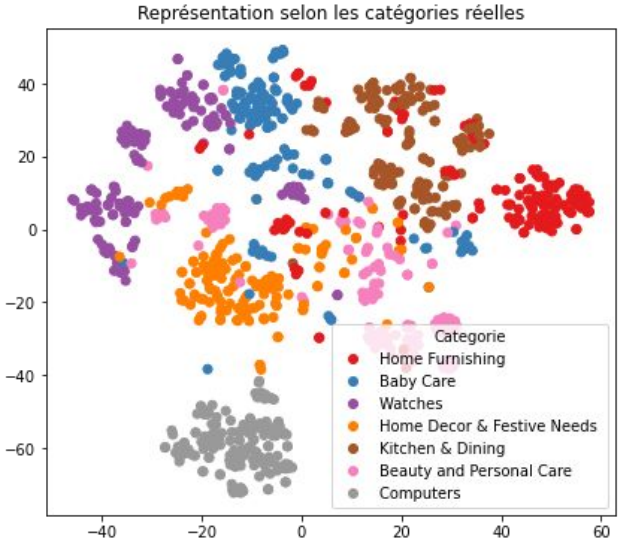


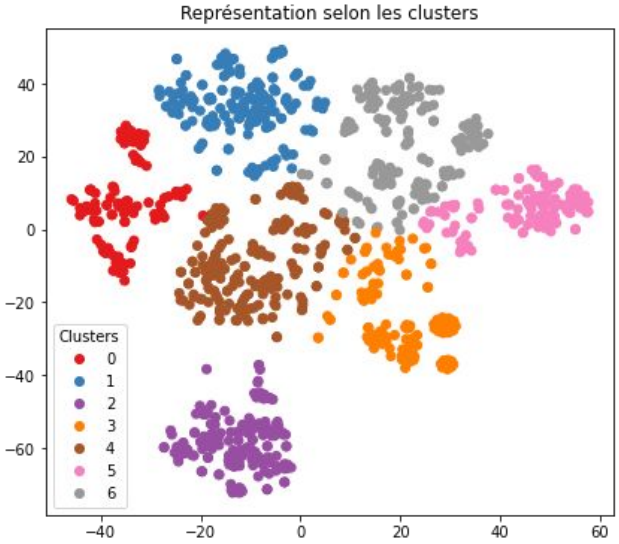


La classification comporte des erreurs et les catégories sont mal attribuées

Classification avec TF-IDF

Score ARI : 0.5351

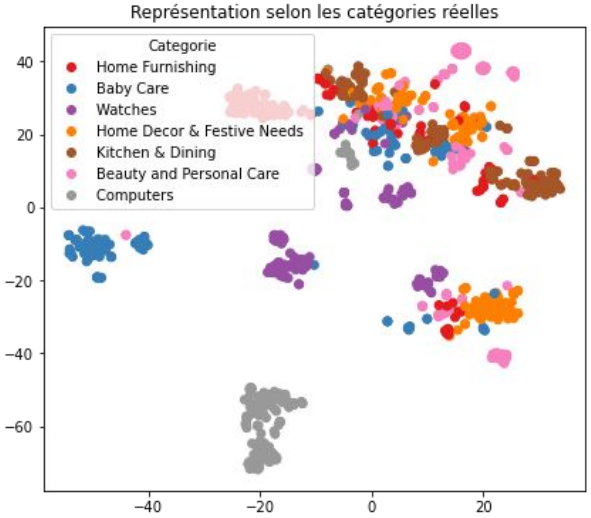


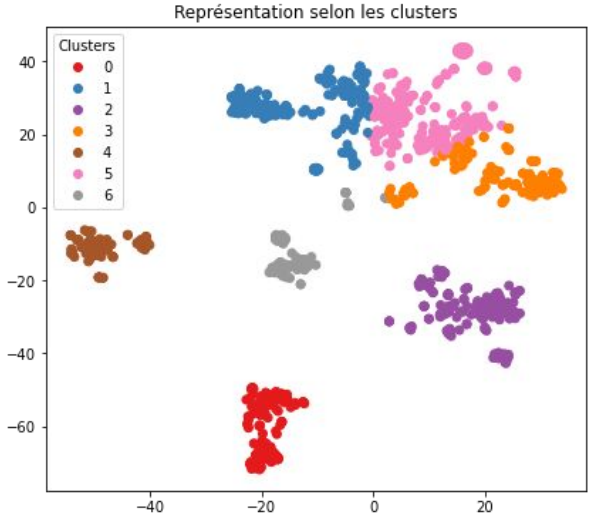


Meilleure classification qu'avec le count-vectorizer. Les catégories sont assez bien retrouvées avec l'algorithme.

Classification avec Word2Vec

Score ARI : 0.3326

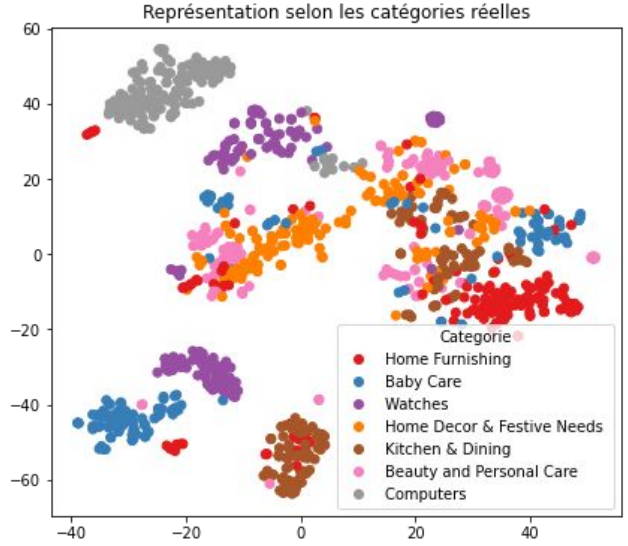


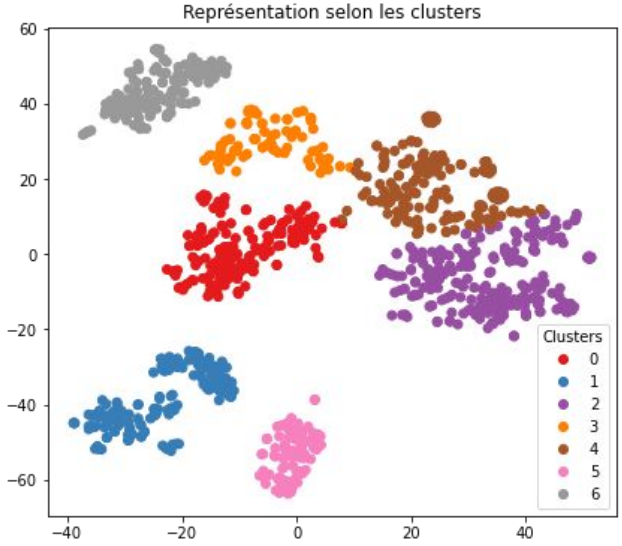


Nous redescendons niveau performance. Les catégories proches sont mal attribuées.

Classification avec BERT

Score ARI : 0.3363

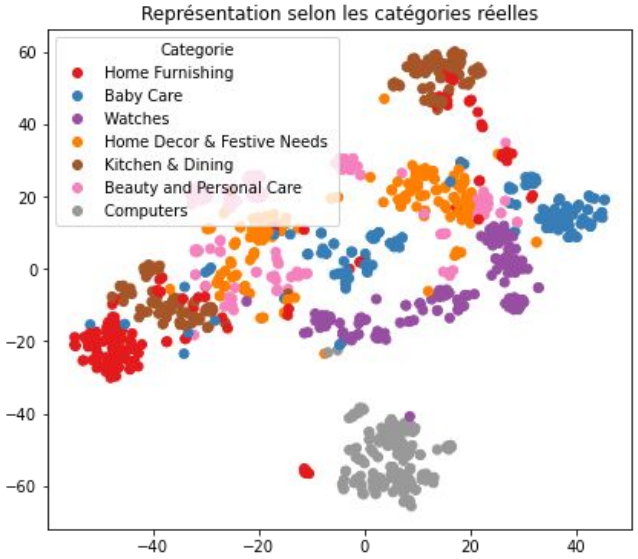


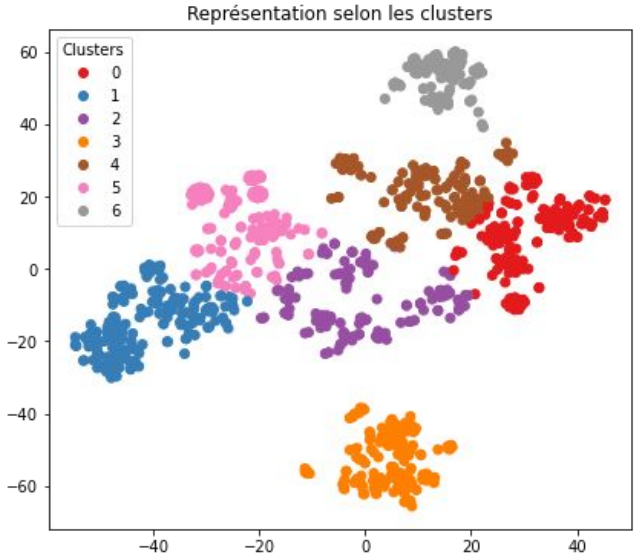


On obtient sensiblement les mêmes résultats qu'avec Word2Vec et Count-Vectorizer

Classification avec USE

Score ARI : 0.4091





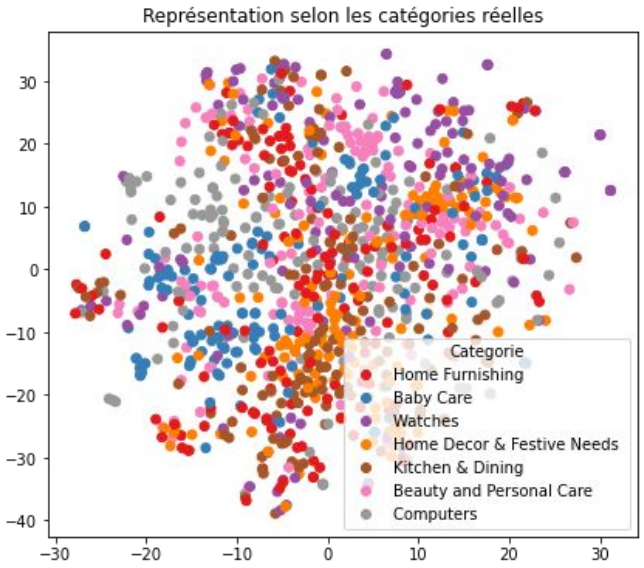
Meilleure classification après le TF-IDF. Le principe d'embedding obtient de meilleurs résultats et minimise les matrices vides.

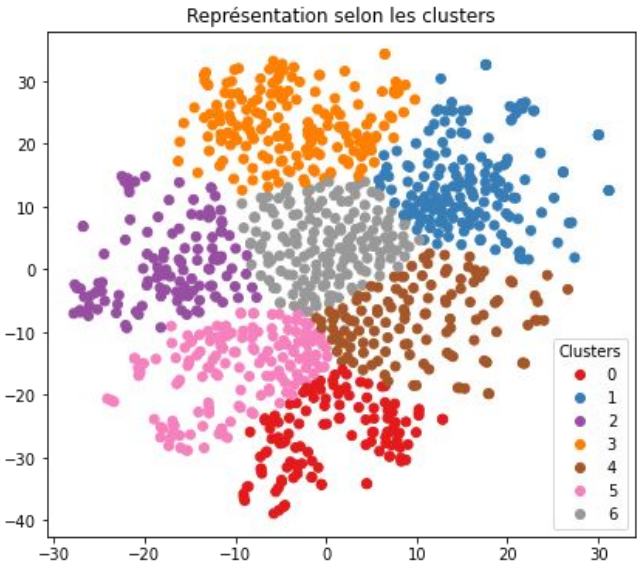
Etude de la classification à partir des images

- Approche SIFT
- Approche CNN

Classification SIFT

Score ARI : 0.0439

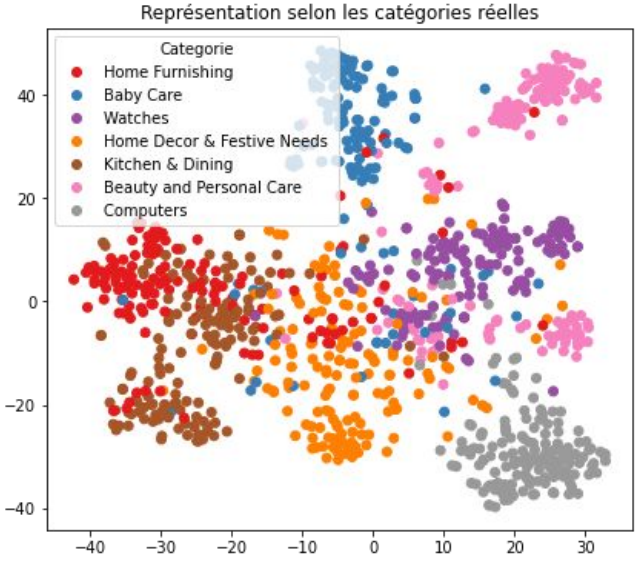


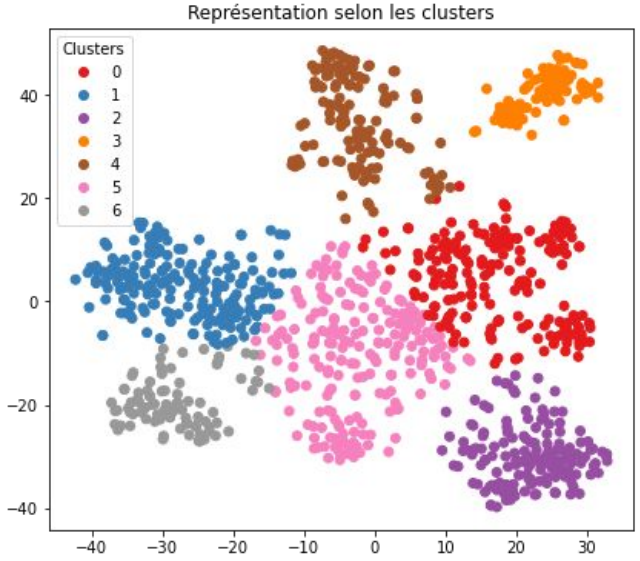


On obtient une très mauvaise classification. Presqu'aucune image n'est bien classée.

Classification Réseau de neurones : CNN

Score ARI : 0.4956





Meilleure classification pour les données graphiques(photos). Les catégories sont assez bien retrouvées avec l'algorithme.

Le score peut être amélioré en adaptant plus le modèle à nos données.

Conclusion

Le moteur de classification est réalisable.
Au vu des résultats obtenus nous pourrions partir sur une combinaison de USE + CNN car ce sont les deux algorithmes qui obtiennent les meilleurs scores ARI et qui fonctionne sur le principe d'embedding.

Pistes d'améliorations :

Ajouter des variables comme le nom du produit et la deuxième catégorie pourrait améliorer la classification .

Pré-entrainer les réseaux de neurones pourrait améliorer les résultats. Affiner les hyperparamètres également.

Merci de votre attention

Questions - Réponses