ANALYSE EXPLORATOIRE & PREPROCESSING

		Valeur	Valeur en %
Vérification	Colonne		
Valeurs manquantes	designation	0	0.000000
	description	29800	35.093504
	productid	0	0.000000
	imageid	0	0.000000
	prdtypecode	0	0.000000
Doublons	Designation	2651	3.121909
	Description	7610	13.807243
	Productid	0	0.000000
	Imageid	0	0.000000
	Prdtypecode	84889	99.968204
Unicité	Designation	82265	96.878091
	Description	47506	55.944698
	Productids	84916	100.000000
	Imageids	84916	100.000000
	Prdtypecode	27	0.031796

Tableau : Résumé de la qualité des données : valeurs manquantes, doublons et unicité des colonnes clés

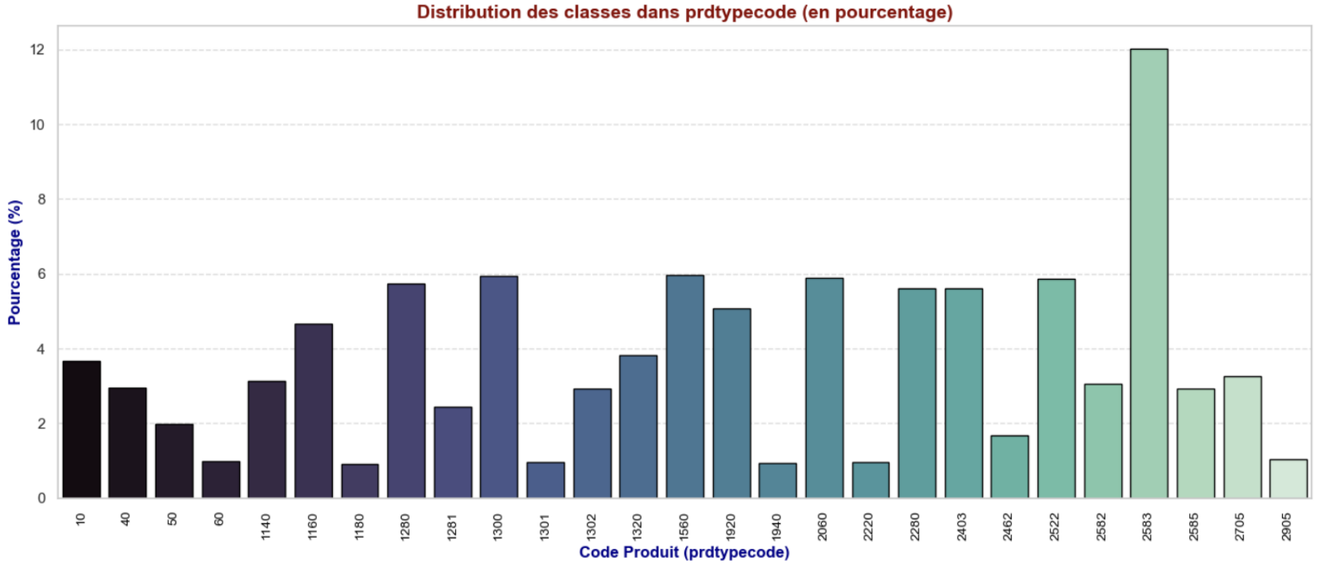


Figure 1 : Distribution catégories de produits (prdtypecode) dans le jeu de données Rakuten

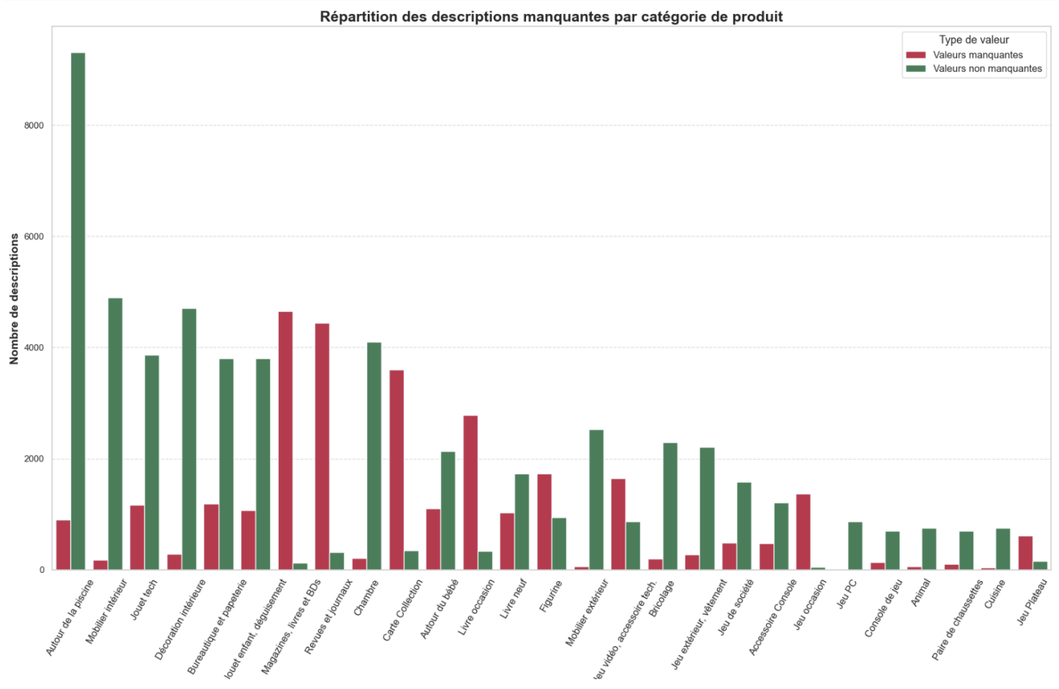


Figure 2 : Répartition des valeurs manquantes dans la colonne description selon la catégorie de produit

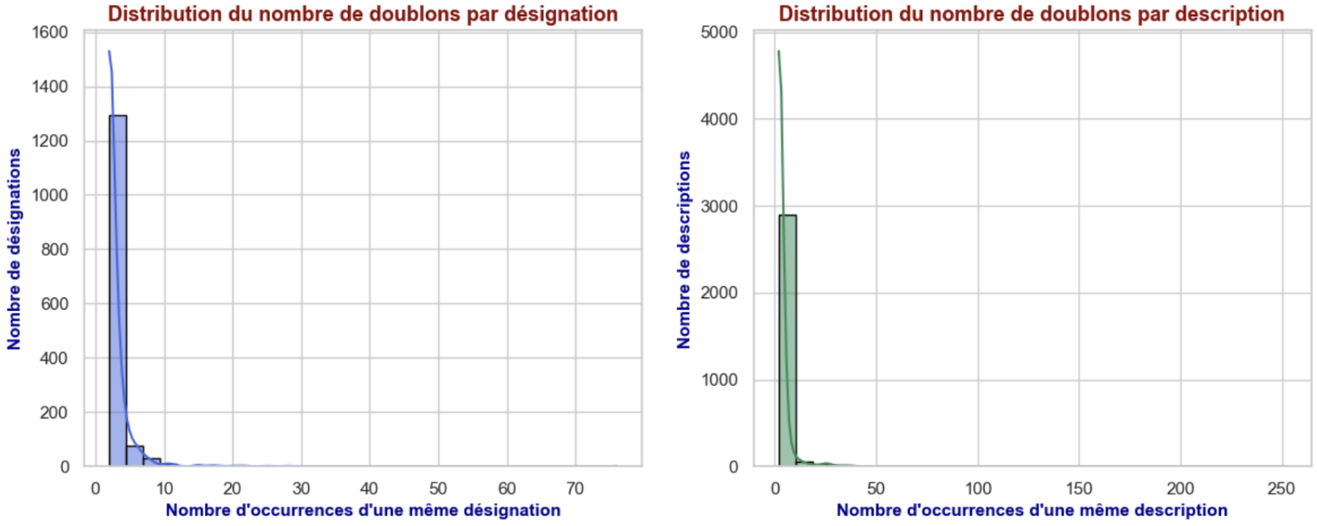


Figure 3 : Distribution du nombre de doublons dans les colonnes textuelles designation et description

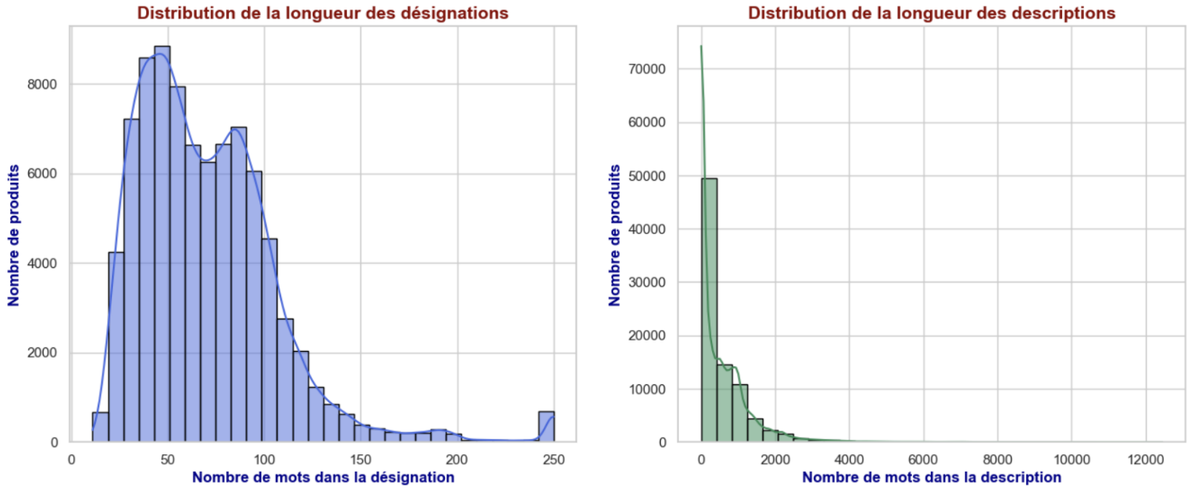


Figure 4: Distribution de la longueur des textes dans les colonnes textuelles designation et description

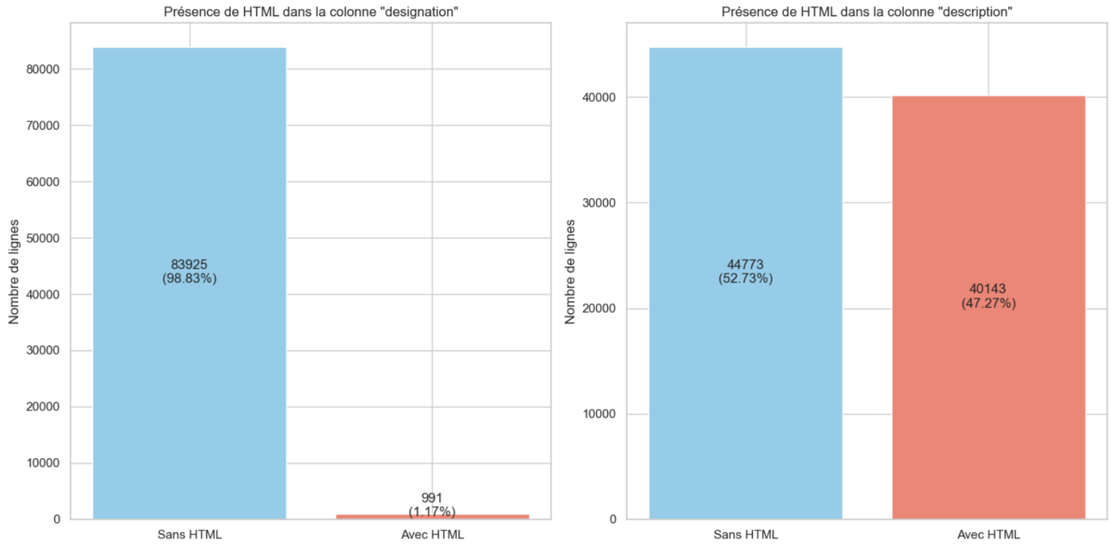


Figure 5: Analyse de la présence de balises HTML dans les colonnes designation et description

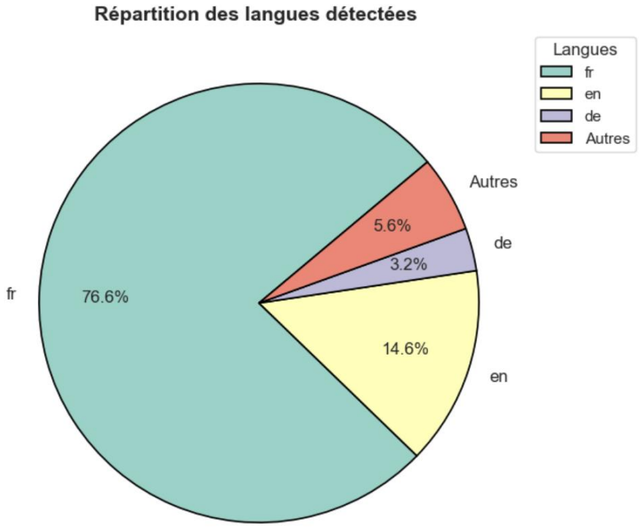


Figure 6: Analyse linguistique des données textuelles : proportion de langues détectées

Figure 7: Nuage de mots combiné : contenu textuel des produits Rakuten **sans préprocessing**

Figure 8: Nuage de mots globale et combiné des textes traduits en français **après preprocessing**

Figure 9: Nuage de mots globale et combiné après preprocessing la catégorie produit 'Console de jeu' : corpus multilingue vs français

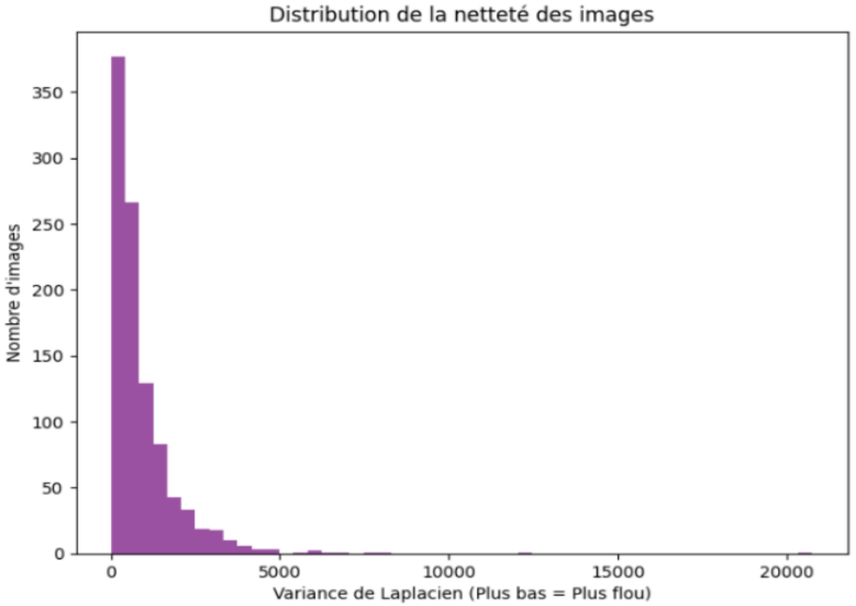


Figure 10: Distribution de la netteté des images mesurée par la variance du Laplacien

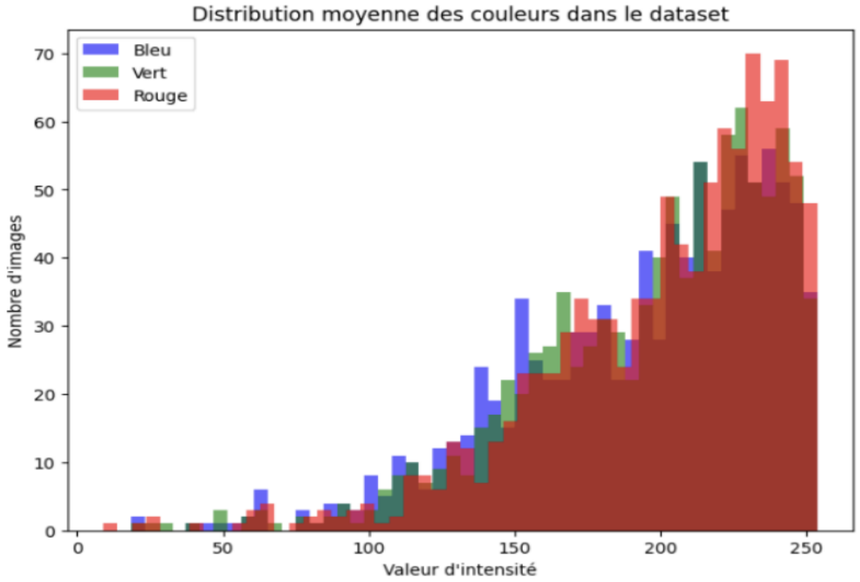
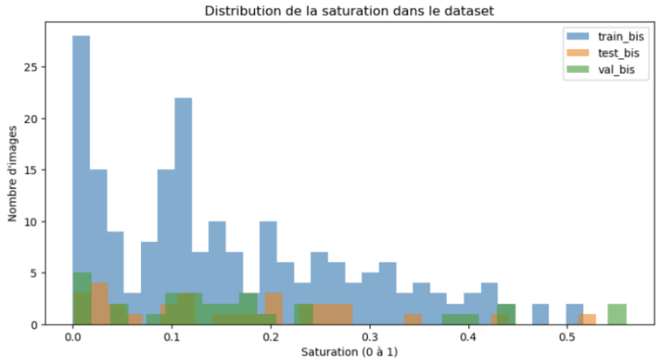
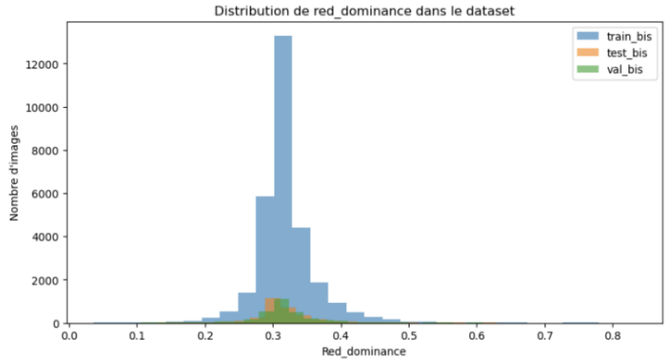


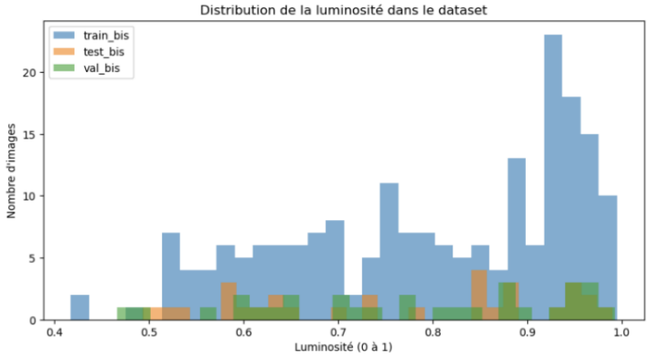
Figure 11: Distribution des intensités moyennes des canaux de couleur (RVB) dans les images du dataset



Dataset	Saturation moyenne
train_bis	0.201
test_bis	0.201
val_bis	0.224



Dataset	Red_dominance moyenne
train_bis	0.318
test_bis	0.320
val_bis	0.318



Dataset	Luminosité moyenne
train_bis	0.782
test_bis	0.766
val_bis	0.775

Figure 12: Analyse des propriétés colorimétriques des images : saturation (gauche), dominance du rouge (milieu) et luminosité (droite)

MODELISATION

Classification Report sur Validation:

	precision	recall	f1-score	support
Accessoire Console	0.74	0.83	0.78	166
Animal	0.82	0.87	0.84	82
Autour de la piscine	0.98	0.97	0.97	989
Autour du bébé	0.83	0.81	0.82	324
Bricolage	0.79	0.82	0.80	244
Bureautique et papeterie	0.93	0.93	0.93	496
Carte Collection	0.89	0.89	0.89	396
Chambre	0.91	0.92	0.91	427
Console de jeu	0.87	0.86	0.86	83
Cuisine	0.88	0.89	0.88	81
Décoration intérieure	0.82	0.77	0.80	497
Figurine	0.73	0.80	0.77	267
Jeu PC	0.92	0.99	0.96	87
Jeu Plateau	0.49	0.63	0.55	76
Jeu de société	0.60	0.56	0.57	207
Jeu extérieur, vêtement	0.85	0.79	0.82	249
Jeu occasion	0.72	0.74	0.73	141
Jeu vidéo, accessoire tech.	0.66	0.66	0.66	251
Jouet enfant, déguisement	0.76	0.66	0.71	487
Jouet tech	0.93	0.89	0.91	505
Livre neuf	0.67	0.72	0.69	276
Livre occasion	0.49	0.56	0.52	311
Magazines, livres et BDs	0.78	0.74	0.76	475
Mobilier extérieur	0.72	0.77	0.74	257
Mobilier intérieur	0.84	0.84	0.84	505
Paire de chaussettes	0.90	0.93	0.91	80
Revues et journaux	0.84	0.84	0.84	473
accuracy			0.82	8432
macro avg	0.79	0.80	0.80	8432
weighted avg	0.82	0.82	0.82	8432

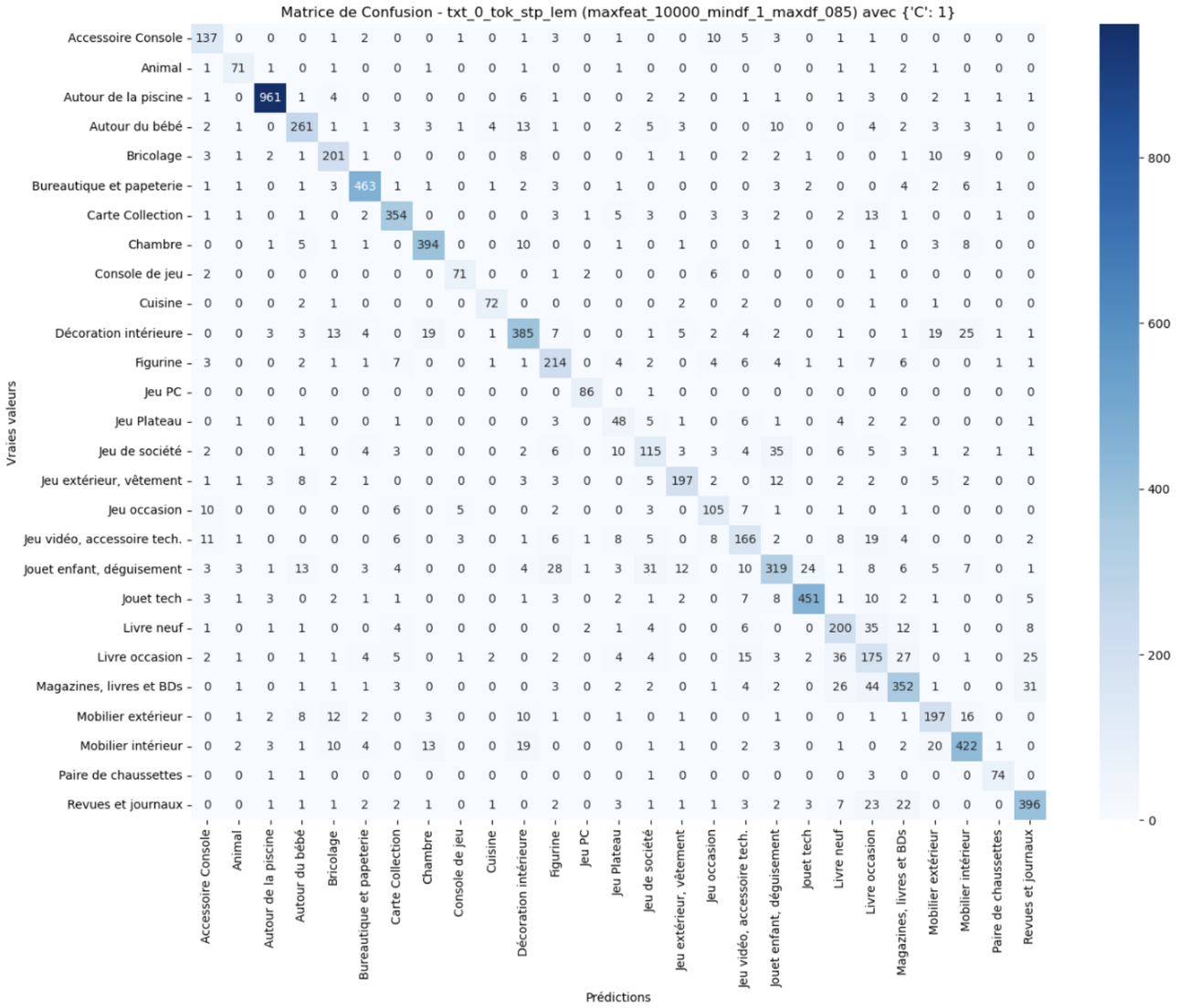


Figure 13: Performance du meilleur modèle Machine Learning : LinearSVC (TF-IDF, class_weight=balanced, C=1) sur le jeu de validation avec rapport de classification et matrice de confusion

Rapport de classification :

	precision	recall	f1-score	support
Accessoire Console	0.86	0.81	0.83	166
Animal	0.90	0.80	0.85	82
Autour de la piscine	0.97	0.97	0.97	989
Autour du bébé	0.79	0.80	0.79	324
Bricolage	0.77	0.79	0.78	244
Bureautique et papeterie	0.92	0.93	0.92	496
Carte Collection	0.90	0.87	0.89	396
Chambre	0.92	0.92	0.92	427
Console de jeu	0.93	0.84	0.89	83
Cuisine	0.88	0.80	0.84	81
Décoration intérieure	0.79	0.83	0.81	497
Figurine	0.76	0.80	0.78	267
Jeu PC	0.95	1.00	0.97	87
Jeu Plateau	0.75	0.47	0.58	76
Jeu de société	0.64	0.54	0.58	207
Jeu extérieur, vêtement	0.82	0.76	0.79	249
Jeu occasion	0.76	0.80	0.78	141
Jeu vidéo, accessoire tech.	0.69	0.63	0.66	251
Jouet enfant, déguisement	0.74	0.70	0.72	487
Jouet tech	0.95	0.91	0.93	505
Livre neuf	0.67	0.68	0.68	276
Livre occasion	0.38	0.64	0.48	311
Magazines, livres et BDs	0.79	0.74	0.77	475
Mobilier extérieur	0.76	0.68	0.72	257
Mobilier intérieur	0.83	0.85	0.84	505
Paire de chaussettes	0.93	0.86	0.90	80
Revues et journaux	0.87	0.83	0.85	473
accuracy			0.81	8432
macro avg	0.81	0.79	0.80	8432
weighted avg	0.82	0.81	0.82	8432

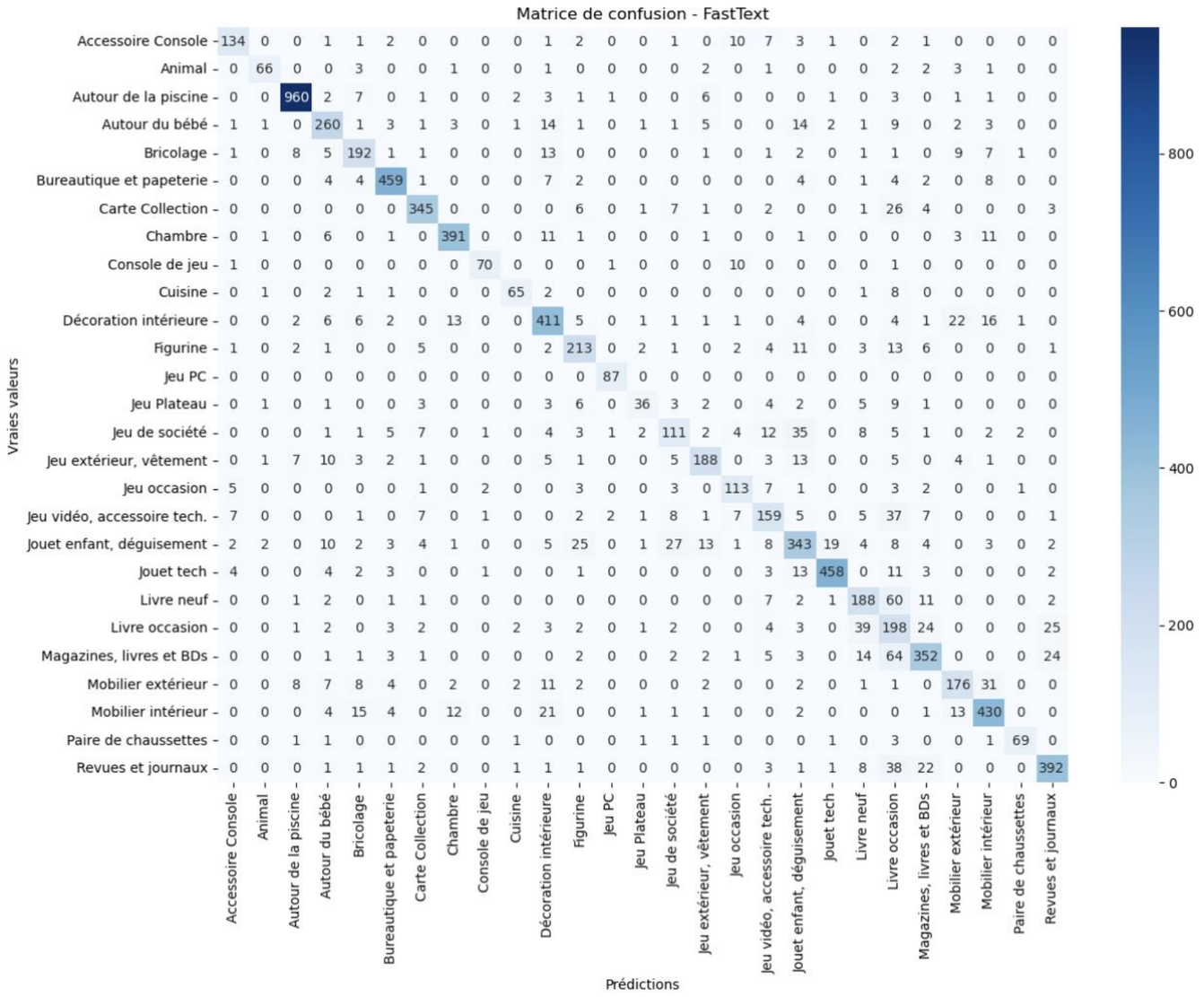
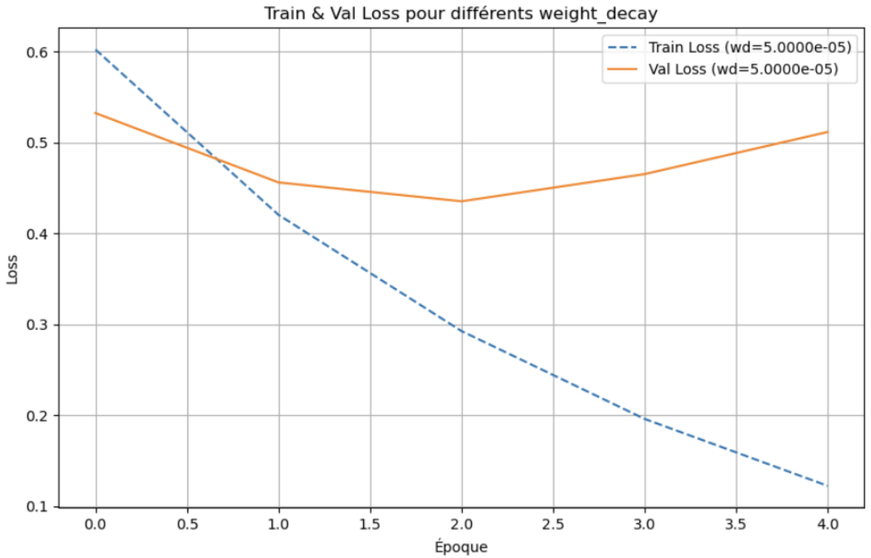


Figure 14: Performance du meilleur modèle FastText supervised (lr=0.5, dim=100, epoch=100) sur le jeu de validation avec rapport de classification et matrice de confusion



Évaluation sur le jeu de validation (X_val) :

	precision	recall	f1-score	support
Livre occasion	0.69	0.77	0.73	311
Jeu vidéo, accessoire tech.	0.77	0.81	0.79	251
Accessoire Console	0.82	0.89	0.85	167
Console de jeu	0.97	0.80	0.87	83
Figurine	0.80	0.82	0.81	267
Carte Collection	0.96	0.93	0.94	396
Jeu Plateau	0.63	0.63	0.63	76
Jouet enfant, déguisement	0.79	0.81	0.80	486
Jeu de société	0.77	0.66	0.71	207
Jouet tech	0.97	0.96	0.97	505
Paire de chaussettes	0.93	0.89	0.91	80
Jeu extérieur, vêtement	0.90	0.77	0.83	249
Autour du bébé	0.84	0.85	0.84	324
Mobilier intérieur	0.93	0.89	0.91	505
Chambre	0.92	0.95	0.93	428
Cuisine	0.91	0.96	0.93	81
Décoration intérieure	0.87	0.90	0.88	497
Animal	0.95	0.93	0.94	82
Revues et journaux	0.89	0.94	0.92	476
Magazines, livres et BDs	0.85	0.87	0.86	477
Jeu occasion	0.80	0.83	0.82	141
Bureautique et papeterie	0.96	0.95	0.95	497
Mobilier extérieur	0.85	0.83	0.84	257
Autour de la piscine	0.98	0.98	0.98	1012
Bricolage	0.93	0.92	0.93	249
Livre neuf	0.86	0.77	0.81	276
Jeu PC	1.00	1.00	1.00	87
accuracy			0.89	8467
macro avg	0.87	0.86	0.87	8467
weighted avg	0.89	0.89	0.89	8467

F1-score (weighted) : 0.8852

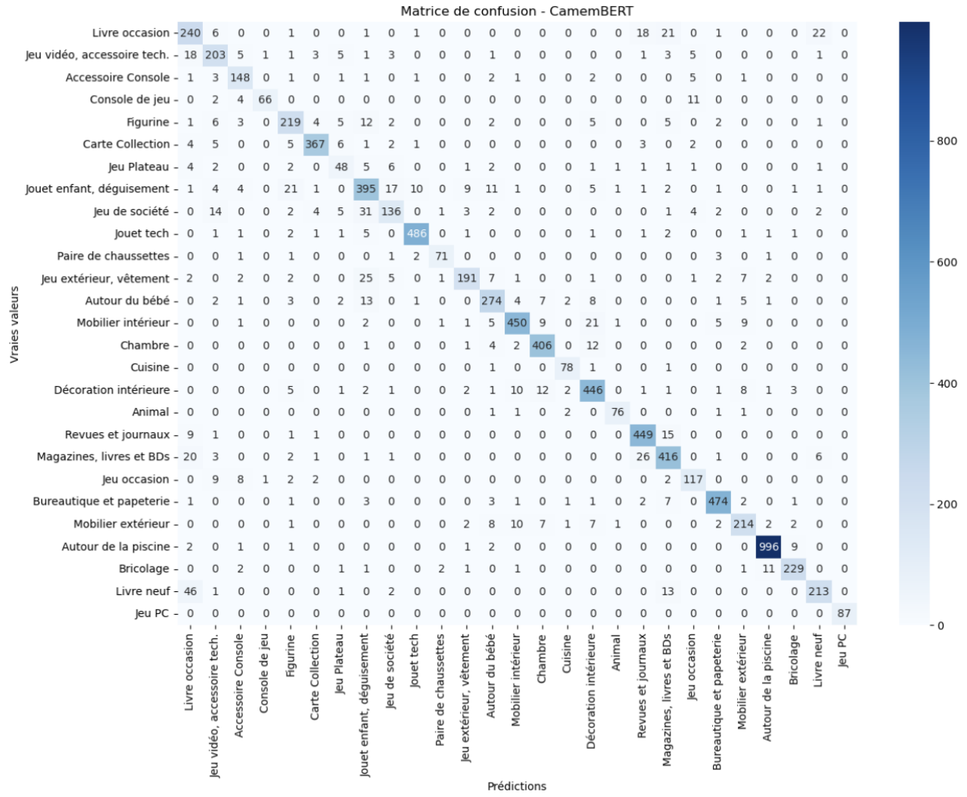


Figure 15: Performance du meilleur modèle CamemBERT (lr = 5e-5, weight_decay = 0.001, batch_size=16) sur le jeu de validation : perte (gauche), rapport de classification (milieu), matrice de confusion (droite)

Classification Report (Validation):

	precision	recall	f1-score	support
Accessoire Console	0.33	0.29	0.31	168
Animal	0.47	0.11	0.18	82
Autour de la piscine	0.84	0.83	0.84	1021
Autour du bébé	0.40	0.48	0.44	324
Bricolage	0.47	0.36	0.41	250
Bureautique et papeterie	0.71	0.62	0.66	499
Carte Collection	0.86	0.90	0.88	396
Chambre	0.67	0.78	0.72	431
Console de jeu	0.56	0.76	0.64	83
Cuisine	0.40	0.78	0.53	81
Décoration intérieure	0.46	0.47	0.46	499
Figurine	0.72	0.50	0.59	267
Jeu PC	0.77	0.78	0.78	87
Jeu Plateau	0.54	0.17	0.26	76
Jeu de société	0.46	0.16	0.24	207
Jeu extérieur, vêtement	0.41	0.34	0.37	249
Jeu occasion	0.76	0.39	0.52	142
Jeu vidéo, accessoire tech.	0.62	0.67	0.65	251
Jouet enfant, déguisement	0.39	0.53	0.45	487
Jouet tech	0.60	0.72	0.65	505
Livre neuf	0.59	0.92	0.72	276
Livre occasion	0.59	0.45	0.51	311
Magazines, livres et BDs	0.76	0.63	0.69	477
Mobilier extérieur	0.45	0.43	0.44	259
Mobilier intérieur	0.55	0.51	0.53	507
Paire de chaussettes	0.41	0.49	0.45	81
Revues et journaux	0.72	0.81	0.76	476
accuracy			0.61	8492
macro avg	0.58	0.55	0.54	8492
weighted avg	0.62	0.61	0.60	8492

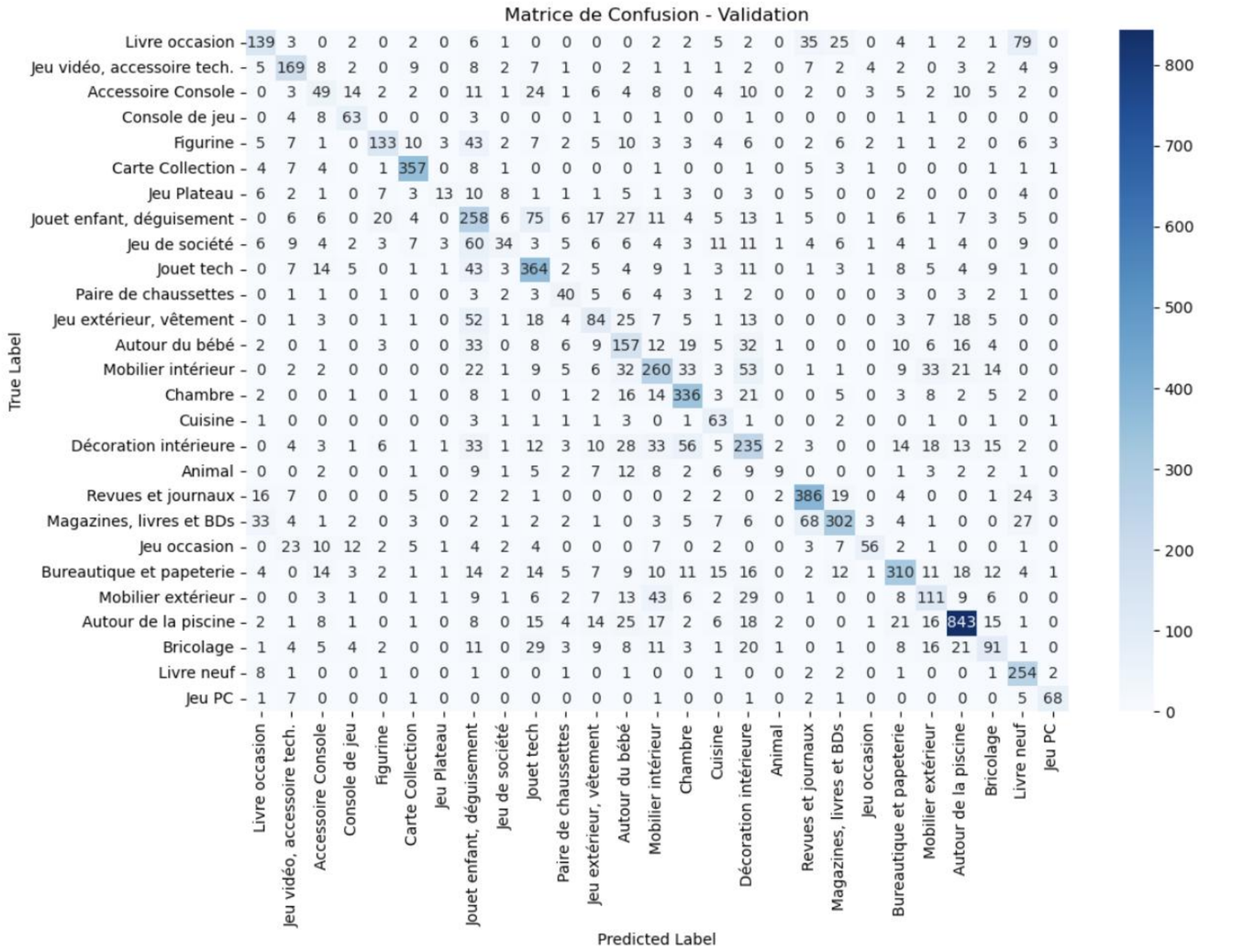


Figure 16: Performance du meilleur modèle RESNET50 (avec data-augmentation) sur le jeu de validation : perte (gauche), rapport de classification (milieu), matrice de confusion (droite)

Rapport de classification - VALIDATION (Stacking avec XGBoost) :

	precision	recall	f1-score	support
Livre occasion	0.80	0.80	0.80	311
Jeu vidéo, accessoire tech.	0.83	0.88	0.85	251
Accessoire Console	0.88	0.85	0.86	167
Console de jeu	0.95	0.89	0.92	83
Figurine	0.84	0.86	0.85	267
Carte Collection	0.97	0.96	0.97	396
Jeu Plateau	0.69	0.63	0.66	76
Jouet enfant, déguisement	0.82	0.81	0.81	486
Jeu de société	0.75	0.74	0.75	207
Jouet tech	0.97	0.96	0.97	505
Paire de chaussettes	0.95	0.89	0.92	80
Jeu extérieur, vêtement	0.87	0.84	0.85	249
Autour du bébé	0.84	0.88	0.86	324
Mobilier intérieur	0.92	0.90	0.91	505
Chambre	0.93	0.93	0.93	428
Cuisine	0.93	0.94	0.93	81
Décoration intérieure	0.87	0.88	0.88	497
Animal	0.90	0.91	0.91	82
Revues et journaux	0.92	0.93	0.92	476
Magazines, livres et BDs	0.86	0.89	0.87	477
Jeu occasion	0.83	0.86	0.85	141
Bureautique et papeterie	0.96	0.96	0.96	497
Mobilier extérieur	0.85	0.84	0.85	257
Autour de la piscine	0.99	0.99	0.99	1012
Bricolage	0.94	0.93	0.94	249
Livre neuf	0.92	0.89	0.90	276
Jeu PC	1.00	1.00	1.00	87
accuracy			0.90	8467
macro avg	0.89	0.88	0.89	8467
weighted avg	0.90	0.90	0.90	8467

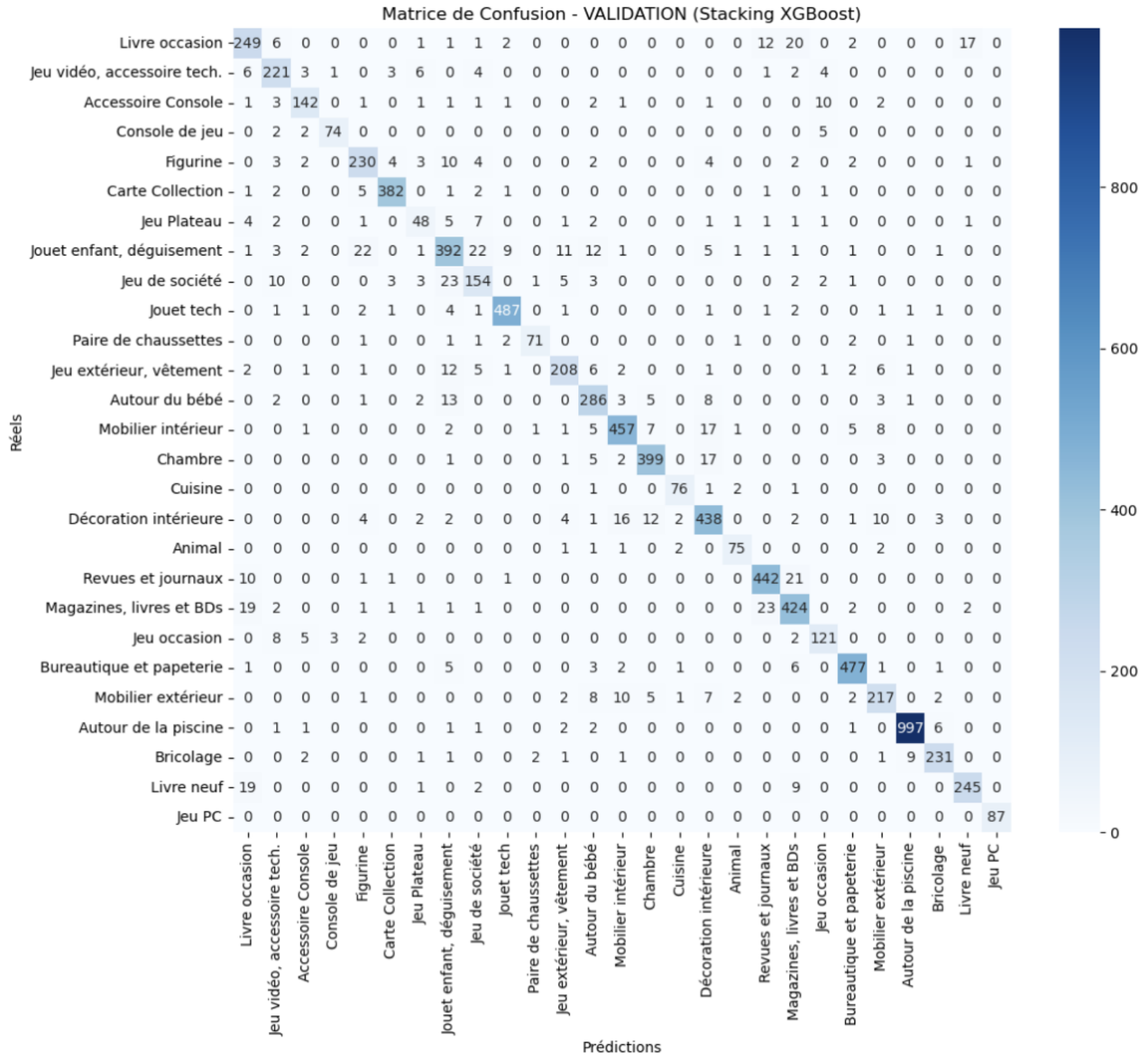


Figure 17: Performance du meilleur modèle multimodale (Stacking Classifier, Level-0: SVC sur probabilités CamemBERT, RandomForest sur probabilités RESNET50), Level-1: XGBoost) sur le jeu de validation : perte (gauche), rapport de classification (milieu), matrice de confusion (droite)