DataScientest

PROJET RAKUTEN CHALLENGE

Xiaosong
Gaby Duchiron
Céline T.
Marta González Rincón

Table des matières

I - Introduction générale

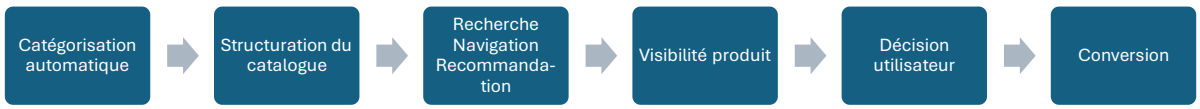
1.1 Contexte et enjeux de la classification e-commerce

Dans un contexte de e-commerce à grande échelle, la capacité à catégoriser automatiquement les produits constitue un enjeu stratégique majeur. Les plateformes de vente en ligne reposent sur une structuration fine de leurs catalogues pour assurer la pertinence de la recherche, la qualité des recommandations, la navigation utilisateur et, in fine, la conversion commerciale.

La catégorisation conditionne directement la visibilité des produits dans les moteurs de recherche internes, leur exposition dans les systèmes de recommandation et la cohérence des parcours utilisateurs. Une erreur de classification peut entraîner un produit mal positionné, difficilement trouvable, voire totalement invisible pour l'utilisateur final.

À l'échelle industrielle, où des milliers de nouveaux produits peuvent être intégrés quotidiennement, l'automatisation de cette tâche devient indispensable. Elle permet de réduire les coûts opérationnels liés à l'annotation manuelle, d'accélérer la mise en ligne des catalogues et de garantir une homogénéité globale de la taxonomie produit.

Figure 1:Impact business de la catégorisation produit sur le parcours d'achat



1.2 Problématique du challenge Rakuten

Les descriptions textuelles sont souvent bruitées, incomplètes ou rédigées de manière hétérogène, mêlant langage marketing, termes techniques, anglicismes et parfois en plusieurs langues. Les images, quant à elles, présentent une forte variabilité en termes de cadrage, de qualité, de fond et de contenu visuel.

La tâche consiste à attribuer chaque produit à l'une des 27 catégories du catalogue Rakuten, couvrant des univers très variés tels que les livres, les jeux, l'équipement de la maison, les loisirs ou encore les produits techniques. Cette diversité rend la séparation entre certaines classes particulièrement délicate.

Dans ce contexte, la difficulté centrale de ce problème réside dans la nature multimodale de l'information disponible. Selon la catégorie, le signal discriminant peut provenir majoritairement du texte, de l'image ou de leur interaction.

L'enjeu du projet dépasse ainsi la simple maximisation d'un score de performance. Il s'agit de comprendre les forces et les limites de chaque modalité, d'analyser leurs erreurs respectives, et de concevoir des stratégies de fusion capables d'exploiter leur complémentarité.

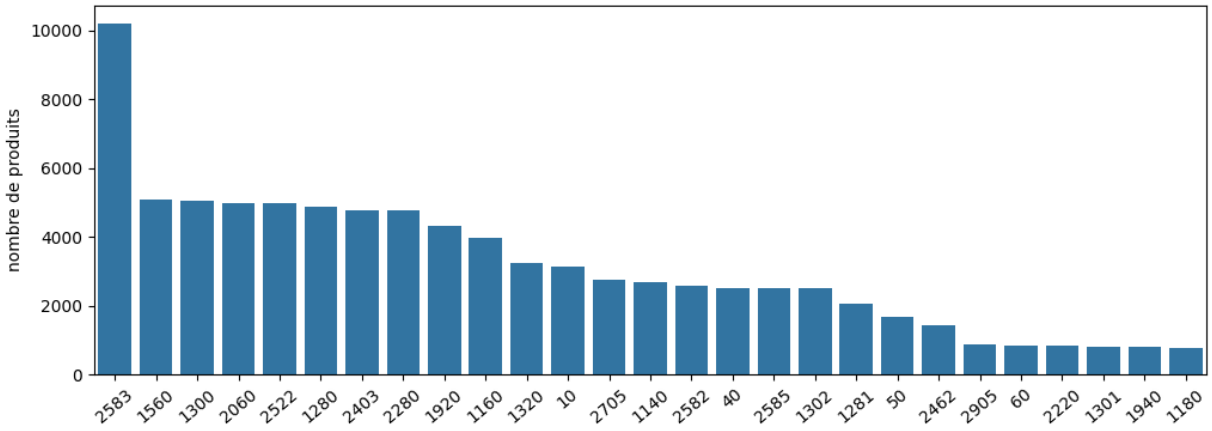
1.3 Présentation du jeu de données et structure du catalogue

Le jeu de données se compose de 84,916 produits. Chaque produit est associé à un titre, une description optionnelle, une image, ainsi qu'une étiquette de catégorie sous forme de code numérique (prdtypecode). La variable cible est prdtypecode, qui regroupe 27 catégories distinctes. À ce stade, ces catégories sont uniquement identifiées par leurs codes, sans interprétation sémantique directe, ce qui impose un travail exploratoire ultérieur pour en comprendre la signification et les relations.

Les données textuelles sont particulièrement volumineuses : les descriptions atteignent en moyenne plusieurs centaines de caractères, avec de fortes disparités selon les catégories. Un point critique à noter est le taux de descriptions manquantes, qui s'élève à environ 35 %. Cette absence d'information textuelle pour une part significative des produits complique la tâche de classification basée uniquement sur le contenu sémantique et nécessite des stratégies de traitement adaptées, comme l'exploitation renforcée de la désignation ou des features visuelles.

Dans le graphique ci-dessous, la distribution des classes met en évidence un déséquilibre marqué du catalogue, avec un ratio d'environ 13:1 entre la catégorie la plus représentée et la moins fréquente. Ce déséquilibre structurel constitue une contrainte majeure pour la modélisation et influence directement le choix des métriques d'évaluation, ainsi que l'adoption systématique de stratégies de découpage stratifié.

Figure 2: Distribution des classes



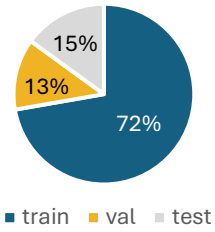
1.4 Stratégie de découpage et protocole d'évaluation

Afin de garantir une évaluation rigoureuse, reproductible et comparable tout au long du projet, une stratégie de découpage fixe a été adoptée dès le départ.

Le jeu de données est d'abord séparé en un ensemble d'entraînement (85 %) et un ensemble de test final (15 %), strictement conservé pour l'évaluation finale.

L'ensemble d'entraînement est ensuite à nouveau scindé en un sous-ensemble d'apprentissage (85 %) et un ensemble de validation (15 %), en appliquant une stratification sur la variable cible afin de préserver la distribution des classes à chaque étape. Cette organisation limite les risques de fuite de données et permet une analyse fiable et cohérente des performances des modèles développés.

Ce déséquilibre structurel a également guidé le choix de la métrique d'évaluation de référence. Afin d'éviter que les catégories minoritaires ne soient pénalisées par leur faible effectif, le F1-score pondéré (F1-weighted) est retenu comme métrique principale tout au long du projet.



Cette métrique permet de tenir compte à la fois de la précision et du rappel, tout en pondérant la contribution de chaque classe par sa fréquence réelle dans le jeu de données.

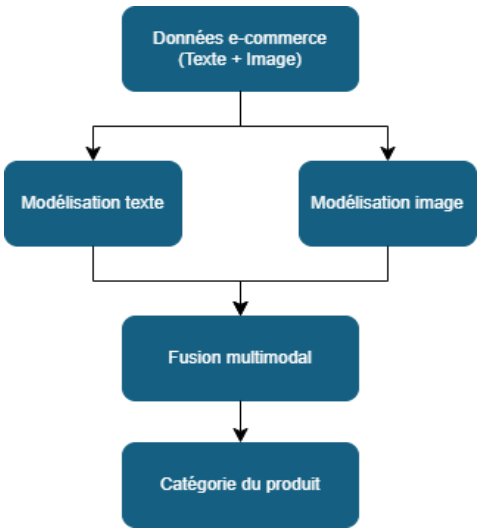
1.6 Organisation générale du projet

Le projet est structuré de manière progressive et modulaire. Le rapport suit cette logique de développement.

Dans un premier temps, nous nous concentrons exclusivement sur la modalité textuelle. Une exploration approfondie des données textuelles est menée afin de caractériser la nature des signaux disponibles, d'évaluer leur qualité et d'identifier les premières pistes de modélisation. Des modèles de référence et des baselines sont alors construits sur le texte seul, servant de points de comparaison solides pour la suite du projet.

Dans un second temps, le même raisonnement est appliqué à la modalité visuelle. Les données d'images sont analysées indépendamment, puis des modèles dédiés sont développés afin d'extraire des représentations pertinentes à partir du signal visuel.

Figure 3: Schéma simplifié du pipeline



Dans un troisième temps, et toujours de manière séparée pour chaque modalité, des modèles plus avancés sont introduits. Nous explorons notamment des architectures de type Transformer pour le texte, ainsi que des réseaux de neurones profonds pour l'image — incluant des modèles convolutionnels et des Transformers visuels —, afin d'améliorer la qualité des représentations apprises au sein de chaque modalité.

Enfin, une étape de fusion tardive permet de combiner les meilleures représentations issues de chaque modalité dans une approche multimodale unifiée, constituant l'aboutissement méthodologique du projet.

II - Partie Texte

2.1 Analyse et Nettoyage des Données Textuelles

2.1.1 Enjeux du Nettoyage dans un Contexte E-commerce

Dans le cadre du challenge Rakuten, les données textuelles proviennent d'un environnement e-commerce réel, caractérisé par une grande hétérogénéité et par un niveau de bruit élevé. Le nettoyage du texte constitue donc une étape critique, car il conditionne directement la qualité des représentations vectorielles et par conséquent, les performances des modèles de classification.

Une contrainte majeure identifiée dès l'analyse initiale est la rareté de l'information : 35,09 % des produits ne disposent pas de description et reposent uniquement sur leur désignation. Dans ce contexte, toute stratégie de nettoyage doit améliorer la lisibilité du texte sans appauvrir excessivement le signal disponible, en particulier pour les catégories déjà pauvres en information.

2.1.2 Diagnostic Initial du Corpus Brut

L'examen du corpus brut met en évidence plusieurs types de bruit.
Sur le plan structurel, de nombreuses balises HTML (
, />), entités encodées (',) et erreurs d'encodage Unicode sont présentes, issues de systèmes de collecte hétérogènes. Ces artefacts génèrent des tokens artificiels lors de la vectorisation et nuisent à la qualité des représentations textuelles.

Sur le plan lexical, l'analyse fréquentielle montre une domination écrasante de mots fonctionnels tels que « de », « et » ou « la », masquant les termes réellement discriminants. À cela s'ajoute un bruit commercial important, mis en évidence par l'analyse en n-grams, avec des expressions promotionnelles standardisées comme « de haute qualité », apparaissant plus de 10 000 fois.

Enfin, l'analyse révèle des incohérences dans les données, notamment la présence de doublons et d'erreurs d'annotation, caractéristiques des catalogues e-commerce à grande échelle.

Exemple d'un début de description brute

```
MATELAS:<br />Â· Accueil : Ferme .<br />Â· Soutien : Très Ferme .<br />Â· Technologie
matelas : Face été &#43; à¢me en Mousse Poli Lattex Dernière Génération Indéformable
Très Haute Résilience - Face Hiver 45 cm de Mousse à\xa0 Mémoire de Forme Très Haute
Densité 60 Kg/m3 &#34; Massante&#34;<br />Â· Épaisseur du matelas : &#43;/- 20 cm.<br
/>Â· REPOS PLUS SAIN grà¢ce au Traitement Anti-acariens / anti-bactérien …
```

2.1.3 Conception d'un Pipeline de Nettoyage Paramétrable

Face à cette diversité de bruit, une fonction de nettoyage paramétrable a été développée afin de tester l'impact de chaque transformation. Cette approche modulaire permet d'activer ou de désactiver indépendamment chaque étape du nettoyage, rendant possible une évaluation empirique rigoureuse. Le pipeline repose sur trois niveaux complémentaires :

1. Le nettoyage structurel vise à restaurer l'intégrité formelle du texte par la correction des encodages, le décodage HTML et la normalisation Unicode.
2. Le raffinement lexical cible des expressions commerciales et des stopwords spécifiques au domaine, identifiés empiriquement lors de l'exploration.
3. Certaines transformations morphologiques, comme la standardisation des dimensions physiques, sont explorées afin d'homogénéiser les formats sans perdre d'information technique.

2.1.4 Impact Qualitatif du Nettoyage sur le Corpus

L'application du pipeline de nettoyage entraîne une transformation profonde du paysage lexical. Avant nettoyage, le vocabulaire est dominé par des artefacts structurels et des mots fonctionnels. Après nettoyage, les termes les plus fréquents correspondent à des concepts techniques et produits concrets, tels que « cm », « piscine », « bois » ou « acier ».

Cette évolution confirme que le nettoyage permet de recentrer le corpus sur un signal sémantique pertinent, condition indispensable pour une modélisation efficace.

Les trois nuages de mots illustrent l'effet du prétraitement textuel : le texte brut reste fortement bruité, l'application des stopwords français réduit partiellement ce bruit mais laisse subsister des balises HTML, tandis que leur suppression permet d'obtenir un vocabulaire plus lisible et interprétable.

2.1.5 Synthèse et Rôle du Nettoyage dans le Pipeline Global

Cette phase de nettoyage constitue une fondation méthodologique essentielle du projet. Elle met en évidence l'importance d'une approche méthodique, évitant à la fois le sous-nettoyage, qui laisse subsister du bruit et le sur-nettoyage, qui risque d'éliminer des indices contextuels utiles.

Les choix opérés à ce stade aboutissent à un corpus textuel nettoyé et structuré, destiné à servir de support aux étapes ultérieures de modélisation. Avant d'introduire des modèles de classification ou d'évaluer des performances, il est cependant indispensable d'examiner empiriquement les caractéristiques de ce corpus.

La section suivante est ainsi consacrée à une exploration approfondie des données textuelles, visant à analyser la disponibilité de l'information, son intégrité, sa structure et sa cohérence sémantique. Cette étape de diagnostic permet d'identifier les contraintes et régularités du corpus, qui orienteront directement la construction des features et le choix des modèles présentés dans les sections ultérieures.

2.2. Exploration Approfondie du Corpus Textuel

2.2.1 Analyse par catégorie : sémantique, qualité et complétude

À ce stade du pipeline, les catégories produits sont uniquement identifiées par des codes, sans interprétation sémantique explicite. Une analyse exploratoire est donc nécessaire afin de labelliser ces catégories, en associant à chaque identifiant une interprétation sémantique fondée sur le contenu textuel des produits correspondants.

Nous visons à expliciter le sens des classes, à analyser leurs proximités lexicales et à vérifier leur cohérence interne, constituant ainsi un prérequis indispensable à toute phase de modélisation supervisée.

L'étude sémantique repose sur des nuages de mots et des heatmaps de mots-clés, permettant de valider la cohérence globale des catégories. Les termes dominants observés correspondent bien aux univers produits associés, confirmant que le signal sémantique reste exploitable malgré le bruit.

Les nuages de mots ont permis d'associer à chaque *prdtypecode* une interprétation sémantique explicite — comme illustré en figure 4, où la catégorie *2583* est clairement identifiée comme relevant de l'univers *Piscine et accessoires* — et de mettre en évidence des proximités lexicales entre certaines catégories, facilitant leur regroupement exploratoire en ensembles thématiques cohérents.

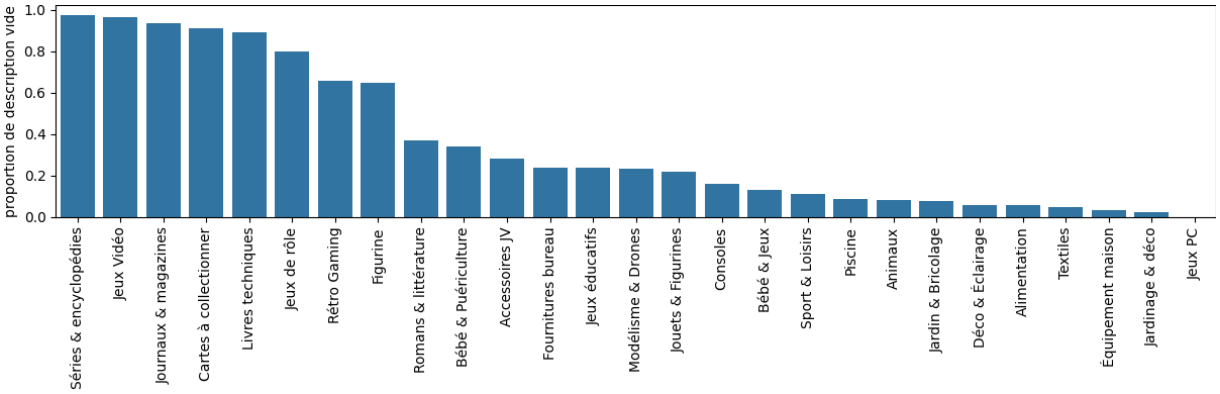
Figure 4: Nuage de mots de la catégorie 2583

Cette analyse a conduit à l'identification de grands ensembles thématiques regroupant plusieurs catégories partageant un socle lexical commun. Ces regroupements, utilisés uniquement à des fins exploratoires, ont notamment permis d'identifier des catégories sémantiquement proches susceptibles d'être confondues, tout en confirmant la cohérence globale du signal textuel.

Dans ce cadre désormais clarifié, l'analyse peut être prolongée par l'étude de la qualité et de la complétude des données textuelles. L'exploration révèle en effet une forte hétérogénéité dans la quantité d'information disponible : bien que 35,09 % des produits ne disposent pas de description, cette proportion varie fortement selon les catégories.

Certaines catégories culturelles reposent presque exclusivement sur la désignation, tandis que d'autres, comme les jeux PC en téléchargement, présentent des descriptions particulièrement longues, parfois supérieures à 2 000 caractères.

Figure 5: Taux de description manquantes par catégorie



2.2.3 Caractérisation Structurelle et Technicité des Produits

Au-delà du contenu lexical, plusieurs indicateurs structurels sont extraits afin de caractériser la nature des descriptions.

La densité de chiffres, la présence d'unités de mesure et les marqueurs de numérotation révèlent une frontière nette entre catégories techniques et catégories narratives.

Les catégories techniques comme Bricolage & Outillage ou Textiles d'intérieur reposent fortement sur des mesures et des spécifications, tandis que les catégories culturelles comme Jouets & Figurines mobilisent principalement un discours descriptif.

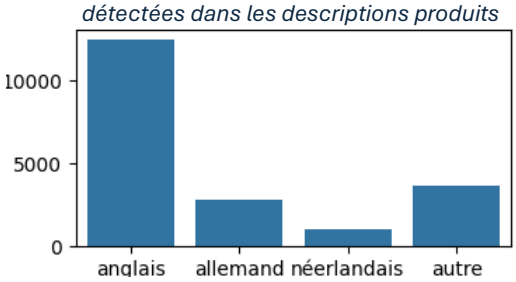
Certaines features, comme le marqueur "n°", se révèlent hautement discriminantes pour des catégories spécifiques come Littérature, Livres spécialisés ou Presse & Magazines.

2.2.4 Analyse Linguistique du Corpus

Un audit linguistique met en évidence une diversité linguistique significative. Bien que le français reste majoritaire, près d'un quart des descriptions (23 %) sont détectées comme rédigées dans une langue étrangère, principalement dans les catégories techniques et de loisirs.

Cette diversité limite l'efficacité de modèles strictement francophones et justifie l'exploration de modèles multilingues dans les étapes ultérieures.

Figure 6: fréquence des langues étrangères détectées dans les descriptions produits



2.2.5 Synthèse et Enseignements pour la Modélisation

Cette phase d'exploration met en lumière des contraintes structurelles fortes : déséquilibre des classes, rareté et hétérogénéité de l'information textuelle, bruit d'annotation. Elles révèlent, également, l'existence de régularités sémantiques et structurelles exploitables.

Les analyses lexicales, structurelles et linguistiques conduites à ce stade apportent des enseignements déterminants pour la conception de features pertinentes, l'adaptation des stratégies de vectorisation et le choix de modèles capables de capturer des signaux plus complexes.

2.3. Modèles de Référence et Baseline Textuelle

2.3.1 Objectifs et Rôle de la Baseline

Avant d'introduire des représentations vectorielles complexes ou des modèles profonds, il est essentiel d'établir une baseline simple, interprétable et reproductible. L'objectif de cette section est d'évaluer jusqu'où des features textuelles élémentaires, issues directement de l'exploration du corpus, peuvent prédire les catégories produits.

Cette baseline ne cherche pas l'optimisation maximale, mais constitue un point de référence solide, permettant de mesurer de manière rigoureuse les gains apportés par des approches plus avancées et d'éclairer le choix des futures stratégies de modélisation.

2.3.2 Construction des Features Textuelles Heuristiques

Les features utilisées dans cette baseline sont directement dérivées des enseignements de l'exploration du corpus. Trois grandes familles de variables sont :

1. Les mots-clés constituent la première source d'information. Ils reposent sur des dictionnaires manuellement définis pour certaines familles de produits et permettent de capter des indices lexicaux très spécifiques.
2. Les indicateurs techniques forment la seconde famille de features. Ils détectent la présence d'unités de mesure, de quantités et de marqueurs de numérotation (comme « n° »), qui se sont révélés caractéristiques des catégories techniques.
3. Enfin, des variables descriptives simples, telles que la longueur du texte en caractères ou en tokens, sont ajoutées afin de capturer les différences structurelles entre descriptions courtes et descriptions longues.

2.3.3 Comparaison des Modèles de Référence

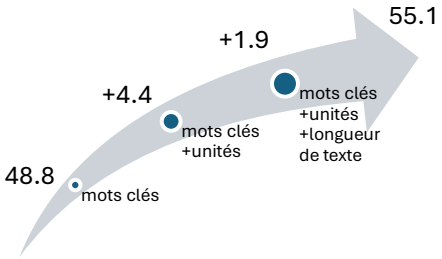
Trois algorithmes de classification sont évalués sur ces features : la Régression Logistique, le Linear SVC et l'Arbre de Décision. Cette comparaison vise à identifier le modèle le plus adapté à des représentations heuristiques de faible dimension.

Les résultats montrent que la Régression Logistique surpasse systématiquement les autres approches, avec un F1-weighted initial de 0,488 basé uniquement sur les mots-clés. Le Linear SVC obtient des performances proches mais légèrement inférieures, tandis que l'Arbre de Décision souffre fortement du bruit et du déséquilibre des classes.

2.3.4 Apport Progressif des Familles de Features

L'analyse des performances met en évidence un effet cumulatif clair lors de l'ajout progressif des différentes familles de features.

Le modèle basé uniquement sur les mots-clés atteint un F1-weighted de 0,488 sur le jeu de validation. L'ajout des indicateurs techniques (unités et numérotation) permet d'atteindre 0,532, confirmant leur fort pouvoir discriminant pour les catégories techniques. L'intégration des variables de longueur de texte améliore encore les performances, portant le score de validation à 0,551.

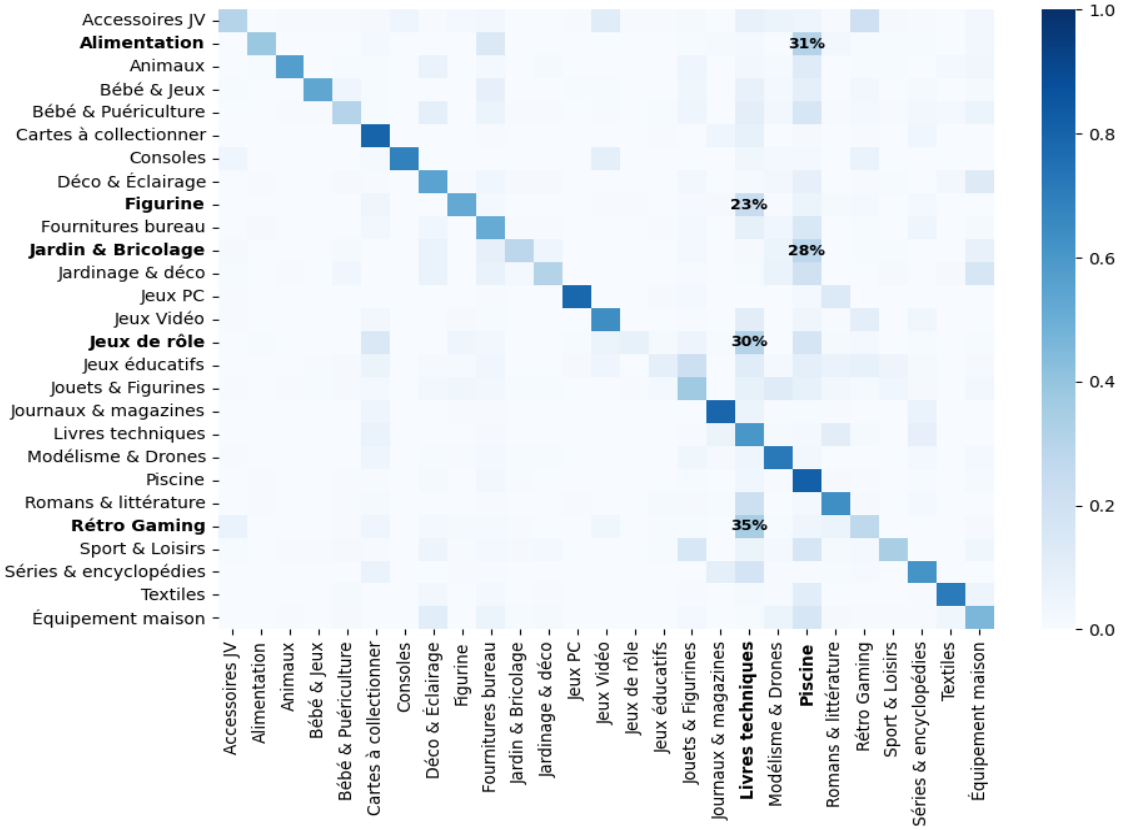


Ce gain progressif démontre que les différentes familles de features apportent des signaux complémentaires.

2.3.5 Analyse des Résultats et Limites de la Baseline

Sur le jeu de test le F1-score pondéré atteint 55,2 %. En regardant plus en détail les résultats mettent en évidence un biais marqué du modèle vers certaines catégories dominantes, en particulier Livres techniques et Piscine. Ces classes présentent un rappel élevé (respectivement 0,60 et 0,82), indiquant qu'elles sont fréquemment prédites, mais au prix d'une précision faible, signe d'une sur-prédiction. Ce phénomène se traduit par des confusions particulièrement élevées et parfois contre-intuitives : par exemple, 35 % des produits de la catégorie Rétro Gaming sont prédits comme Livres techniques, ou encore 31 % des produits Alimentation sont confondus avec Piscine. L'analyse des scores F1 par catégorie confirme cette tendance. Les catégories fortement impactées par ces confusions, telles que Jeux de rôle, Jeux éducatifs ou Rétro Gaming, affichent des rappels très faibles et des F1-scores dégradés.

Figure 7:matrice de corrélation normalisé par ligne (vraie classe)



2.3.6 Conclusion : Le Plafond Sémantique des Features Simples

Cette baseline démontre que des features simples, manuellement construites, offrent une base robuste, interprétable et reproductible. Cependant, les performances atteignent rapidement un plafond structurel, notamment pour les catégories narratives où le sens ne peut être réduit à la présence de mots-clés ou de marqueurs techniques.

Ces résultats justifient le recours à des méthodes de vectorisation automatique et à des modèles capables de capturer le contexte et la sémantique du texte, qui seront abordés dans les sections suivantes.

2.4. Vectorisation et Optimisation des Représentations Textuelles

2.4.1 Objectifs et Enjeux de la Vectorisation

Après l'établissement d'une baseline fondée sur des features heuristiques, cette section vise à optimiser la représentation textuelle afin de capturer des régularités sémantiques plus fines. L'objectif n'est pas uniquement d'augmenter les performances, mais de comprendre l'impact de chaque choix méthodologique (vectorisation, n-grams, pondération) sur la capacité du modèle à généraliser.

Cette étape constitue un pivot du projet, car elle fournit un modèle texte robuste et calibré, destiné à servir de composant principal dans une stratégie de fusion ou de stacking ultérieure.

2.4.2 Choix du Modèle et Cadre Expérimental

Une phase de benchmark initial a été menée afin de comparer plusieurs classifieurs linéaires sur une vectorisation TF-IDF fixée. Les modèles évalués incluent un SVM linéaire (LinearSVC), une régression logistique, un classifieur Ridge, ainsi que des approches probabilistes de type Naive Bayes.

Les résultats mettent en évidence la supériorité du LinearSVC, qui atteint un F1-score de 0,838 sur le jeu de validation, contre 0,833 pour le RidgeClassifier et 0,816 pour la régression logistique, tout en conservant un temps d'entraînement significativement plus faible que ce dernier (28 s contre plus de 230 s sur la même machine). À l'inverse, les modèles de type Naive Bayes, bien que rapides, présentent des performances sensiblement inférieures, traduisant une capacité de séparation plus limitée dans cet espace de représentation.

Ce compromis favorable entre performance, stabilité et coût computationnel justifie le choix du LinearSVC comme modèle de référence. Il s'inscrit dans une logique de robustesse : les modèles linéaires sont particulièrement adaptés aux espaces de grande dimension et fortement clairsemés induits par la vectorisation textuelle TF-IDF, où une frontière de décision linéaire suffit souvent à capturer l'essentiel du signal discriminant.

2.4.3 Optimisation de la Vectorisation Textuelle

La phase d'optimisation de la vectorisation met en évidence des écarts significatifs entre les approches basées sur le simple comptage et la pondération TF-IDF. La vectorisation TF-IDF associée à un SVM linéaire atteint un F1-score de 0,838 sur le jeu de validation, constituant la meilleure performance observée.

À l'inverse, les représentations par comptage binaire et par comptage brut obtiennent des scores quasi parfaits sur l'entraînement (F1 $\simeq$ 0,998), mais présentent une dégradation marquée en validation, avec des F1-scores respectifs de 0,818 et 0,808, soit une perte de 2 à 3 points par rapport à TF-IDF. Cet écart

important entre entraînement et validation traduit un sur-apprentissage prononcé, indiquant que ces représentations capturent des régularités trop spécifiques au jeu d'entraînement. Par ailleurs, les méthodes par comptage se révèlent également plus coûteuses en temps d'entraînement, avec des durées qui peuvent être 3 fois plus longue que pour TF-IDF, sans gain de performance en généralisation.

Concernant les n-grams[1], l'évaluation montre que la combinaison des unigrams et bigrams (1,2) constitue le meilleur compromis entre expressivité et généralisation. L'ajout des trigrams (1,3) n'apporte aucun gain mesurable, entraîne une légère dégradation des performances, et augmente inutilement la complexité du modèle, en raison d'un espace de représentation plus vaste et plus clairsemé.

2.4.4 Apport des N-grams de Caractères

L'introduction des n-grams de caractères constitue l'un des apports majeurs de cette section. Ces représentations permettent de capturer des motifs morphologiques locaux et offrent une forte robustesse aux variations orthographiques et lexicales fréquentes dans les descriptions produits, ce qui les rend particulièrement adaptées au contexte e-commerce.

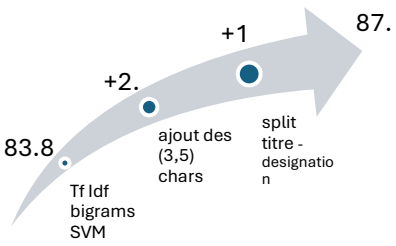
À titre d'illustration, les descriptions « piscine gonflable enfant » et « piscines gonflables enfants » diffèrent au niveau des mots, mais partagent de nombreux motifs communs lorsqu'elles sont représentées par des character n-grams (3–5), tels que « pisc », « gonfl » ou « enfan ». Cette approche permet ainsi de capturer efficacement les variations morphologiques tout en conservant un fort pouvoir discriminant.

D'un point de vue quantitatif, l'ajout des n-grams de caractères a été évalué selon deux configurations, (3–5) et (4–6) caractères. La configuration (3–5) conduit à une amélioration significative du F1-weighted, qui passe de 0,838 avec une vectorisation par mots seule à 0,860, tandis que la configuration (4–6) atteint un F1-score de 0,857, sans gain supplémentaire. Ces résultats confirment l'intérêt des character n-grams pour exploiter les régularités locales présentes dans des textes courts, hétérogènes et riches en abréviations, typiques des descriptions produits.

2.4.5 Pondération Différentielle du Titre et de la Description

Plutôt que de fusionner aveuglément l'ensemble du texte, le titre (désignation) et la description sont traités comme deux sources d'information distinctes. En effet, l'analyse exploratoire et les premières expérimentations montrent que le titre concentre, dans la majorité des cas, les signaux les plus directement discriminants alors que la description apporte une information plus détaillée, mais également plus hétérogène et parfois bruitée.

C'est pour cette raison qu'une pondération explicite du titre est introduite dans le modèle, afin de refléter son importance relative dans le processus de classification. L'expérimentation met en évidence l'existence d'un optimum pour un poids de 1,2 : en deçà, l'information portée par le titre n'est pas suffisamment valorisée ; au-delà, le modèle tend à sur-apprendre, en sur-exploitant des signaux déjà dominants.



[1] Un n-gramme est un ensemble de n éléments successifs dans un document texte pouvant comprendre des mots, des nombres, des symboles et de la ponctuation

Ce choix méthodologique permet ainsi d'améliorer significativement les performances globales du modèle, conduisant à un F1-weighted proche de 0,87. Ce résultat confirme qu'un traitement différencié et raisonné des composantes textuelles est préférable à une fusion uniforme.

2.4.6 Régularisation et Généralisation

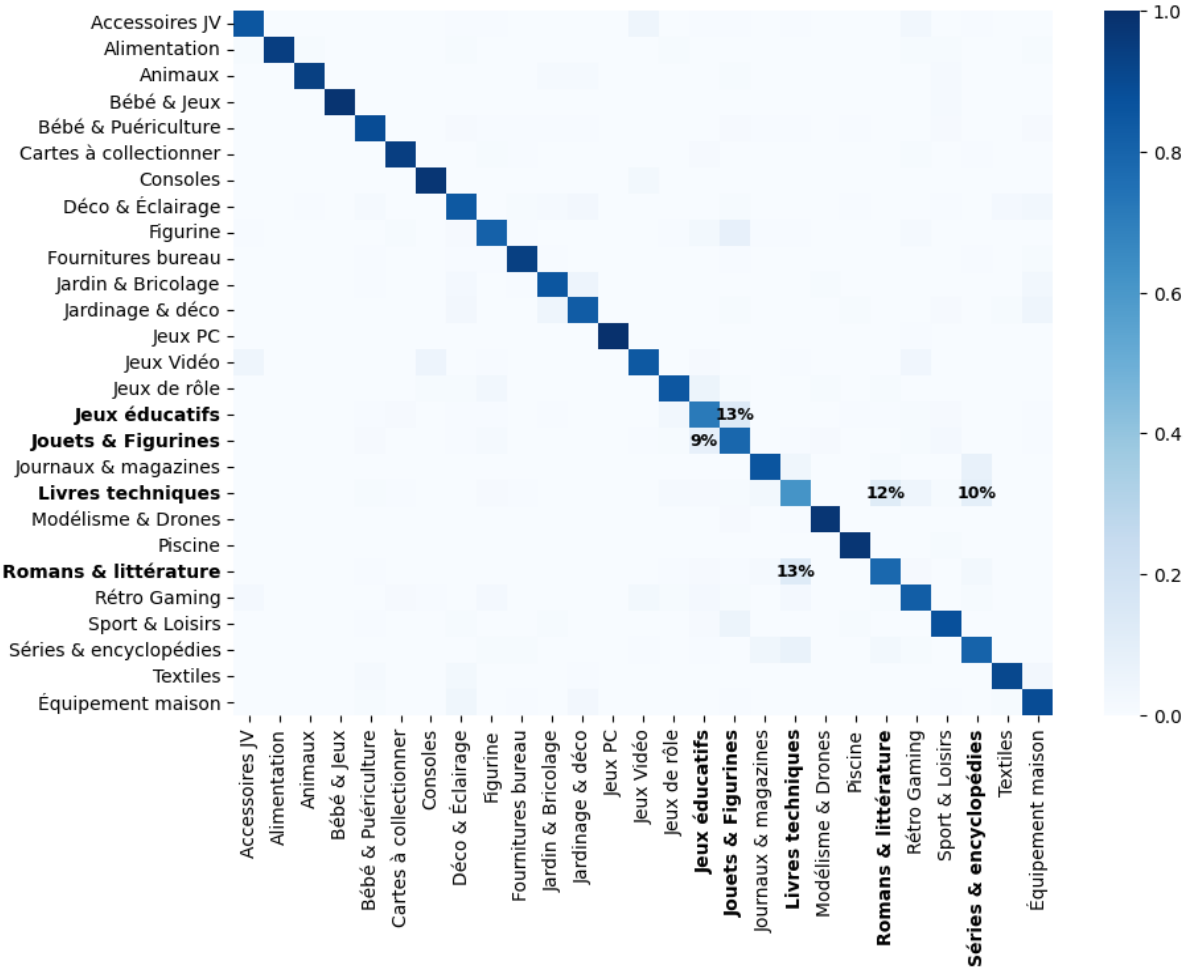
L'analyse de l'influence du paramètre de régularisation *C* du LinearSVC met en évidence l'existence d'un plateau de performance pour des valeurs comprises entre 0,2 et 1,0. Cette stabilité des résultats suggère que, dans cet intervalle, le modèle est relativement peu sensible à de légères variations du niveau de régularisation.

C'est pour cette raison qu'une valeur conservatrice de *C = 0,2* est retenue. Ce choix permet de maintenir des performances élevées tout en limitant l'écart observé entre les scores d'entraînement et de validation, indiquant un meilleur contrôle du risque de sur-apprentissage.

Ce choix méthodologique illustre ainsi une volonté assumée de privilégier la capacité de généralisation du modèle plutôt que l'optimisation locale des scores sur les données d'entraînement

2.4.7 Évaluation finale du modèle texte

Figure 8: Matrice corrélation du SVM



L'évaluation finale du modèle texte sur le jeu de test conduit à un F1-score pondéré de 0,873, confirmant une bonne capacité de généralisation et la pertinence des choix méthodologiques effectués lors des phases de benchmark.

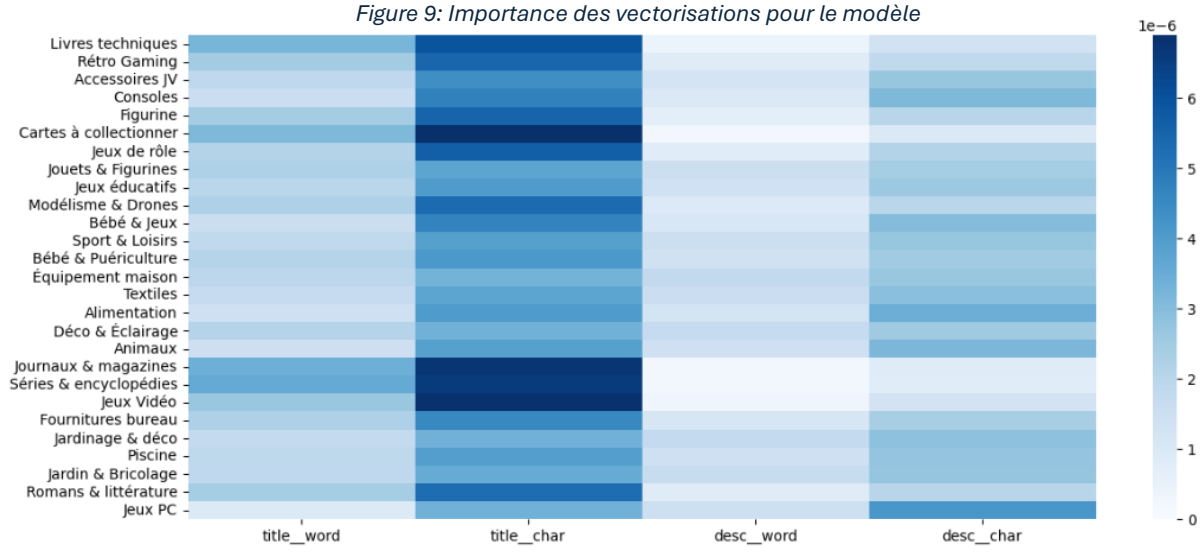
L'analyse des erreurs montre que les principales confusions concernent des catégories sémantiquement proches, partageant des vocabulaires et des usages similaires. Les confusions les plus marquées apparaissent ainsi par binômes cohérents :

- Jeux éducatifs ↔ Jouets & Figurines : ces deux catégories présentent des descriptions souvent génériques et orientées vers l'usage du produit, ce qui rend leur distinction difficile sur la seule base du texte.
- Romans & littérature ↔ Livres techniques : le vocabulaire commun lié au format livre et la faible spécialisation de certaines descriptions conduisent à des confusions réciproques entre ces deux classes.
- Jeux de rôle ↔ catégories de jeux proches : la présence de marqueurs lexicaux spécifiques entraîne une tendance du modèle à sur-prédire cette catégorie, avec un rappel élevé mais une précision plus faible.

Ces confusions apparaissent cohérentes avec les observations issues de l'analyse exploratoire et traduisent davantage des chevauchements sémantiques intrinsèques aux catégories que des limites structurelles du modèle texte. Elles fournissent un cadre d'interprétation utile pour l'analyse qualitative présentée dans la section suivante, ainsi que pour l'étude des gains attendus lors de la fusion multi-modale.

2.4.8 Analyse Qualitative et Interprétabilité

Figure 9: Importance des vectorisations pour le modèle



La heatmap ci-dessus présente le poids moyen des coefficients du modèle SVM pour les différentes modalités de vectorisation textuelle considérées. Les quatre représentations analysées sont :

- la vectorisation du titre à l'aide de mots unigrammes et bigrammes (title__word);
- la vectorisation du titre par n-grams de caractères (title__char);
- la vectorisation de la description à l'aide de mots unigrammes et bigrammes (desc__word);
- la vectorisation de la description par n-grams de caractères (desc__char).

Cette visualisation met en évidence le rôle prépondérant du titre dans la construction des hyperplans de séparation du modèle. En particulier, la vectorisation par n-grams de caractères appliqués au titre apparaît comme la source de signal la plus discriminante, contribuant fortement à la séparation des catégories de produits. Ce résultat souligne l'importance de motifs lexicaux courts, robustes aux variations orthographiques et morphologiques, fréquemment présents dans les titres produits.

Néanmoins, ce comportement n'est pas homogène selon les catégories. Pour les produits de type jeux PC, la description joue un rôle plus important dans la décision du modèle. Cette observation est cohérente avec les résultats de l'analyse exploratoire, qui ont montré que les descriptions associées à cette catégorie sont généralement longues, détaillées et riches en informations techniques.

À l'inverse, pour des catégories telles que les cartes à collectionner, où les descriptions sont le plus souvent absentes ou très succinctes (voir figure 4), la séparation repose quasi exclusivement sur les informations contenues dans le titre. Dans ce cas, la vectorisation textuelle du titre constitue l'unique source de signal exploitable par le modèle pour discriminer les produits.

2.5. Modélisation Transformer : Approches Monolingues et Multilingues

2.5.1 Objectifs et Positionnement Méthodologique

Cette section marque une rupture méthodologique avec les approches précédentes fondées sur des représentations statistiques du texte. L'objectif est d'évaluer l'apport de modèles de type Transformer et d'analyser dans quelle mesure les stratégies de prétraitement influencent leurs performances. Au-delà du score final, l'enjeu est de comprendre l'impact du prétraitement sur la dynamique d'apprentissage.

2.5.2 Cadre Expérimental et Modèles Évalués

L'ensemble des expériences repose sur le modèle pré-entraîné CamemBERT, dans ses versions base et large. Afin de garantir une comparaison équitable, les hyperparamètres sont maintenus constants pour toutes les expériences : taux d'apprentissage fixé à 2e-5, batch size de 64, huit époques d'entraînement et une longueur maximale de séquence de 384 tokens.

Cette standardisation permet d'attribuer les variations de performance observées uniquement aux choix de prétraitement ou à la capacité du modèle.

2.5.3 Stratégies de Prétraitement Comparées

Plusieurs stratégies de prétraitement sont évaluées de manière systématique afin d'analyser leur impact sur les performances des modèles de type Transformer. Le texte brut constitue la référence, exploitant directement les capacités du modèle pré-entraîné sans transformation explicite.

En complément, deux niveaux de prétraitement ciblant spécifiquement les quantités numériques sont considérés : une stratégie light, reposant sur un remplacement partiel et conservateur des nombres, et une stratégie full, consistant à transformer la quasi-totalité des expressions numériques en tokens abstraits.

Texte brute

```
CONQUERANT CLASSIQUE Cahier 240 x 320 mm seyès incolorecouverture en Polypro 96 pages
agrafé papier de 90 g/m2(400006764)
```

Texte avec token numérique version « light »

```
CONQUERANT CLASSIQUE Cahier [petite_surface] seyès incolorecouverture en Polypro 96 pages
agrafé papier de [poids_leger]/m2(400006764)
```

Texte avec token numérique version « full »

```
CONQUERANT CLASSIQUE Cahier [petite_surface] seyès incolorecouverture en Polypro
[grand_nombre] pages agrafé papier de [poids_leger]/m2([très_grand_nombre])
```

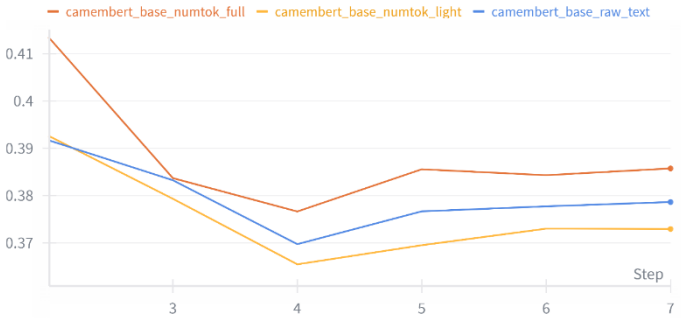
Ces trois configurations — brut, numérique light et numérique full — permettent d'évaluer dans quelle mesure une abstraction explicite des quantités numériques apporte un signal discriminant supplémentaire ou, au contraire, dégrade la compréhension contextuelle. Elles visent également à analyser la capacité intrinsèque des modèles Transformer à capturer, interpréter et exploiter l'information numérique présente dans les descriptions produits.

2.5.4 Analyse des Performances et Dynamique d'Apprentissage

Les résultats obtenus avec CamemBERT-base montrent que la stratégie de prétraitement *light* constitue un compromis particulièrement pertinent. Elle conduit à une minimisation de la cross-entropy plus grande tout en conservant un F1-score comparable à celui obtenu avec le texte brut. Ce comportement suggère que le modèle bénéficie d'une abstraction partielle des quantités numériques, qui réduit le bruit sans altérer la compréhension contextuelle.

À l'inverse, la stratégie *full* apparaît trop destructrice pour un modèle de type Transformer. En discrétisant de manière systématique les informations numériques, elle supprime des indices quantitatifs et relationnels que le modèle est pourtant capable d'exploiter nativement grâce à son pré-entraînement. Cette sur-normalisation introduit ainsi un appauvrissement du signal textuel, se traduisant par une dégradation des performances.

Figure 10 : courbe de l'entropie croisée avec les 3 pré-traitements

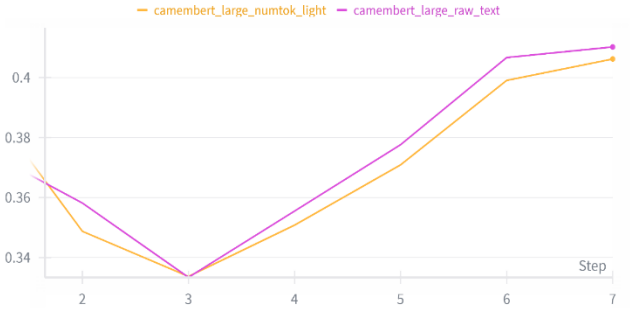


Ces observations confirment que, pour des modèles fortement contextualisés, un prétraitement excessif peut être contre-productif, et que des transformations légères, respectant la structure naturelle du texte, sont à privilégier.

2.5.5 Influence de la Capacité du Modèle

Afin d'évaluer l'influence de la capacité du modèle sur l'efficacité des stratégies de prétraitement numérique, l'expérience est reproduite avec CamemBERT-large, qui dispose de 24 couches d'attention, contre 12 pour CamemBERT-base. Deux configurations sont comparées : le texte brut, utilisé comme référence, et la stratégie de discrétisation numérique légère (num tok light). Cette analyse vise à déterminer si l'augmentation du contexte et de la profondeur des représentations modifie les effets observés précédemment avec CamemBERT-base.

Figure 11: courbe de l'entropie croisée de camembert large

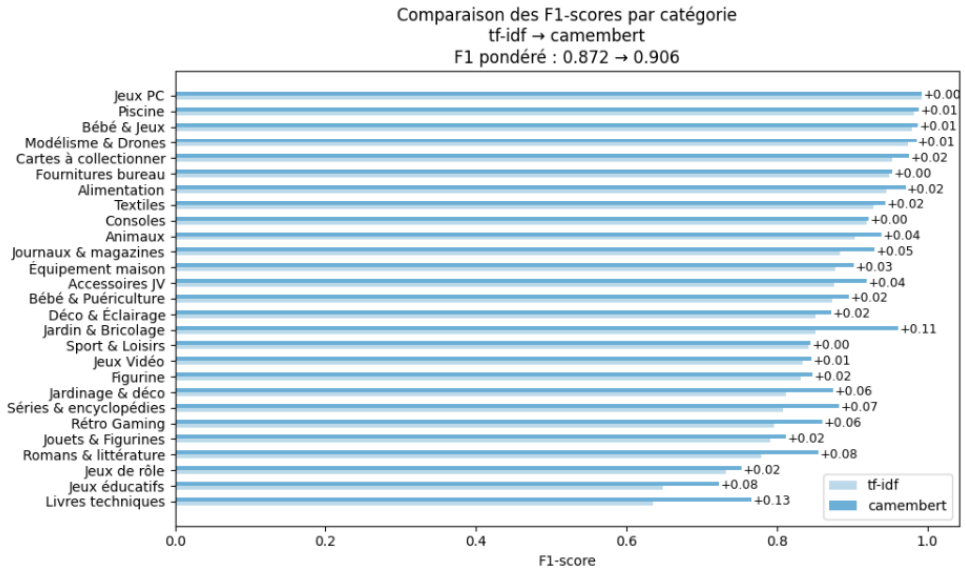


À l'instar des résultats obtenus avec la version base, la stratégie num tok light conduit à des performances légèrement supérieures. Le F1-score de validation atteint 0,9106 à l'epoch 8 pour num tok light, contre 0,9096 pour le texte brut. L'écart observé sur la fonction de perte est moins marqué que celui constaté avec CamemBERT-base. Cela pourrait suggérer que l'augmentation de la capacité du modèle et du contexte disponible avec Camembert-large réduit la sensibilité aux stratégies de prétraitement numérique.

Par ailleurs, la comparaison entre les versions base et large met en évidence un gain global substantiel : le F1-score de validation passe d'environ 0,896 avec CamemBERT-base à environ 0,91 avec CamemBERT-large, indépendamment de la stratégie de prétraitement. Ce résultat confirme l'apport déterminant de la capacité du modèle sur les performances globales.

Dans la suite du travail, le modèle CamemBERT-large avec la stratégie num tok light est retenu, ce choix reposant sur le meilleur F1-score observé, tout en conservant une dynamique d'apprentissage plus stable. Ce modèle fera l'objet de l'évaluation finale et des analyses approfondies présentées dans les sections suivantes.

2.5.6 Comparaison avec les Approches Classiques



Une comparaison directe avec la meilleure approche TF-IDF met en évidence un saut de performance majeur. Le F1-weighted global passe de 0,872 à 0,906 sur le jeu de test, avec une amélioration stable ou positive pour l'ensemble des catégories.

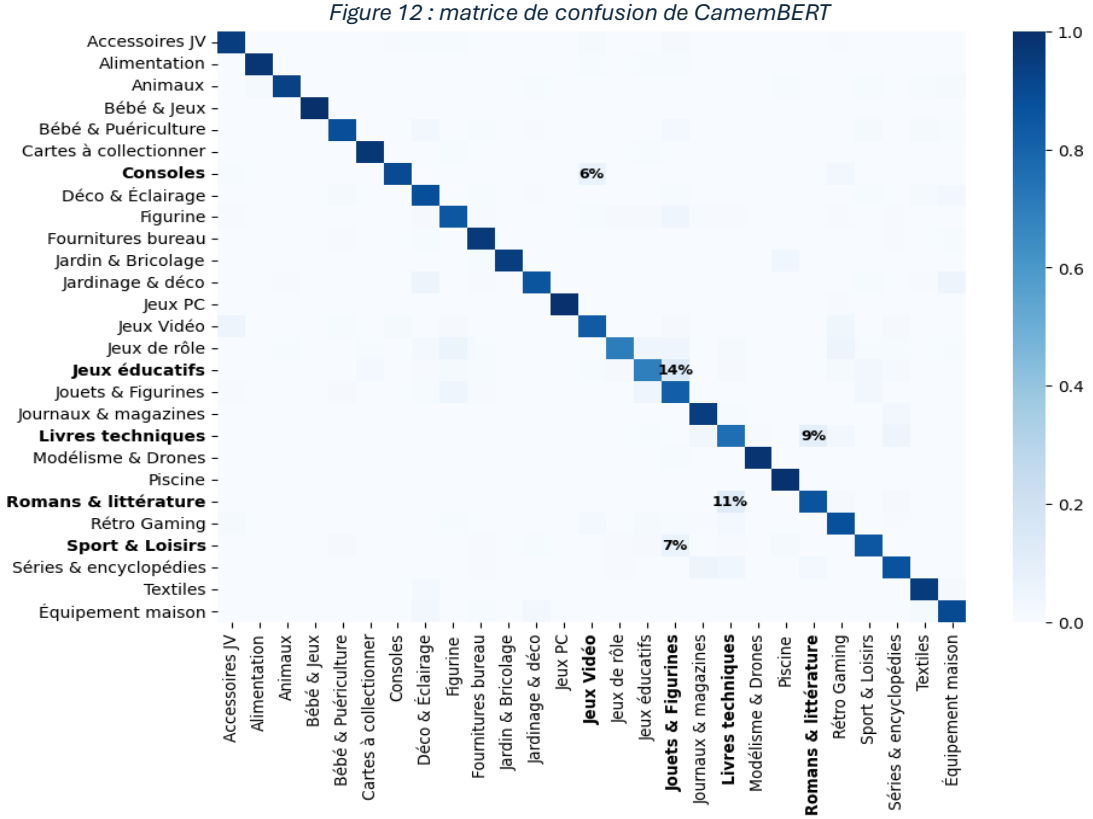
Les améliorations les plus marquées concernent des catégories plus complexes ou plus hétérogènes, telles que Jardin & Bricolage, Jardinage & déco, ainsi que des catégories historiquement difficiles à discriminer car souvent confondues entre elles, comme Livres techniques, Romans & littérature et Séries & encyclopédies ou Jeux éducatifs et Jouets & Figurines. Sur ces classes, les gains peuvent dépasser +10 points de F1.

2.5.7 Analyse des Erreurs Résiduelles

Malgré les améliorations significatives obtenues, certaines confusions subsistent. Les catégories Jeux éducatifs et Jouets ou figurines, jeux de rôle demeurent difficiles à différencier. De même, bien que le modèle Camembert apporte une amélioration notable par rapport à la simple vectorisation, les catégories Romans et Livres techniques restent fréquemment confondues.

Ces erreurs mettent en évidence les limites actuelles d'une approche reposant uniquement sur le signal textuel et suggèrent l'intérêt d'une fusion multimodale pour améliorer les performances.

Figure 12 : matrice de confusion de CamemBERT



2.5.8 Ouverture Multilingue : XLM-RoBERTa et Complémentarité

Après l'identification de CamemBERT-large avec tokenisation numérique légère comme meilleure configuration monomodèle, l'objectif n'est plus d'optimiser un modèle pris isolément, mais d'explorer une architecture complémentaire susceptible d'enrichir la diversité des représentations textuelles.

Dans ce cadre, XLM-RoBERTa-base est introduit comme modèle Transformer multilingue, pré-entraîné sur un large corpus couvrant de nombreuses langues et registres rédactionnels, afin d'évaluer sa capacité à capturer des signaux textuels différents ou complémentaires de ceux appris par CamemBERT. Son caractère multilingue constitue un atout dans le contexte du catalogue Rakuten, où, malgré la prédominance du français, les analyses exploratoires ont mis en évidence la présence de descriptions partiellement ou entièrement rédigées en anglais, allemand ou italien, ainsi qu'un usage fréquent d'anglicismes.

Dans une optique de fusion avec CamemBERT, le choix a été fait d'utiliser XLM-RoBERTa sur du texte brut, sans tokenisation numérique spécifique, afin d'exploiter une représentation textuelle différente de celle utilisée par CamemBERT.

Enfin, l'analyse des erreurs montre que XLM-RoBERTa parvient à produire des prédictions correctes pour environ 30 % des erreurs commises par CamemBERT sur le jeu de validation, ce qui suggère que ce modèle est en mesure d'apporter une contribution non négligeable dans une stratégie de fusion des modèles textuels.

2.6. Fusion des Modèles Texte : Blending et Stacking

2.6.1 Objectifs et préparation probabiliste à la fusion

Cette section constitue l'aboutissement de la chaîne de modélisation textuelle. Après l'exploration de modèles statistiques (TF-IDF/SVM), de Transformers monolingues (CamemBERT) et multilingues (XLM-RoBERTa), l'objectif n'est plus d'optimiser un modèle pris isolément, mais de combiner leurs forces respectives afin d'obtenir un système plus robuste et plus fiable. La fusion vise en particulier à exploiter la complémentarité entre des modèles aux biais inductifs différents, capables de corriger mutuellement certaines erreurs.

Dans cette perspective, un enjeu central réside dans la comparabilité et la fiabilité des sorties probabilistes utilisées pour la fusion. Les différents modèles considérés ne produisent pas naturellement des probabilités homogènes : le modèle TF-IDF/SVM fournit des scores de décision, tandis que les Transformers génèrent des logits issus de fonctions de sortie non calibrées. Or, une fusion fondée sur les probabilités nécessite que ces sorties reflètent un niveau de confiance cohérent.

Pour répondre à cette contrainte, une étape de calibration des sorties est systématiquement appliquée. Les scores du LinearSVC sont transformés en probabilités via une calibration sigmoïde, permettant de rendre ces sorties exploitables dans un cadre probabiliste. Cette opération a un impact marginal sur le F1-score, du modèle mais améliore significativement la qualité des probabilités. De manière analogue, les logits produits par les modèles Transformers sont calibrés a posteriori par temperature scaling.

Cette harmonisation des sorties constitue un prérequis essentiel pour les méthodes de fusion tardive, en garantissant que les probabilités combinées traduisent des niveaux de confiance comparables et exploitables, indépendamment de l'architecture sous-jacente des modèles.

2.6.2 Cadre Expérimental

La fusion repose sur un protocole expérimental rigoureux visant à éviter toute fuite d'information. Le jeu de validation est scindé selon un ratio 70/30 : la première partie est utilisée pour estimer les paramètres de fusion, tandis que la seconde est réservée exclusivement à l'évaluation finale des différentes stratégies.

Deux approches de fusion tardive sont comparées :

- Blending pondéré : cette approche consiste en une combinaison linéaire des probabilités produites par chaque modèle. Les poids sont déterminés à partir de l'inverse de la log loss de chaque composant, de manière à favoriser les modèles dont les estimations probabilistes sont les plus fiables.
- Stacking : cette seconde approche repose sur l'entraînement d'un méta-modèle de type régression logistique à partir des probabilités fournies par les modèles de base. Ce choix vise à conserver une architecture simple et interprétable, tout en limitant les risques de sur-apprentissage.

Pour chacune de ces approches, trois combinaisons de modèles sont évaluées afin d'analyser l'apport relatif de chaque composant et le degré de complémentarité entre les représentations textuelles :

- CamemBERT + TF-IDF/SVM
- CamemBERT + XLM-RoBERTa
- CamemBERT + TF-IDF/SVM + XLM-RoBERTa

2.6.3 Résultats Comparatifs et Analyse Globale

Figure 13:Tableau comparatif de fusion de modèles textes

		f1 (%)	log loss
CamemBERT + TF-IDF	blending	91,05	0,3418
	stacking	91,06	0,3432
CamemBERT + XLM-R	blending	91,19	0,3107
	stacking	91,05	0,3510
CamemBERT + XLM-R + TF-IDF	blending	91,57	0,3226
	stacking	91,27	0,3261

Les combinaisons impliquant les deux Transformers (CamemBERT + XLM-R) obtiennent des résultats largement meilleurs en blending qu'en stacking, en particulier pour la combinaison double. En revanche, la combinaison CamemBERT + TF-IDF montre des performances très proches dans les deux stratégies. Cette différence peut s'expliquer par la proximité des sorties des deux modèles Transformers.

En effet, lorsque les modèles de base sont fortement corrélés, un méta-modèle de stacking peut sur-ajuster des différences spécifiques au jeu d'entraînement qui ne reflètent pas de vrais signaux. Si les probabilités des modèles sont très proches, le méta-modèle dispose de peu de variations pour apprendre, ce qui limite sa capacité à généraliser. Dans ce cas, une combinaison simple pondérée (blending) reste plus robuste.

Parmi toutes les configurations testées, les meilleures performances sont obtenues avec le blending CamemBERT + XLM-R, avec ou sans TF-IDF :

- Sans TF-IDF : log loss la plus basse (-1.2)
- Avec TF-IDF : F1-weighted le plus élevé (+0.4 pt)

Dans notre contexte, où le F1-score est prioritaire, nous privilégions la configuration CamemBERT + XLM-R + TF-IDF, qui sera donc retenue pour la suite.

2.6.4 Evaluation

Le modèle texte final repose sur une stratégie de blending pondéré combinant CamemBERT-large, XLM-R et des features TF-IDF. Il atteint un F1-score pondéré global de 0,912 sur le jeu de test final, contre 0,906 pour CamemBERT-large seul, soit un gain absolu de +0,006. Bien que modéré en valeur absolue, ce gain est stable et obtenu sans dégradation significative sur d'autres classes.

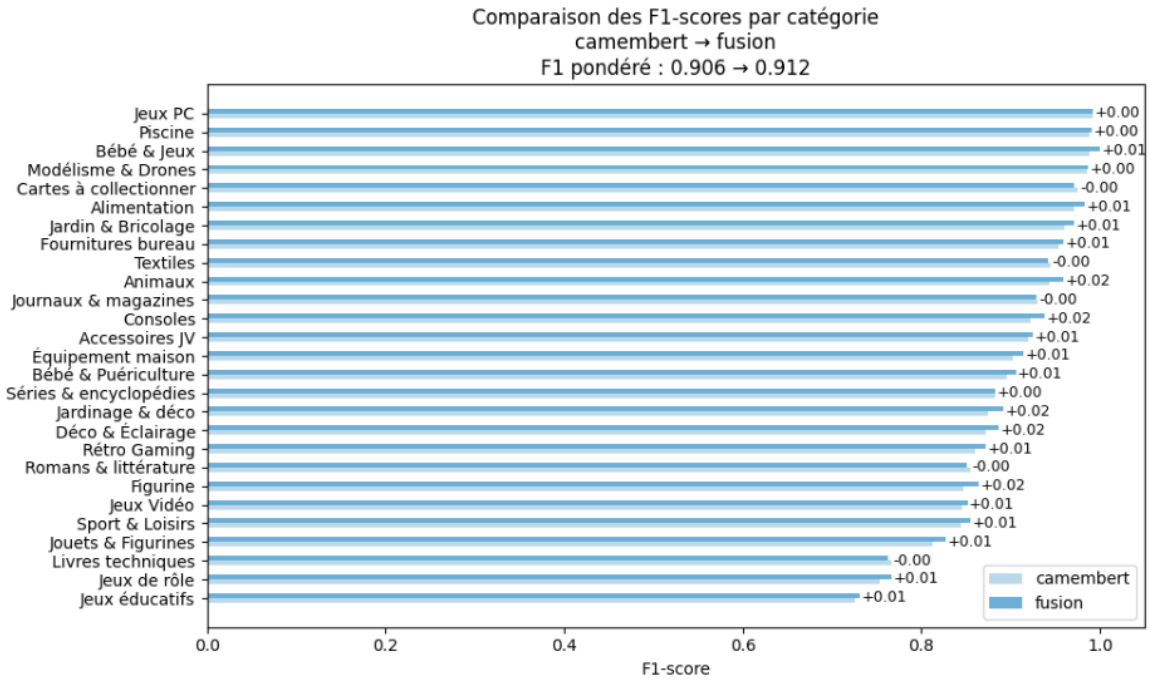
L'analyse par catégorie montre que la fusion bénéficie à une large majorité des classes. Des améliorations régulières, comprises entre +0,01 et +0,02 de F1-score, sont observées sur des segments intermédiaires tels que Jeux de rôle, Jouets & Figurines, Déco & Éclairage, Équipement maison, Rétro Gaming, Animaux ou Consoles, correspondant aux classes où la complémentarité des signaux textuels est la plus exploitable.

Quelques catégories très performantes initialement présentent des variations marginales, tandis que certaines classes aux frontières sémantiques floues, notamment Jeux éducatifs, Livres techniques, Romans & littérature, conservent des F1-scores plus faibles. Ces résultats reflètent principalement les ambiguïtés inhérentes aux descriptions produits, plutôt que des limitations du modèle lui-même.

L'analyse des confusions confirme cette interprétation. Les erreurs les plus fréquentes sont cohérentes sur le plan sémantique :

- Jeux éducatifs confondus avec Jouets & Figurines (14,5 % des erreurs de la classe) ;
- Romans & littérature et Livres techniques présentent des confusions croisées (11,1 % et 9,9 %) ;
- Les Jeux de rôle sont souvent confondus avec les Jeux éducatifs, reflétant des descriptions textuelles proches dans leur intention.

Ces confusions traduisent davantage des zones grises du problème de classification que des erreurs systématiques du modèle, et elles justifient pleinement l'introduction d'une fusion multimodale texte–image pour améliorer la discrimination dans ces segments.



III - Image

3.1. Exploration et Modèle de Référence

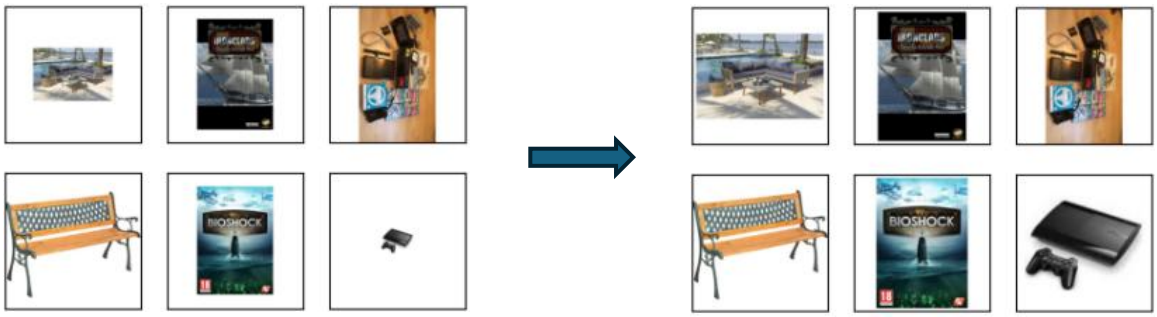
3.1.1 Objectifs et Positionnement de l'Analyse Image

Cette section ouvre la seconde partie du projet consacrée à la modalité image. L'objectif n'est pas uniquement d'évaluer la performance d'un modèle image, mais aussi de comprendre quelles caractéristiques visuelles sont réellement exploitées par le modèle pour discriminer les catégories produits.

L'analyse exploratoire met également en évidence la présence d'un volume significatif de doublons visuels, avec 8 956 images identifiées à l'aide d'un hachage perceptuel. Ces doublons n'ont toutefois pas été supprimés, dans la mesure où ils ne correspondent pas nécessairement à des doublons textuels et afin de préserver la cohérence et la stabilité des jeux d'entraînement, de validation et de test utilisés tout au long du projet.

3.1.2 Prétraitement et Focalisation sur l'Objet

Le prétraitement des images inclut une étape de recadrage automatique simple, visant à recentrer l'objet d'intérêt tout en réduisant l'influence du fond. Cette méthode repose sur l'hypothèse, largement vérifiée dans le jeu de données, d'un fond majoritairement uniforme, généralement blanc ou noir. Le recadrage est effectué à partir d'un seuillage global permettant d'identifier les zones de fond, puis de déterminer la région minimale englobant les pixels d'intérêt. Afin de préserver la cohérence géométrique des images, les proportions originales sont conservées, et une marge de sécurité est ajoutée autour de la zone recadrée. Cette approche, volontairement simple et déterministe, s'est révélée robuste et stable sur l'ensemble du jeu de données, et a été retenue comme solution de prétraitement dans le pipeline final.



Par ailleurs, une méthode de recadrage plus avancée, fondée sur la détection de contours, a également été explorée. Cette approche visait à isoler précisément l'objet d'intérêt en détectant ses contours principaux (via l'algorithme de Canny), puis à uniformiser le fond avant recadrage, tout en conservant les proportions de l'image. Bien que cette stratégie se soit révélée efficace sur certains exemples, elle a montré des limites de robustesse à l'échelle du jeu de données, notamment en raison d'une forte sensibilité aux seuils globaux et de difficultés à gérer des variations locales d'éclairage et de contraste. En l'absence de mécanismes de seuillage adaptatif pleinement satisfaisants dans ce cadre, cette approche n'a finalement pas été retenue dans le pipeline final, au profit d'une solution plus simple et plus stable.

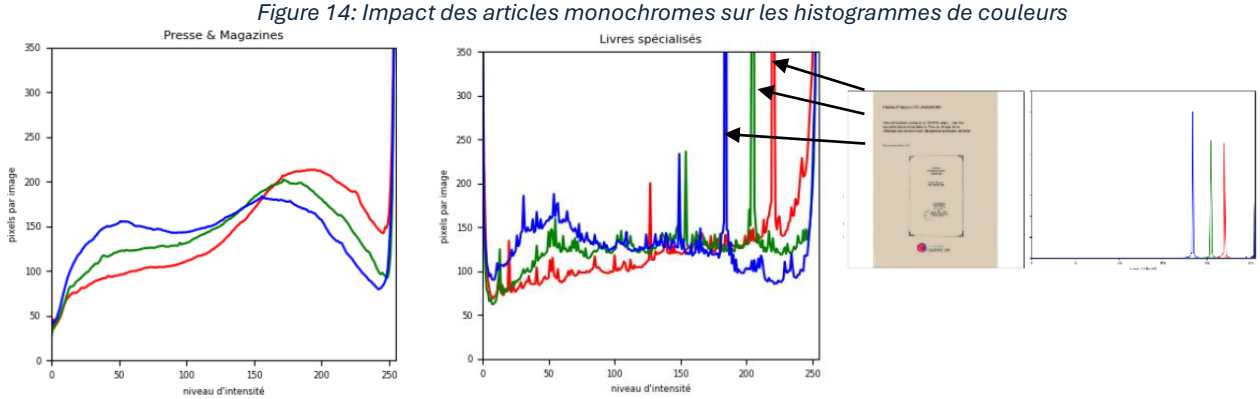
3.1.4 Exploration des couleurs

L'analyse exploratoire des distributions de couleurs met en évidence l'existence de signatures visuelles spécifiques à certaines catégories de produits. L'étude des histogrammes de couleurs révèle notamment une dominance des teintes rouges dans la majorité des catégories. À l'inverse, la catégorie Piscine se distingue par des intensités de bleu plus élevées, cohérentes avec la présence fréquente d'éléments aquatiques.

Par ailleurs, certaines catégories présentent des images globalement plus sombres que d'autres. C'est notamment le cas de la catégorie Jeux PC, dont les distributions de luminosité sont plus basses comparativement à des catégories comme Textile d'intérieur, caractérisées par des visuels plus clairs.

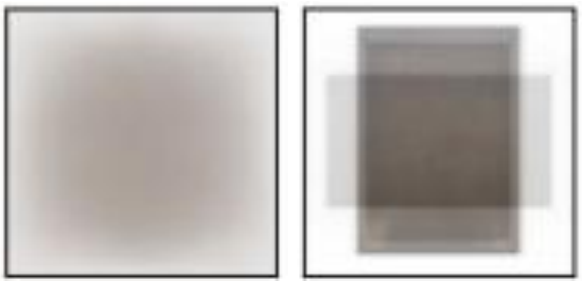
La forme des histogrammes varie également selon les catégories : certaines présentent des distributions lissées, traduisant une grande diversité chromatique, tandis que d'autres affichent des pics marqués à certaines intensités. Ces pics indiquent qu'une proportion importante de pixels partage une même valeur d'intensité, phénomène souvent lié à la présence d'objets quasi monochromes. Par exemple, dans la catégorie Livres techniques, ces pics correspondent fréquemment à des couvertures uniformes, comme des manuscrits à dominante chromatique unique, illustrés dans l'exemple ci-dessous.

Figure 14: Impact des articles monochromes sur les histogrammes de couleurs



L'analyse des images moyennes par catégorie, obtenues à partir des intensités moyennes des pixels, met en évidence des différences structurelles marquées entre les catégories. Pour certaines d'entre elles, des formes géométriques distinctes émergent clairement, tandis que pour d'autres, l'image moyenne apparaît plus diffuse et moins structurée.

Figure 15: Image des intensités moyennes des catégories décoration & luminaire et jeux PC



Ce phénomène est particulièrement prononcé pour les catégories Jeux de cartes, Jeux PC et Littérature, dans lesquelles des formes rectangulaires se détachent nettement. Ces observations suggèrent la présence récurrente d'objets aux contours réguliers, tels que des boîtes ou des couvertures, caractéristiques de ces catégories. Sur la base de cette analyse exploratoire, une feature dédiée a été construite, consistant à compter le nombre de rectangles (ou parallélogrammes) détectés dans chaque image, afin de capturer explicitement cette information structurelle dans le pipeline de modélisation.

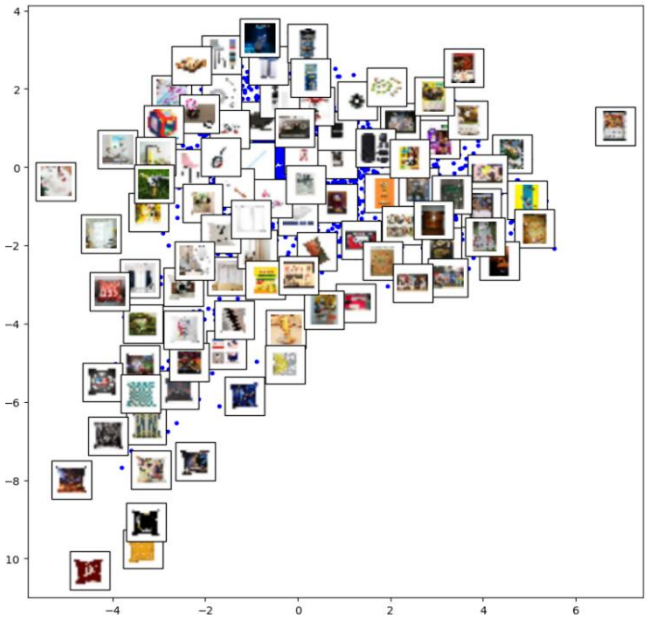
Figure 16: Détéction des parrallèlogrammes

3.1.6 Projection Supervisée et Séparabilité des Catégories

Afin d'analyser la séparabilité des classes, une projection supervisée de type LDA est réalisée. Cette méthode permet de visualiser la capacité des features image à discriminer les différentes catégories, en tenant explicitement compte des labels.

Comme l'illustre le graphique, la projection sur les deux premières composantes de la LDA met en évidence une séparation nette pour les Coussins et oreillers, clairement isolées dans la partie inférieure gauche du plan, ce qui permet d'identifier une grande partie des articles relevant du textile d'intérieur. Les catégories liées aux livres et revues quant à elles apparaissent regroupées sur la droite de la projection, mais restent fortement imbriquées entre elles. Ces articles étant présents dans plusieurs catégories du catalogue, cette projection ne permet pas de les distinguer finement les unes des autres.

Figure 17: projection sur le plan généré par les deux premières composantes de la LDA



3.1.7 Modèle image de référence

Afin d'évaluer la contribution de l'information visuelle seule à la tâche de classification, nous mettons en place un modèle image de référence, servant de baseline avant l'introduction de modèles plus complexes.

Les images font l'objet d'un recadrage automatique, puis plusieurs familles de descripteurs visuels sont ensuite extraites en parallèle afin de capturer différentes dimensions de l'information contenue dans les images :

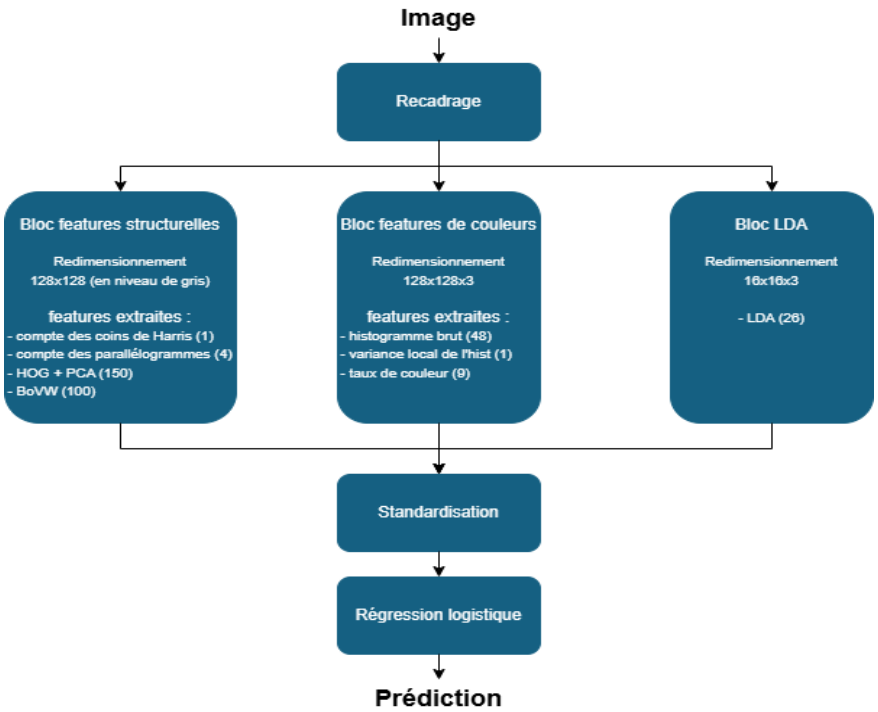
- Les descripteurs chromatiques reposent sur des histogrammes RGB ainsi que sur un encodage discret des couleurs dominantes, permettant de caractériser les principales signatures colorimétriques des produits.
- Les descripteurs structurels s'appuient sur les Histogrammes of Oriented Gradients (HOG), la détection de coins de Harris, le comptage de motifs géométriques, ainsi que sur des descripteurs SIFT, capables de capturer à la fois la structure locale et les motifs texturaux présents dans les images.
- En complément, une projection supervisée par Analyse Discriminante Linéaire (LDA) est appliquée en parallèle afin de produire une représentation projetée optimisant la séparabilité entre les classes.

L'ensemble des représentations obtenues est ensuite concaténé et normalisé afin de former une description visuelle dense et homogène. Sur cette représentation finale, un modèle de régression logistique est entraîné, choisi pour sa simplicité, sa rapidité d'apprentissage et son caractère interprétable.

Le modèle image de référence n'atteint toutefois qu'un F1-score pondéré de 37,3 %, traduisant des performances globalement limitées. Certaines catégories, telles que Animaux, Jeux éducatifs et Jeux de rôle, présentent des F1-scores inférieurs à 10 %, ce qui peut s'expliquer par une forte hétérogénéité visuelle des images associées.

À l'inverse, les meilleures performances sont observées pour les catégories liées aux textiles d'intérieur et aux jeux de cartes. Lors de l'analyse exploratoire, la projection par LDA avait déjà mis en évidence une bonne séparabilité visuelle des articles tels que les coussins et oreillers, tandis que les jeux de cartes présentent des formes et structures visuelles distinctives.

Figure 18: Pipeline modèle image baseline



3.1.8 Importance relative des features

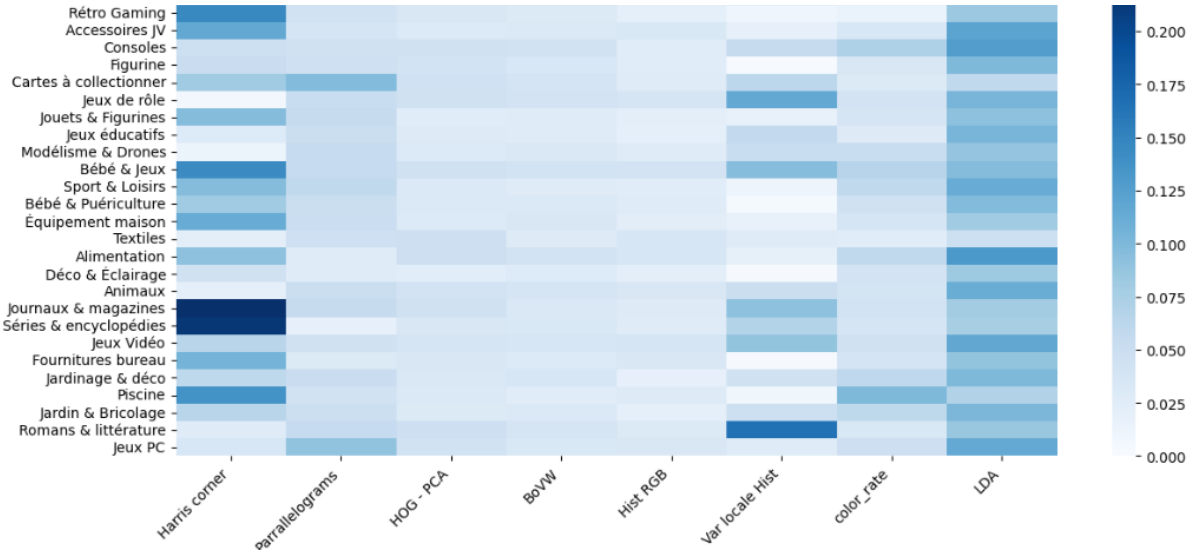
L'analyse des coefficients du modèle permet d'interpréter les performances observées par catégorie. Afin de faciliter l'analyse les coefficients ont été regroupés par familles de features, à l'aide de la moyenne de leurs valeurs absolues.

Les coins de Harris apparaissent comme fortement discriminants pour les catégories de supports imprimés (journaux, encyclopédies, livres techniques), cela pourrait être en raison de la densité élevée de structures anguleuses liées aux textes présents sur les couvertures.

La détection de parallélogrammes contribue fortement à l'identification des objets de format rectangulaire standardisés, tels que les jeux PC, les livres et les cartes à collectionner.

Les descripteurs chromatiques jouent un rôle central pour certaines classes spécifiques. La catégorie « Piscine » est ainsi fortement corrélée à une dominance du bleu. Les couvertures monochromes de certains romans semblent également jouer un rôle discriminant, comme en témoigne la valeur élevée de la feature var locale hist, sensible à la présence de pics marqués dans les histogrammes de couleur.

Figure 19: Matrice des valeur moyenne des coefficients absolues par bloc de features



3.2. Modélisation Image par Deep Learning

Cette section prolonge l'analyse visuelle amorcée avec les approches de Machine Learning classiques et vise à évaluer l'apport de modèles de Deep Learning capables d'apprendre automatiquement des représentations visuelles complexes de bout en bout. Contrairement aux méthodes précédentes fondées sur des descripteurs construits manuellement, ces architectures apprennent directement à partir des pixels, ouvrant la voie à une meilleure généralisation sur des catégories visuellement riches et hétérogènes.

3.2.1 Point de départ : LeNet, une référence minimale convolutive

LeNet-5 constitue le point d'entrée historique vers l'apprentissage profond. Bien que conçu à l'origine pour la reconnaissance de chiffres manuscrits en niveaux de gris, ce modèle a été adapté afin de servir de première borne de performance pour la classification d'images produits du jeu Rakuten.

L'architecture a été modernisée pour accepter des images RVB de petite résolution (32×32), permettant au modèle d'exploiter les signaux colorimétriques identifiés lors de la phase d'exploration visuelle. Un

mécanisme de dropout (0,5) a été introduit dans les couches entièrement connectées afin de limiter le surapprentissage, et l'entraînement a été réalisé sur GPU avec un taux d'apprentissage de 1e-3 pendant 30 époques, en utilisant l'Automatic Mixed Precision pour optimiser l'efficacité mémoire.

Les performances obtenues restent limitées, avec un meilleur F1 de validation de 0,2978 et une précision de validation d'environ 35 %. Néanmoins, ces résultats dépassent largement le hasard pur (≈3,7 %) et confirment que les réseaux convolutifs sont capables de capter des motifs locaux pertinents tels que contours et textures de base.

LeNet joue ainsi le rôle de borne inférieure : un modèle sensible mais à vision limitée, utile pour contextualiser les gains apportés par des architectures plus modernes.

3.2.2 Passage à un standard industriel : ResNet50 et apprentissage par transfert

ResNet50 marque une rupture majeure par rapport à LeNet grâce à l'introduction des connexions résiduelles, qui permettent d'entraîner des réseaux profonds sans dégradation du gradient. Le modèle est utilisé dans un cadre d'apprentissage par transfert, avec des poids pré-entraînés sur ImageNet, afin de transférer des connaissances visuelles génériques vers le contexte e-commerce.

Dès les premières époques, ResNet50 dépasse très largement les performances de LeNet : un F1 supérieur à 0,56 est atteint dès l'époque 3, soit presque le double du maximum obtenu par LeNet après 30 époques. Les performances finales atteignent environ 0,64 en F1 de validation et 67 % de précision.

Cependant, l'analyse de la dynamique d'apprentissage met en évidence un écart important entre les performances d'entraînement (quasi parfaites) et celles de validation. Ce comportement traduit une tendance marquée au surapprentissage, soulignant les limites des CNN profonds classiques face à la forte diversité visuelle du jeu de données.

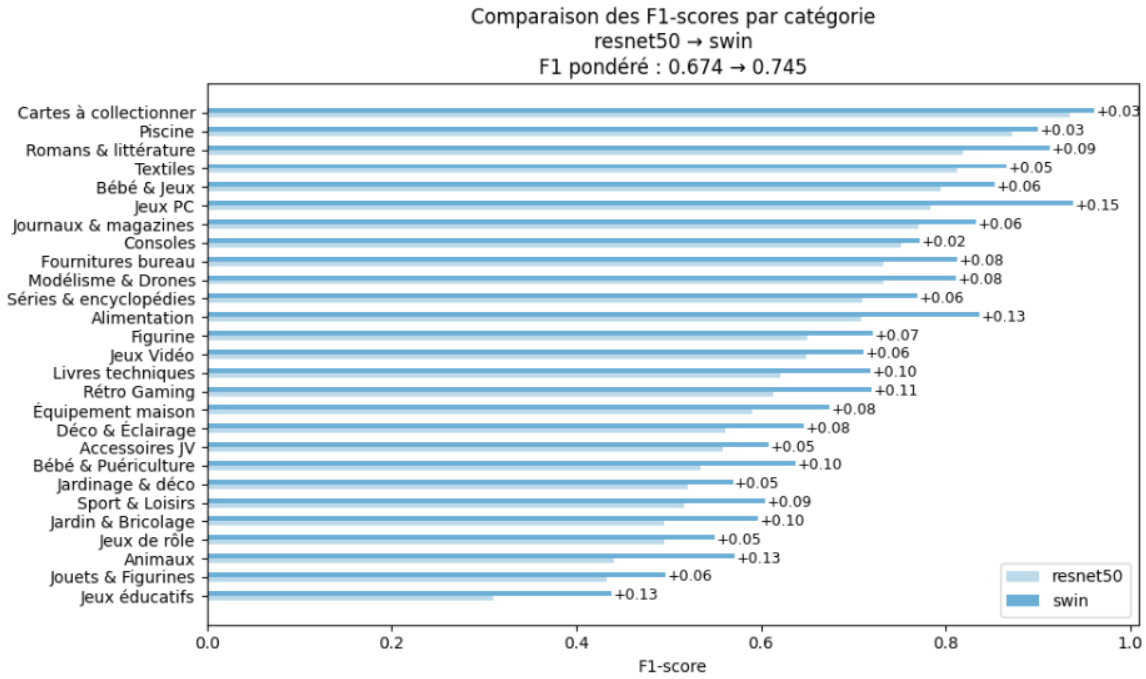
3.2.3 Changement de paradigme : Vision Transformers et Swin Transformer

Les Vision Transformers introduisent une rupture conceptuelle en traitant l'image comme une séquence de patchs, sur lesquels s'appliquent des mécanismes d'auto-attention. Cette approche permet une modélisation globale des relations spatiales, contrairement aux CNN qui reposent sur des réceptifs locaux.

Le Swin Transformer constitue une évolution clé de ce paradigme en combinant attention globale et structure hiérarchique par fenêtres glissantes. Cette architecture permet de capturer simultanément les silhouettes globales et les micro-détails (textures, étiquettes, zones textuelles).

L'entraînement repose sur des stratégies de régularisation avancées : Mixup, CutMix et Drop Path, destinées à renforcer la capacité de généralisation. Les résultats obtenus marquent un saut qualitatif majeur, avec un meilleur F1 de validation de 0,7311 et une précision de 75,45 %. Le modèle atteint des performances élevées dès les premières époques, indiquant une capacité d'apprentissage très rapide.

Les logits issus de Swin constituent ainsi un signal visuel de grande qualité, directement exploitable pour la fusion multimodale.



Le passage d'un CNN standard de type ResNet-50 vers un Transformer visuel de type Swin se traduit par une amélioration systématique des performances, avec des gains de F1-score positifs pour l'ensemble des catégories. Les progressions les plus marquées concernent des classes initialement difficiles pour les CNN, telles que Jeux PC (+0,15), Animaux (+0,13), Jeux éducatifs (+0,13) ou Alimentation (+0,13). Ces catégories présentent souvent une forte variabilité visuelle ou des structures globales complexes, pour lesquelles la capacité des Transformers à modéliser des dépendances spatiales à différentes échelles constitue un avantage déterminant.

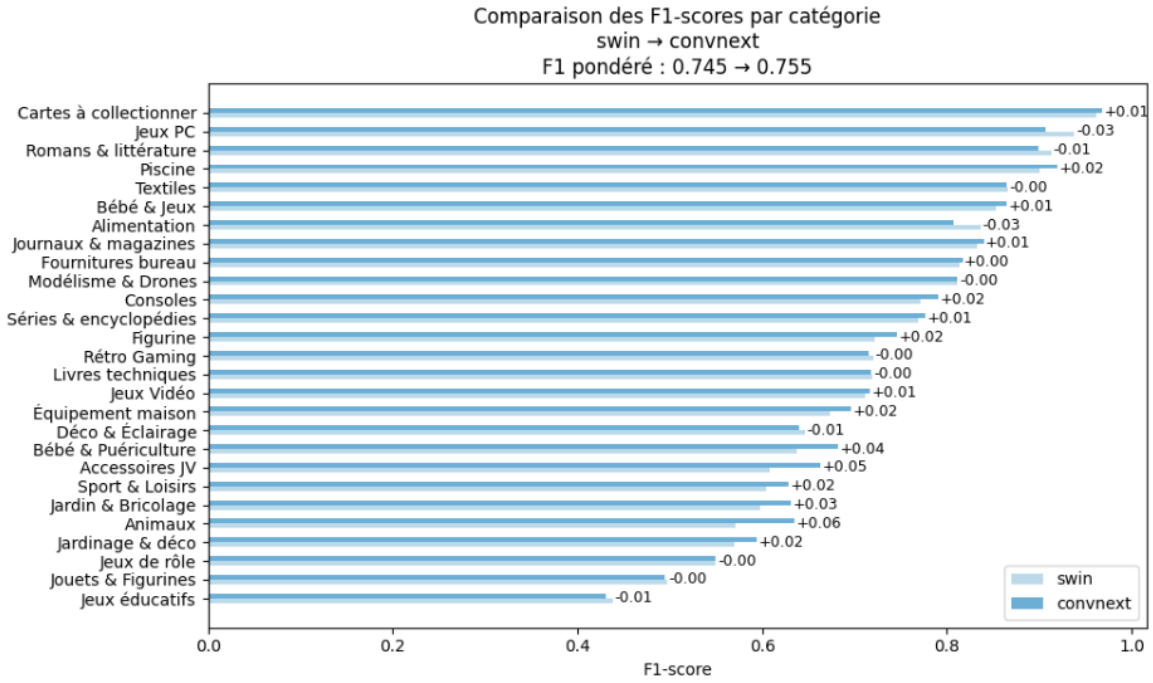
À l'inverse, les gains sont plus modérés pour des catégories déjà bien séparées par le modèle de référence, comme Cartes à collectionner, Piscine ou Consoles. Ces classes bénéficiaient déjà de motifs visuels très distinctifs (formes, couleurs dominantes), laissant peu de marge d'amélioration. Globalement, ces résultats confirment que les Transformers visuels, grâce à leur attention hiérarchique et à leur meilleure prise en compte du contexte global, permettent une représentation plus riche et plus robuste que les CNN classiques, en particulier pour les catégories hétérogènes ou ambiguës.

3.2.4 Renaissance de la convolution : ConvNeXt

ConvNeXt démontre qu'un réseau purement convolutif, modernisé par les principes issus des Transformers, peut rivaliser avec les architectures à attention. Le modèle exploite une résolution élevée (384×384), une régularisation agressive via Drop Path et une Exponential Moving Average (EMA) des poids afin de stabiliser l'apprentissage.

Ces choix permettent de corriger les limitations observées sur ResNet50, en réduisant le surapprentissage tout en conservant une forte capacité expressive.

ConvNeXt atteint un meilleur F1 de validation de 0,7395 et une précision EMA de 76,24 %, en faisant le modèle convolutif le plus robuste évalué. Sa complémentarité avec Swin Transformer constitue le socle de la composante image pour la fusion multimodale finale.



Globalement, ConvNeXt surpasse légèrement Swin Transformer, avec un F1-score pondéré de 0.755 contre 0.745. Ce gain reste modéré mais cohérent, et s'explique par des améliorations localisées sur plusieurs catégories spécifiques.

Les gains les plus marqués en faveur de ConvNeXt sont observés sur les catégories Animaux (+0.06), Accessoires JV (+0.05), Bébé & Puériculture (+0.04) et Jardin & Bricolage (+0.03). Ces catégories présentent des images visuellement très diverses (multiplicité des objets, variations de contexte, d'échelle et de cadrage), suggérant que ConvNeXt est plus robuste à cette hétérogénéité visuelle et parvient à mieux généraliser à partir de motifs locaux récurrents, malgré une forte variabilité globale.

À l'inverse, Swin Transformer conserve un léger avantage ou une stabilité sur certaines catégories telles que Jeux éducatifs, Déco & Éclairage, Alimentation, Jeux PC ou Romans & littérature, où les écarts restent faibles mais parfois défavorables à ConvNeXt.

3.2.5 Synthèse et perspectives pour la fusion multimodale

Cette étude met en évidence une progression claire des performances à mesure que les architectures gagnent en capacité et en sophistication. LeNet fournit une borne minimale, ResNet50 introduit la profondeur et le transfert de connaissances, tandis que Swin Transformer et ConvNeXt atteignent un niveau de généralisation compatible avec une exploitation en production.

Les résultats soulignent également la complémentarité entre architectures à attention et convolutives modernes. Les différences de comportement observées entre Swin Transformer et ConvNeXt, bien que modérées, suggèrent une complémentarité potentielle entre ces deux architectures.

3.3. Fusion des modèles image

3.3.1 Objectifs et positionnement méthodologique

Cette section correspond à l'étape finale de la modélisation image et vise à exploiter la complémentarité des différents modèles visuels étudiés précédemment. Après l'évaluation séparée de modèles convolutifs modernes (ConvNeXt), de Transformers visuels (Swin Transformer).

Comme pour la modalité texte, les sorties des modèles sont calibrées puis combinées à l'aide de stratégies de fusion tardive, selon un protocole strict évitant toute fuite d'information. Deux approches sont comparées : un blending pondéré et un stacking, reposant sur un classifieur de méta-niveau entraîné à partir des probabilités des modèles de base.

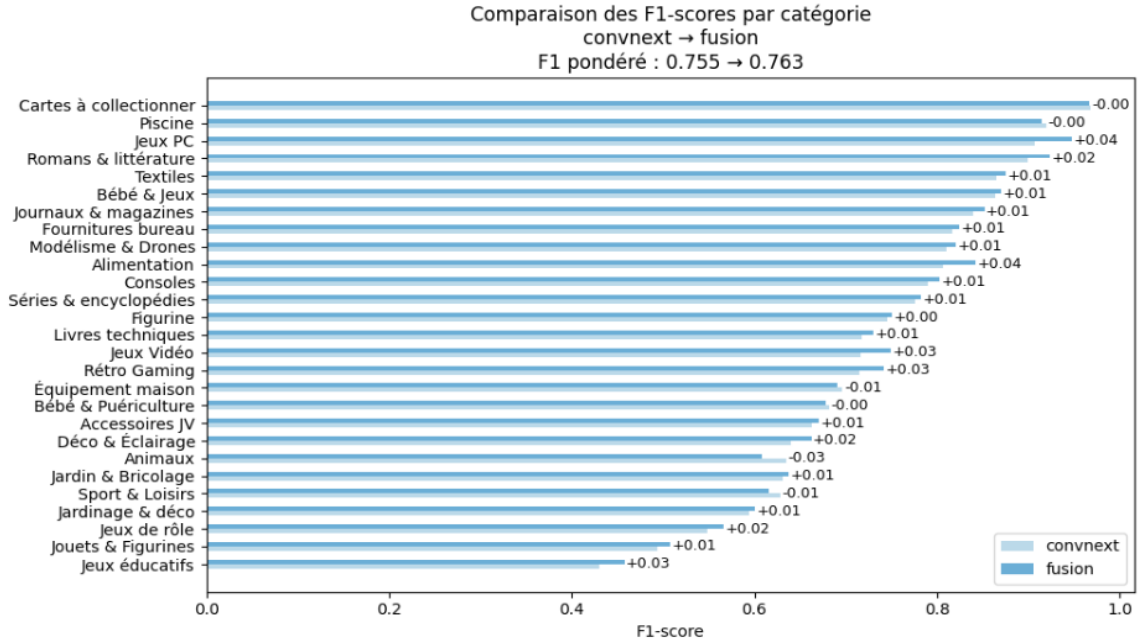
3.3.2 Analyse des performances et résultats

Les stratégies de blending et de stacking atteignent des performances similaires en termes de F1-score, mais le blending présente un log loss nettement plus faible. Ce résultat suggère une meilleure qualité des probabilités produites, avec des prédictions correctes associées à des niveaux de confiance plus élevés et des erreurs moins fortement confiantes.

La fusion retenue est donc le blending pondéré, pour lequel les poids estimés — définis proportionnellement à l'inverse de la log loss de chaque modèle — sont quasi équivalents, proches de 0,5. Cette répartition équilibrée confirme la complémentarité des architectures ConvNeXt et Swin Transformer.

Évaluée sur le jeu de test, cette stratégie de fusion permet d'améliorer systématiquement les performances par rapport aux modèles pris individuellement. Le F1-score pondéré progresse ainsi d'environ 0,755 pour ConvNeXt et 0,745 pour Swin Transformer, pour atteindre 0,763 avec la meilleure configuration fusionnée.

Les gains sont particulièrement visibles sur certaines catégories comme Jeux PC ou Accessoires JV, pour lesquelles les améliorations atteignent +0,04 à +0,06 en F1-score. À l'inverse, les catégories visuellement très hétérogènes, telles que Jeux éducatifs ou Jouets & Figurines, restent difficiles à discriminer malgré la fusion.



IV - Fusion Multimodale Texte–Image

4.1. Construction et Sélection du Modèle Multimodal

4.1.1 Objectifs et Positionnement Méthodologique

Cette section constitue l'aboutissement du projet et vise à dépasser les limites observées dans les approches monomodales. Malgré les performances élevées du modèle texte final (F1 = 0,912), certaines confusions persistent dans des zones sémantiques ambiguës, notamment lorsque les descriptions textuelles sont pauvres. L'objectif de la fusion multimodale est donc d'introduire l'information visuelle comme un signal correctif, capable de lever ces ambiguïtés sans dégrader les catégories déjà bien maîtrisées par le texte.

4.1.2 Modèles d'Entrée et Hypothèse de Complémentarité

Les stratégies de fusion retenues sont identiques à celles explorées précédemment pour les fusions intra-modalité : un blending pondéré et un stacking basé sur une régression logistique. Les entrées de ces modèles correspondent aux probabilités de prédiction issues :

- du modèle texte, résultant d'un triple blending entre CamemBERT, XLM-R et d'un modèle TF-IDF ;
- du modèle image basé sur sur un blending entre ConvNeXt et Swin Transformer.

L'hypothèse centrale est que les erreurs commises par ces deux modalités ne sont pas fortement corrélées : l'image est supposée apporter un signal discriminant précisément là où le texte atteint un plafond sémantique.

4.1.3 Résultats et Choix du Modèle Final

Figure 20: comparaison blending / stacking multimodal

	f1_weighted	log_loss
blending *	0,922	0,360
stacking	0,922	0,297

** avec poids du texte d'environ 0,73*

Les deux stratégies atteignent des performances proches en F1 pondéré (≈ 0,922), confirmant l'apport global de la multimodalité. Toutefois, la stratégie de stacking se distingue nettement par une log-loss significativement plus faible (0,297 contre 0,36 pour le blending).

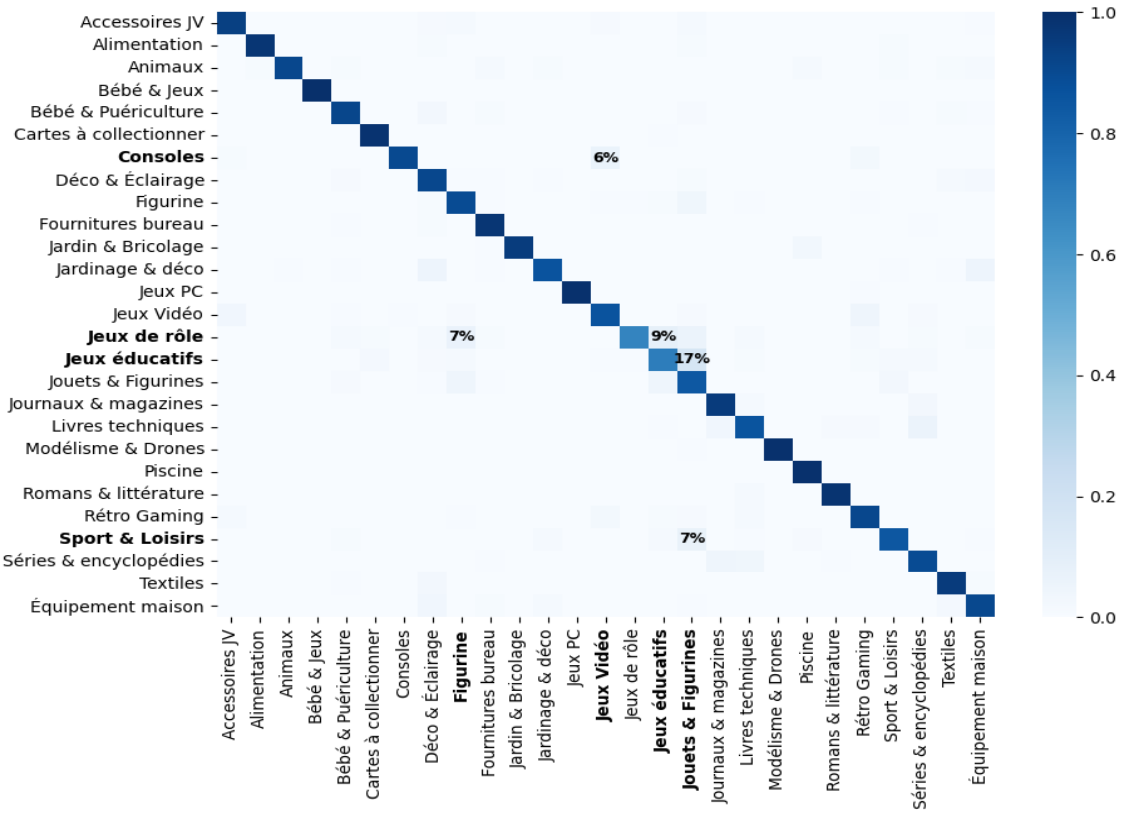
Une log-loss plus faible traduit des probabilités mieux calibrées, le modèle attribue des probabilités plus élevées aux prédictions correctes et limite la sur-confiance sur les erreurs. Cette propriété favorise une meilleure robustesse face à de nouvelles données, ce qui motive le choix du stacking multimodal comme modèle final.

4.2. Analyse Fine de la Fusion Multimodale

4.2.1 Apport Visuel

Sur le jeu de test final, le modèle multimodal empilé atteint un F1 pondéré de 92,6%, soit un gain absolu de +1,4% par rapport au meilleur modèle texte seul.

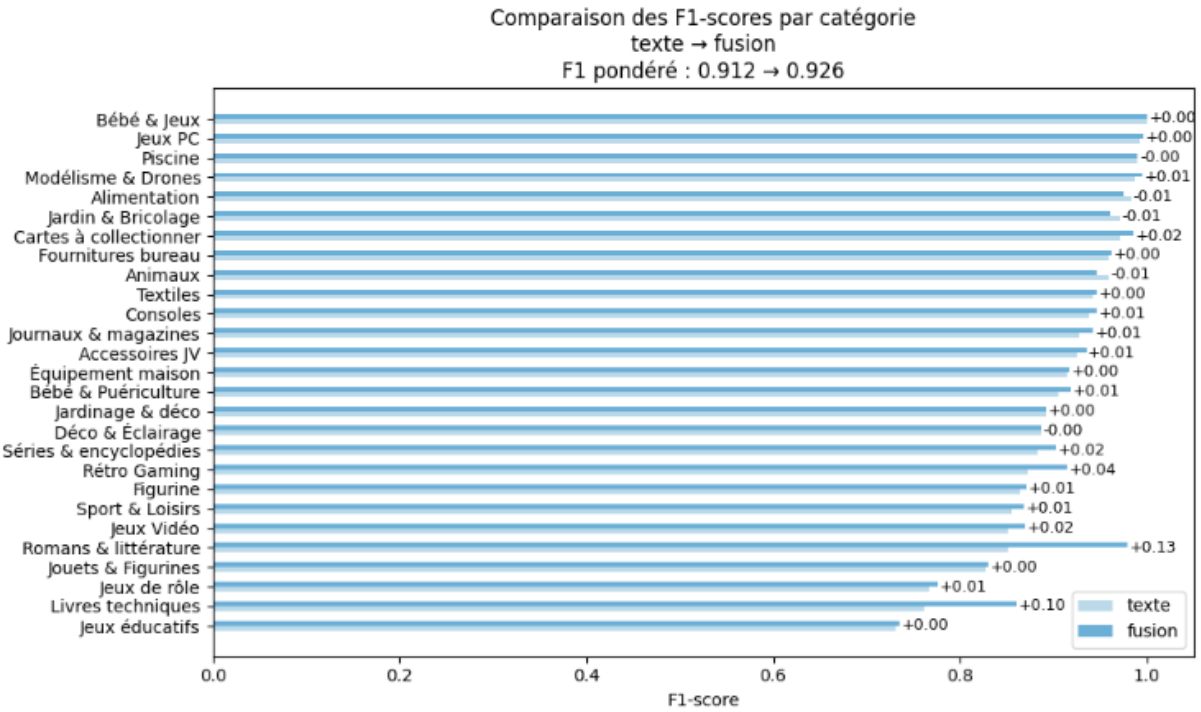
Figure 21 : Matrice de confusion du modèle final



L'analyse des confusions montre que les catégories Jeux de rôle, Jeux éducatifs et Jouets et figurines restent fortement confondues. En revanche, on observe une nette amélioration concernant les confusions entre les catégories « Romans & littérature » et « Livres techniques et éducatifs », drastiquement réduites : elles passent de plus de 9 % dans le modèle texte seul à moins de 2 % après fusion. Cette baisse confirme que les indices visuels (couvertures, mise en page, iconographie) jouent un rôle déterminant pour différencier des descriptions linguistiquement proches.

Figure 22 : confusions s'améliorant le plus (% des prédictions de la classe réelle)

Confusion (classe réelle → classe prédite)	Modèle texte	Modèle final
Romans & littérature → Livres techniques	11.11	1.69
Livres techniques → Romans & littérature	9.85	1.50



Le F1-score de ces deux catégories se trouve logiquement fortement amélioré (+13 % pour « Romans & littérature » et + 10 % pour « Livres techniques et éducatifs »). Seules les catégories Alimentation, Jardin & Bricolage et Animaux présentent une très légère dégradation par rapport au modèle texte seul (de l'ordre de 1 % de F1-score), ce qui indique que la fusion multimodale agit comme un mécanisme correctif sans effet déstabilisant.

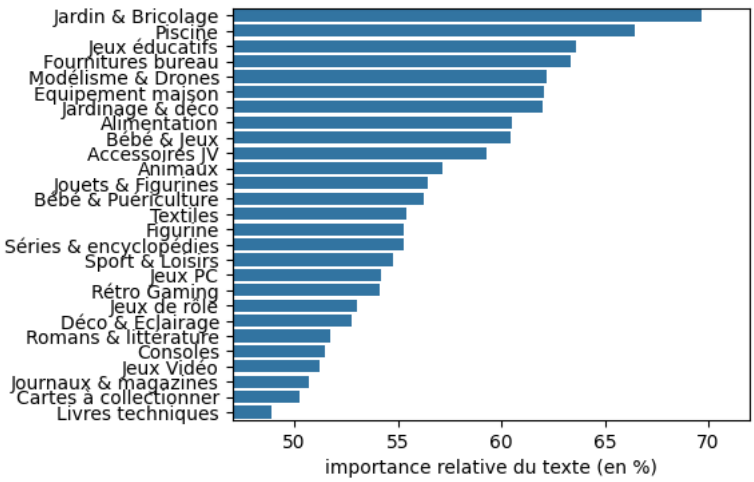
4.2.2 Importance Relative des Modalités et Limites

L'inspection des coefficients du méta-modèle permet de quantifier l'importance relative des modalités. En moyenne, le texte représente 57 % de la décision finale contre 43 % pour l'image, confirmant le rôle dominant du texte tout en soulignant l'apport substantiel du visuel.

Cette importance varie toutefois fortement selon les catégories. De nombreuses classes restent majoritairement texte-dépendantes (par exemple Jardin & Bricolage, Équipement maison ou Piscine), tandis que d'autres présentent une contribution plus équilibrée entre les deux modalités, traduisant une complémentarité effective.

Un point particulièrement intéressant concerne les catégories Livres techniques et Romans & littérature, qui figuraient parmi les plus fortement confondues avec le modèle texte seul. Pour ces deux classes, l'importance de l'image est comparable, voire supérieure à celle du texte, ce qui est cohérent avec la forte réduction des confusions croisées observée précédemment.

Figure 23: Importance du texte dans le méta-modèle



4.3. Préparation du modèle final

4.3.1 Maximisation des données

Afin d'améliorer la capacité de généralisation du modèle final, une stratégie de réentraînement exploitant l'ensemble des données disponibles a été mise en œuvre.

Dans un premier temps, les paramètres de fusion — températures de calibration, poids de blending et coefficients du méta-modèle — sont estimés à partir d'un ensemble combinant le jeu de validation et le jeu de test interne. Cette approche permet de maximiser le volume de données supervisées.

Une fois ces paramètres estimés, ils sont figés et sauvegardés dans un fichier de configuration afin de garantir la stabilité de la logique de fusion. Les modèles sont ensuite réentraînés sur l'ensemble du corpus d'apprentissage (train, validation et test interne).

Le modèle de machine learning classique (TF-IDF et SVM) est entièrement réentraîné sur ce corpus étendu. En revanche, en raison de contraintes de temps et de ressources computationnelles, parmi les quatre modèles de deep learning utilisés, seul le modèle CamemBERT a été réentraîné. Ce réentraînement a été réalisé avec 10 % des données utilisé comme jeu de supervision

Cette stratégie permet de tirer pleinement parti des données disponibles. Dans ce cadre, une soumission finale a été effectuée sur la plateforme du challenge. Le score obtenu sur le jeu de test externe atteint 92,73 % de F1 pondéré, contre 92,6 % sur le test interne, confirmant la bonne capacité de généralisation du modèle.

4.3.2 Pipeline final

À partir des paramètres figés et des modèles enregistrés, un pipeline d'inférence final est construit. Celui-ci repose sur l'assemblage des pipelines texte et image, chacun intégrant ses modèles respectifs, les poids de blending associés et les températures de calibration correspondantes.

Le pipeline texte produit des probabilités conditionnelles à partir des entrées textuelles, tandis que le pipeline image génère des probabilités à partir des chemins des images associées aux produits. Ces sorties sont ensuite combinées par le méta-modèle de fusion afin de produire la prédiction finale.

Le pipeline global est initialisé à partir du chemin du dossier contenant les images. Une fois initialisé, l'appel à la méthode predict_proba permet d'obtenir les prédictions de catégories (prdtypecode) à partir d'un DataFrame d'entrée contenant les quatre champs de base : designation, description, imageid et productid. Cette abstraction garantit une utilisation simple et reproductible modèle final.

Figure 24: Utilisation du modèle

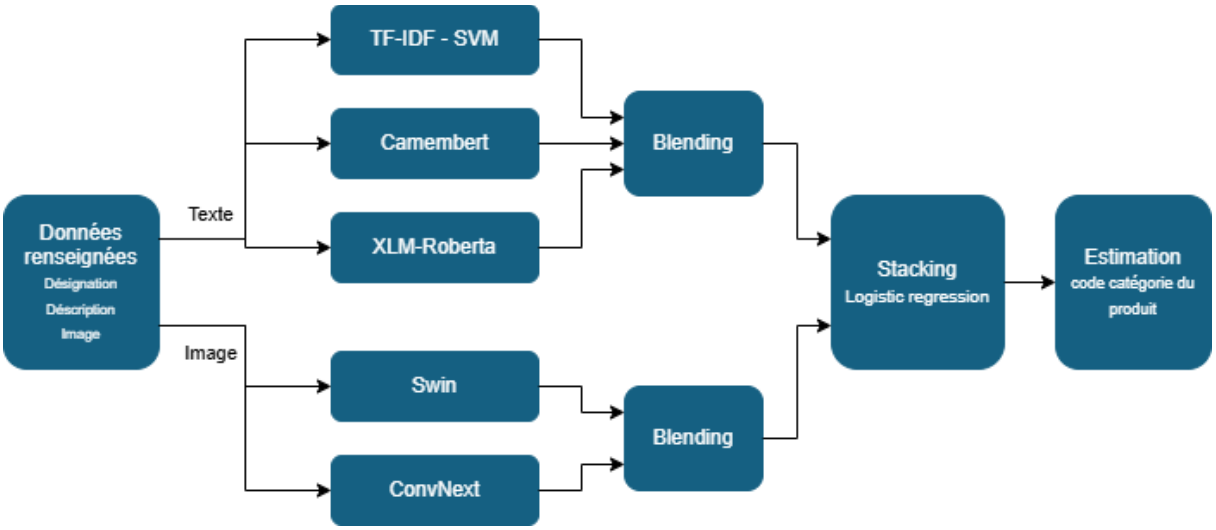
```
from pipeline.multimodal import FinalPipeline

df = pd.read_csv('../data/raw/X_test_update.csv')
model = FinalPipeline('../data/raw/images/image_test')
df.iloc[:5]
```

	Unnamed: 0	designation	description	productid	imageid
0	84916	Folkmanis Puppets - 2732 - Marionnette Et Théâ...	NaN	516376098	1019294171
1	84917	Porte Flamme Gaxix - Flamebringer Gaxix - 136/...	NaN	133389013	1274228667
2	84918	Pompe de filtration Speck Badu 95	NaN	4128438366	1295960357
3	84919	Robot de piscine électrique	<p>Ce robot de piscine d'un design innovan...	3929899732	1265224052
4	84920	Hsm Destructeur Securio C16 Coupe Crois¿E: 4 X...	NaN	152993898	940543690

```
model.predict_labels(df.iloc[:5])
```

```
array([1280, 1160, 2583, 2583, 2522], dtype=int64)
```

Figure 25: Pipeline du modèle



V - Conclusion générale

Ce projet a porté sur la classification automatique de produits e-commerce à partir de données textuelles et visuelles, dans le cadre du challenge Rakuten. L'objectif principal était d'analyser la contribution de chaque modalité, d'en identifier les limites respectives, puis de concevoir une stratégie de fusion capable d'exploiter leur complémentarité.

Les premières expérimentations ont montré que le texte constitue le signal le plus informatif pour une grande partie des catégories, en particulier lorsque les descriptions sont riches et structurées. Les modèles de type Transformer, et notamment CamemBERT, ont permis d'atteindre des performances élevées, tout en mettant en évidence l'importance d'un prétraitement minimaliste et ciblé. La combinaison de CamemBERT avec XLM-R et des signaux lexicaux classiques a permis d'atteindre un niveau de performance général élevé, tout en produisant des probabilités mieux calibrées. Toutefois, certaines confusions persistantes ont révélé les limites d'une approche textuelle seule.

L'analyse du signal visuel a montré que l'information image constitue un apport pertinent, mais partiellement limité lorsqu'elle est exploitée seule. Les modèles de deep learning modernes, en particulier ConvNeXt et Swin Transformer, surpassent nettement les architectures convolutionnelles plus classiques telles que ResNet50, mais restent confrontés à des catégories visuellement hétérogènes. La fusion des modèles image a néanmoins permis d'exploiter la complémentarité entre architectures convolutionnelles modernes et mécanismes d'attention, aboutissant à un modèle visuel plus robuste et cohérent.

La fusion multimodale constitue l'aboutissement du projet. En combinant les probabilités issues des pipelines texte et image au moyen d'une stratégie de stacking calibrée, le modèle final parvient à corriger des ambiguïtés textuelles sans dégrader les catégories déjà bien maîtrisées. Les gains observés sont ciblés mais significatifs, notamment pour les catégories de livres, confirmant la pertinence d'une approche multimodale raisonnée plutôt qu'une simple addition de signaux. L'analyse des coefficients du méta-modèle met en évidence un équilibre clair entre les modalités, avec une contribution moyenne de 56 % pour le texte et de 44 % pour l'image, confirmant le rôle correctif et complémentaire du signal visuel.

VI - Perspectives et pistes d'amélioration

Malgré les améliorations apportées par la fusion multimodale, certaines catégories restent plus difficiles à distinguer, en particulier *Jeux éducatifs*, *Jeux de rôle* et *Jouets & Figurines*. Ces classes présentent des frontières sémantiques et visuelles intrinsèquement floues, ce qui limite la capacité du modèle à les séparer de manière fiable à partir des seules informations textuelles et visuelles actuellement disponibles.

Un premier axe d'amélioration consisterait à enrichir les signaux textuels en intégrant davantage d'informations structurées lorsque celles-ci sont disponibles, telles que l'âge recommandé, le nombre de joueurs, la durée de jeu ou le type d'apprentissage visé.

Dans ce projet, une première étape a déjà été engagée dans cette direction à travers l'introduction de Numeric Tokens lors de l'entraînement de CamemBERT, permettant de discrétiser certaines grandeurs numériques afin de préserver leur valeur informative tout en limitant le bruit lexical.

Toutefois, cette stratégie est restée volontairement générique, avec un nombre limité de schémas de discrétisation. Une exploration plus approfondie et plus ciblée des découpages numériques, éventuellement adaptée à des attributs spécifiques ou à des sous-familles de produits, pourrait renforcer la capacité du modèle à distinguer des catégories proches telles que les jeux éducatifs, les jeux de rôle et les jouets ou figurines.

Une autre perspective d'amélioration serait l'adoption d'une classification hiérarchique, distinguant dans un premier temps les grandes familles de produits (jeux, livres, équipements), avant d'affiner la prédiction au sein de chaque famille. Une telle approche permettrait de réduire les confusions locales entre classes proches en introduisant une structure explicite dans le processus de décision.

Dans une classification multi-classes directe, l'apprentissage d'une classe difficile comme *Jeux éducatifs* consiste à la séparer simultanément de l'ensemble des autres catégories. Cette optimisation globale favorise les séparations faciles vis-à-vis de classes éloignées, au détriment d'une discrimination fine vis-à-vis des classes réellement proches telles que *Jeux de rôle* ou *Jouets & Figurines*.

Une approche hiérarchique permet de restreindre l'espace de décision à ces seules classes ambiguës, conduisant à une optimisation plus ciblée et plus efficace des frontières de décision.

Une alternative pragmatique à une classification hiérarchique complète consisterait à conserver le modèle actuel, puis à introduire un modèle spécialisé activé conditionnellement. Concrètement, lorsque la prédiction du modèle principal appartient à l'ensemble {Jeux éducatifs, Jeux de rôle, Jouets & Figurines}, un classifieur secondaire, entraîné exclusivement sur ces catégories, serait utilisé pour raffiner la décision.

Cette stratégie permet de concentrer la capacité de modélisation sur les zones de confusion réelles, sans dégrader les performances globales du système. Elle constitue un compromis efficace entre robustesse globale et optimisation locale.

En conclusion, ce travail met en évidence l'intérêt d'une approche progressive, interprétable et rigoureusement évaluée pour la classification de produits à grande échelle. La solution multimodale proposée constitue une base solide et cohérente, et ouvre des perspectives naturelles d'amélioration, notamment par l'intégration de métadonnées supplémentaires et de structures de décision plus hiérarchisées.

Disponibilité du code et de la documentation

L'ensemble du code source, des notebooks d'expérimentation ainsi que la documentation associée au projet sont disponibles sur le dépôt GitHub suivant :

https://github.com/beautiful-pixel/DS_rakuten

Le dépôt contient notamment :

- les scripts de prétraitement des données,
- les notebooks d'exploration et d'entraînement,
- les pipelines finaux de prédiction.

La documentation du code source est disponible ici :

https://beautiful-pixel.github.io/DS_rakuten/

Le challenge lié au projet :

https://challengedata.ens.fr/participants/challenges/35/

Table des figures