Note méthodologique : preuve de concept

I- Dataset retenu

Nous utilisons le dataset du **P6** Classification automatique des produits en catégories en utilisant le texte (**sentence_lem**) et la catégorie cible (**Categ_A**).

📊 Statistiques générales

Nombre total d'échantillons : **1050**
Variables :

- **sentence_lem** : Texte nettoyé décrivant le produit
- **Categ_A** : Catégorie attribuée au produit
- **Autres colonnes utiles** : product_name, description, brand, retail_price, discounted_price, etc.

📌 Qualité des données

- **Valeurs manquantes** : sentence_lem (5 valeurs), Categ_A (aucune).
- **Classes cibles** : Les catégories de produits.

🛠️ Utilisation pour le Benchmark BERT vs ModernBERT

- **Entrée modèle** : ["sentence_lem"]
- **Label à prédire** : ["Categ_A"]
- **Prétraitement** : Nettoyage des données et encodage des catégories avant fine-tuning des modèles.

II- Les concepts de ModernBERT

1. Introduction

ModernBERT est une version optimisée de **BERT** (Bidirectional Encoder Representations from Transformers), conçue pour améliorer l'efficacité et la rapidité des modèles de traitement du langage naturel (**NLP**) tout en conservant des performances compétitives sur diverses tâches.
Il repose sur le mécanisme des transformers, une architecture qui a révolutionné le NLP en capturant efficacement les dépendances contextuelles entre les mots.

2. Architecture et Optimisations

C'est un modèle conçu pour remplacer de manière transparente toute architecture de type BERT(**110M ou 340M** de paramètres). Il existe en deux configurations : un **modèle de base** avec **139 millions** de paramètres et un **grand modèle** avec **395 millions** de paramètres.

🔹 2.1. Structure générale

ModernBERT conserve la structure de BERT classique, avec plusieurs couches d'encodeurs basés sur l'auto-attention multi-tête et les feed-forward networks. Cependant, il apporte plusieurs modifications pour améliorer son efficacité.

🔹 2.2. Principales améliorations

1️⃣ Réduction du nombre de paramètres

- Contrairement à BERT, il utilise des variantes plus légères (comme DistilBERT, TinyBERT, ALBERT).
- Compression du modèle via distillation de connaissances (knowledge distillation).

2️⃣ Optimisation de l'attention

- Remplacement de l'auto-attention classique par des versions plus rapides
- Réduction du coût en calcul de $O(n^2)$ à $O(n \log n)$ ou $O(n)$ selon les implémentations.

3️⃣ Factorisation et décomposition des matrices

- Techniques comme la factorisation des embeddings pour limiter l'explosion du nombre de paramètres.

4️⃣ Pré-entraînement plus efficace

- Utilisation de masquage dynamique au lieu du masquage statique de BERT.
- Génération de faux tokens pour entraîner le modèle à détecter des erreurs, ce qui réduit le besoin de données et accélère l'apprentissage.

III- La modélisation

🔹 Prétraitement des Données

- **Nettoyage** : Suppression des valeurs manquantes, normalisation du texte (minuscule, suppression des caractères spéciaux).
- **Encodage des catégories** : Conversion des labels "**Categ_A**" en indices numériques.
- **Tokenisation** : Utilisation des tokenizers propres à chaque modèle.

🔹 Modèles Utilisés

- **BERT** (`'bert-base-uncased'`): Modèle de référence, puissant mais coûteux en calcul.
- **ModernBERT** (`"answerdotai/ModernBERT-base"`) : version plus rapide et plus légère.

🔹 Métrique retenue :

✅ **Accuracy** : Indicateur global.
✅ **Matrice de confusion** : Analyse fine des erreurs de classification.

🔹 Optimisation des Hyperparamètres

Les hyperparamètres suivants sont ajustés avec **GridSearchCV** :

- **Learning rate** : Test de valeurs entre 1e-5 et 5e-5.
- **Batch size** : Comparaison entre 8, 16 et 32 pour un bon compromis entre convergence et temps d'entraînement.
- **Nombre d'époques** : Entre 3 et 5, avec **early stopping** pour éviter l'**overfitting**.

IV- Une synthèse des résultats

1- BERT

1-1- Rapport de classification

Rapport de classification :

	precision	recall	f1-score	support
Baby Care	0.75	0.67	0.71	27
Beauty and Personal Care	0.80	0.95	0.87	21
Computers	0.94	0.87	0.90	38
Home Decor & Festive Needs	0.82	0.93	0.88	30
Home Furnishing	0.83	0.86	0.85	35
Kitchen & Dining	0.90	0.73	0.81	26
Watches	0.94	1.00	0.97	33
accuracy			0.86	210
macro avg	0.86	0.86	0.85	210
weighted avg	0.86	0.86	0.86	210

Précision globale : 0.86

Les résultats obtenus avec **BERT montrent une bonne performance globale (accuracy de 86 %** et **F1-score macro de 85 %)**, avec des variations selon les catégories : certaines classes comme **"Watches"** et **"Computers" sont très bien reconnues (F1-score > 90 %)**, tandis que d'autres, comme **"Baby Care"** et **"Kitchen & Dining", ont un rappel plus faible**, indiquant une certaine difficulté à identifier correctement tous les exemples de ces catégories.

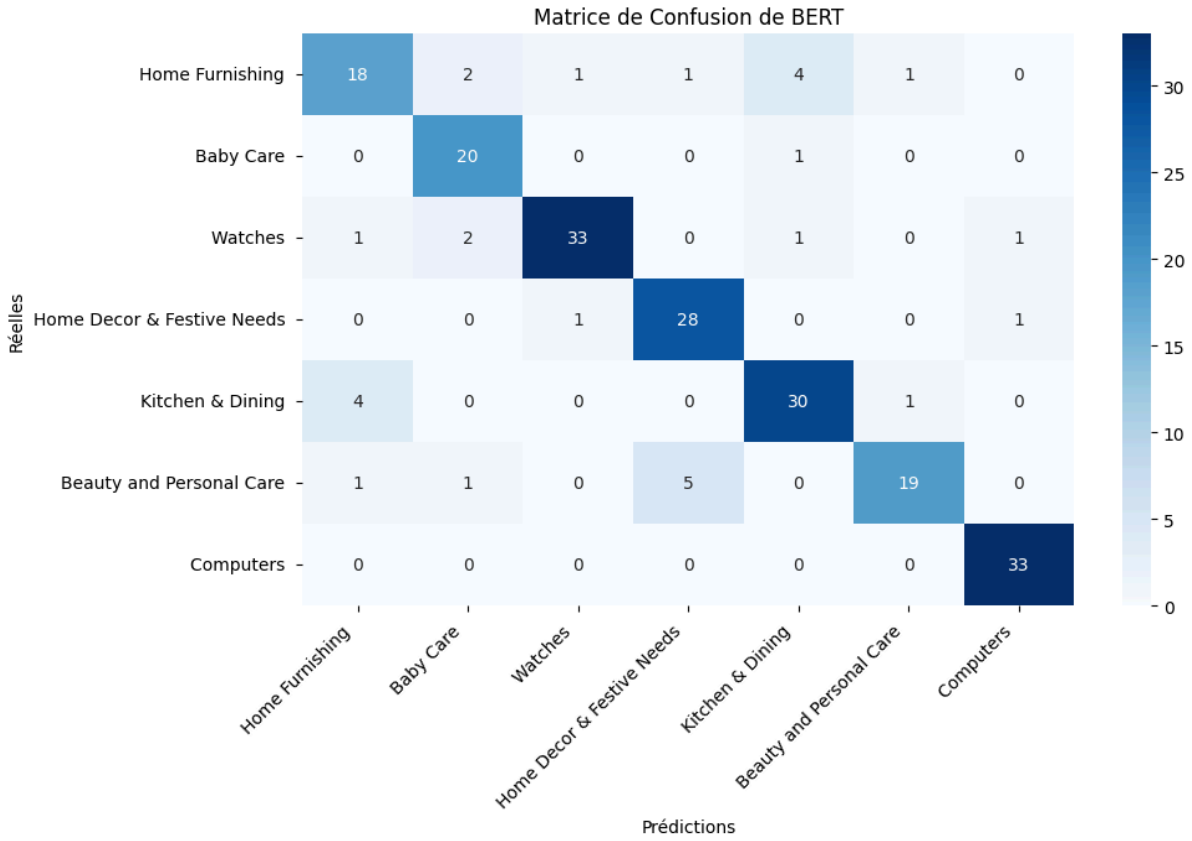
1-2 - Matrice de confusion

La matrice de confusion, ci-dessous, montre que le modèle **classe globalement bien les catégories**, notamment **"Watches"**.

Cependant, on observe quelques erreurs de classification :

- **"Home Furnishing"** est parfois confondu avec **"Kitchen & Dining"** (4 erreurs).

- **"Beauty and Personal Care"** est confondu dans 6 cas (5 erreurs avec **"Kitchen & Dining" et 1 avec "Baby Care"**).



2- ModernBERT

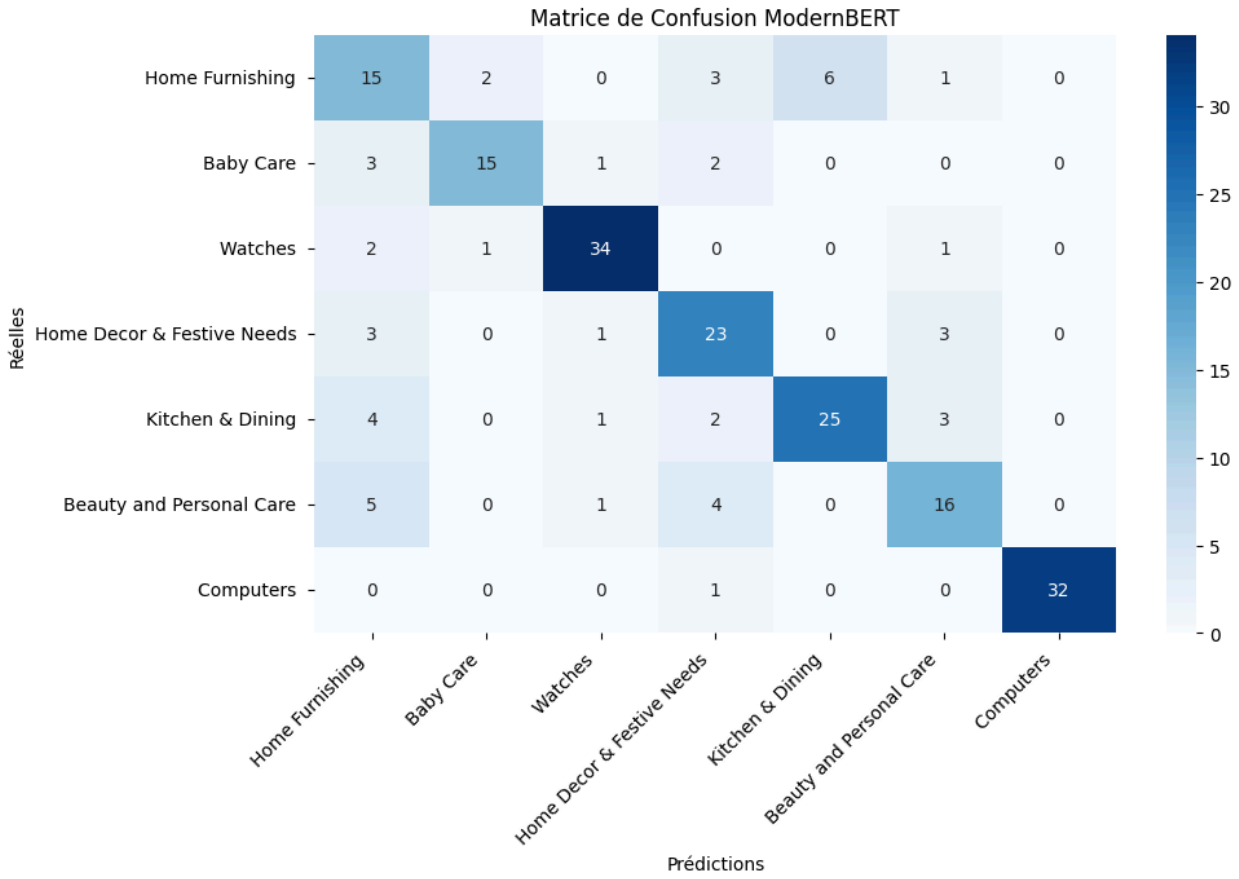
2-1- Rapport de classification

```
Rapport de classification :
                           precision    recall  f1-score   support

                Baby Care       0.47      0.56      0.51        27
  Beauty and Personal Care       0.83      0.71      0.77        21
                Computers       0.89      0.89      0.89        38
Home Decor & Festive Needs       0.66      0.77      0.71        30
           Home Furnishing       0.81      0.71      0.76        35
          Kitchen & Dining       0.67      0.62      0.64        26
                   Watches       1.00      0.97      0.98        33

                  accuracy                           0.76       210
                 macro avg       0.76      0.75      0.75       210
              weighted avg       0.77      0.76      0.77       210

Précision globale : 0.76
```

On remarque une **baisse du F1-score moyen (75 % contre 85 % pour BERT)**, avec des performances plus faibles sur certaines catégories comme **"Baby Care" (F1-score = 0.51 contre 0.71 avec BERT)** et **"Home Furnishing" (0.76 contre 0.85)**. En revanche, la catégorie **"Watches" reste très bien classée avec un F1-score de 0.98**.

Cela indique que **ModernBERT**, bien que plus léger et rapide, **compromet la performance de classification par rapport à BERT** sur ce dataset.

2-2 - Matrice de confusion



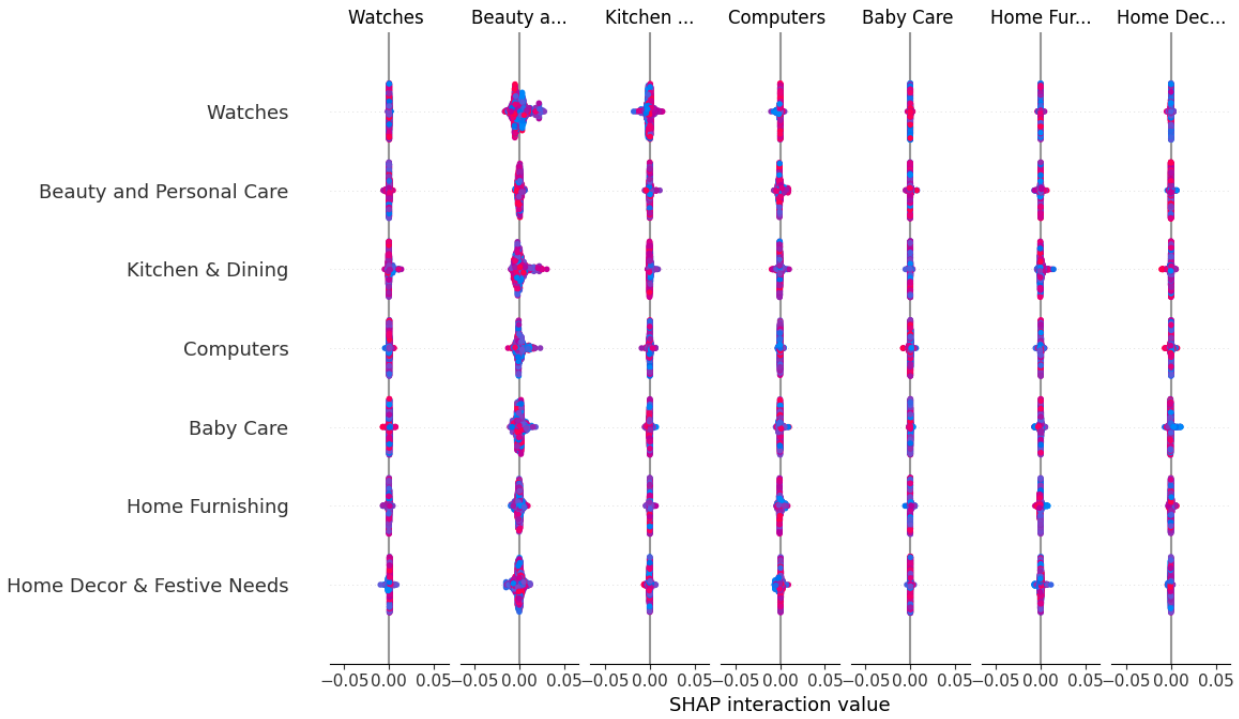
Comparée à BERT, ModernBERT montre :

- Plus d'erreurs pour **"Home Furnishing"**, qui est souvent confondu avec **"Kitchen & Dining" et "Home Decor & Festive Needs"**.
- Une légère amélioration pour **"Watches" (34 bonnes prédictions sur 35)**.
- Une baisse de performance pour **"Home Decor & Festive Needs" (23 bonnes prédictions sur 30,** contre **28 pour BERT**).
- Une plus grande confusion entre **"Beauty and Personal Care" et d'autres catégories**.

Globalement, **ModernBERT** semble légèrement **moins performant que BERT** sur ces catégories, ce qui correspond à la baisse de l'accuracy observée (**0.76** contre **0.86** pour **BERT**).

V- L'analyse de la feature importance globale et locale du nouveau modèle

1- Feature importance globale



Analyse des Features

a- "Beauty and Personal Care" a une dispersion plus forte que les autres. Cela signifie qu'elle influence plus fortement les prédictions du modèle.

- Les valeurs élevées (en rose) tendent à avoir un effet positif ou négatif significatif selon les points.
- Cela peut signifier que cette catégorie est bien différenciée dans le modèle.

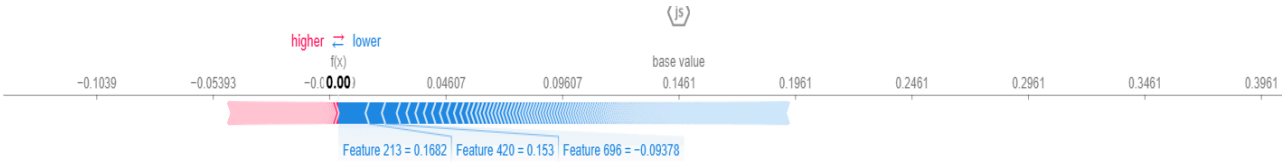
b- "Kitchen & Dining" et "Computers" ont aussi une distribution large des valeurs SHAP, ce qui indique qu'elles influencent bien la classification.

c- "Watches", "Baby Care", "Home Furnishing" et "Home Decor & Festive Needs" semblent avoir moins d'impact sur les décisions du modèle.

- Leur distribution est plus resserrée autour de zéro, ce qui signifie que le modèle ne se base pas fortement sur ces features pour la classification.

2- Feature importance locale

Le **feature important locale** montre comment les features influencent la prédiction d'un échantillon donné.



- **Feature 213 et 420** sont les plus influentes pour pousser la prédiction vers le haut.
- **Feature 696** réduit la prédiction.
- **Mais globalement, l'effet des features annule presque totalement la prédiction, qui reste proche de 0**.

VI- Les limites et les améliorations possibles

1- Perte de précision dans certains cas :

- Si l'optimisation a réduit la capacité du modèle, il peut être moins performant sur des tâches complexes nécessitant un fort pouvoir de généralisation.

- ModernBERT peut sacrifier un peu de précision au profit de la rapidité.

2- Moins de ressources et de communauté

- Contrairement à **BERT**, qui a une énorme communauté et des milliers de variantes sur ***Hugging Face***, **ModernBERT** est moins documenté et moins compatible avec certains frameworks.

3 - Adaptation aux datasets spécifiques

- Les trainsets de **ModernBERT** sont principalement en anglais et en code, les performances peuvent donc être inférieures pour d'autres langues.
- Comme tout modèle **LLM**, ModernBERT peut produire des représentations qui reflètent les biais présents dans ses données d'entraînement. Il faut vérifier les résultats critiques ou sensibles avant de vous y fier.

4- Peu d'avantages si l'on a du GPU puissant

- ModernBERT est surtout utile pour des ressources limitées (CPU, TPU edge…).
- Si tu as un GPU performant, BERT classique peut être tout aussi rapide et plus précis après fine-tuning.

Sources bibliographiques

- Arxiv : https://arxiv.org/pdf/2412.13663
- Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., ... & Poli, I. (2024). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.
- Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.