Note Méthodologique - Proof of Concept (POC) : Classification de Produits avec all-MiniLM-L6-v2

1. Objectif du POC

L'objectif de ce Proof of Concept (POC) est de démontrer l'efficacité du modèle all-MiniLM-L6-v2 pour la classification de produits à partir de leurs descriptions, en utilisant des techniques avancées de traitement du langage naturel (NLP). Ce POC a également pour objectif de comparer all-MiniLM-L6-v2 avec un modèle préalablement testé, bert-base-uncased, en termes de performances pour la tâche de classification de produits.

Plus spécifiquement, l'objectif est de mettre en évidence la capacité d'all-MiniLM-L6-v2 à générer des embeddings sémantiques pertinents (représentations vectorielles) à partir des descriptions de produits et à les utiliser pour prédire la catégorie du produit à l'aide d'un modèle de régression logistique. Ce POC inclut également une évaluation détaillée en utilisant plusieurs métriques classiques telles que le Classification Report de sklearn.metrics, mais également une analyse visuelle des embeddings à l'aide de la méthode t-SNE, ainsi qu'un clustering via KMeans.

Justification du choix du modèle :

all-MiniLM-L6-v2 a été sélectionné pour sa légèreté en termes de calcul et son efficacité sur des tâches de similarité de texte et de recherche sémantique. En dépit de sa taille réduite, il maintient une performance solide sur des tâches de classification, ce qui est particulièrement pertinent dans un contexte e-commerce où l'efficacité computationnelle est primordiale.

2. Jeu de données

Le jeu de données utilisé dans ce POC provient de Flipkart, une plateforme de commerce électronique. Il contient plusieurs informations sur les divers produits, dont :

- product_name : Le nom du produit.
- description : La description textuelle du produit
- product_category_tree : La catégorie hiérarchique du produit.
- product_category : La catégorie principale du produit extraite du product_category_tree.

La tâche consiste à prédire la catégorie principale du produit à partir de sa description textuelle, en utilisant les embeddings générés par le modèle all-MiniLM-L6-v2.

Distribution des catégories : Le jeu de données contient plusieurs catégories de produits, dont certaines peuvent potentiellement présenter de forts déséquilibres, ce qui poserait des défis supplémentaires pour la classification, en particulier pour les modèles qui ne gèrent pas bien les classes sous-représentées.

3. Modèle utilisé

MiniLM : Un modèle compact et performant pour le traitement du langage naturel (NLP)

Dans le domaine de l'intelligence artificielle, les modèles de traitement du langage naturel (Natural Language Processing ou NLP) comme BERT et RoBERTa sont extrêmement puissants, mais leur taille les rend lents et difficiles à utiliser sur des machines classiques.

Le principe de distillation des connaissances permet de compresser ces modèles massifs (professeurs) en versions plus légères (élèves), tout en conservant un haut niveau de performance. MiniLM [1,2,3] est une approche innovante qui réduit considérablement la taille des modèles tout en maintenant jusqu'à 99 % de précision sur certaines tâches, avec une vitesse d'inférence deux fois plus rapide !

Les créateurs de MiniLM ont également rendu leurs modèles accessibles en open source [3]. Initialement, ces modèles étaient conçus pour des tâches de compréhension du langage naturel (NLU), comme répondre à des questions à partir d'un texte. Mais ils peuvent aussi être utilisés pour des tâches de génération de texte (NLG), comme le résumé automatique, en s'inspirant de l'approche UniLM qui utilise des mécanismes d'attention spécifiques.

MiniLM : Une approche de distillation efficace

Le premier article sur MiniLM a été publié en 2019 et a posé les bases d'une distillation efficace des grands modèles NLP. D’autres modèles plus légers de BERT existaient déjà (DistillBERT, TinyBERT, MobileBERT), mais MiniLM se distingue par plusieurs innovations majeures :

1. Une double approche de distillation

MiniLM propose deux stratégies pour entraîner un modèle élève :

- Si le modèle élève a autant de couches que le professeur mais avec une taille de représentation plus petite, il peut apprendre directement du professeur.
- Si le modèle élève a moins de couches, un assistant professeur intermédiaire est utilisé pour faciliter l’apprentissage.

Cette méthode est recommandée lorsque $M < \frac{1}{2}L \quad \text{et} \quad d_i < \frac{1}{2}d$

Avec L et M représentant le nombre de couches, et d et d_i la taille des représentations pour le professeur et l'élève respectivement.

2. Une distillation ciblée sur la dernière couche du modèle

Contrairement à d'autres approches qui forcent l'élève à imiter plusieurs couches du professeur, MiniLM ne transfère que la dernière couche du Transformer. Cela rend l'architecture plus flexible et permet à l'élève d'avoir moins de couches que le professeur.

3. Un transfert d'attention amélioré

L'apprentissage est optimisé grâce à un transfert d'attention double :

- Q-K (Questions-Clés)
- V-V (Valeurs-Valeurs)

La distillation est réalisée en minimisant la perte KL (Kullback-Leibler divergence), une mesure de différence entre les distributions des couches finales du professeur et de l'élève.

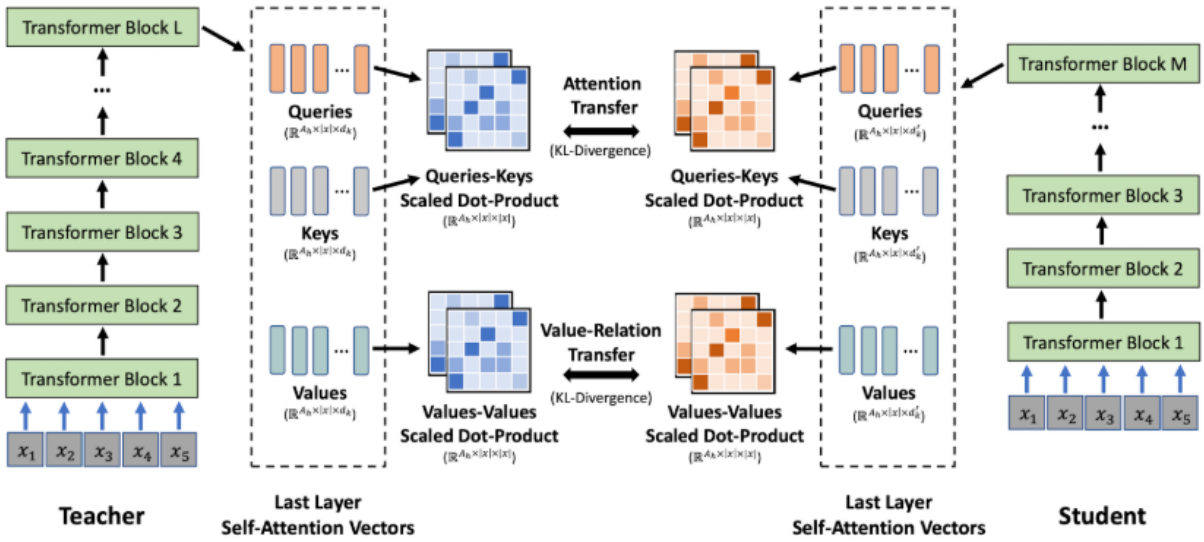

Tableau des performances de MiniLM vs modèles distillés concurrents

4. Performances exceptionnelles

MiniLM a été comparé aux autres techniques de distillation existantes. Il offre une précision élevée sur plusieurs ensembles de données tout en étant beaucoup plus compact.

Approach	Teacher Model	Distilled Knowledge	Layer-to-Layer Distillation	Requirements on the number of layers of students	Requirements on the hidden size of students
DistillBERT	$\text{BERT}_{\text{BASE}}$	Soft target probabilities Embedding outputs			✓
TinyBERT	$\text{BERT}_{\text{BASE}}$	Embedding outputs Hidden states Self-Attention distributions	✓		
MOBILEBERT	$\text{IB-BERT}_{\text{LARGE}}$	Soft target probabilities Hidden states Self-Attention distributions	✓	✓	✓
MINILM	$\text{BERT}_{\text{BASE}}$	Self-Attention distributions Self-Attention value relation			

Model	#Param	SQuAD2	MNLI-m	SST-2	QNLI	CoLA	RTE	MRPC	QQP	Average
$\text{BERT}_{\text{BASE}}$	109M	76.8	84.5	93.2	91.7	58.9	68.6	87.3	91.3	81.5
DistillBERT	66M	70.7	79.0	90.7	85.3	43.6	59.9	87.5	84.9	75.2
TinyBERT	66M	73.1	83.5	91.6	90.5	42.8	**72.2**	**88.4**	90.6	79.1
MINILM	66M	**76.4**	**84.0**	**92.0**	**91.0**	**49.2**	71.5	**88.4**	**91.0**	**80.4**

Tableaux des performances de MiniLM vs modèles distillés concurrents

Enfin, par rapport à BERT-Base, MiniLM parvient à réduire le nombre de paramètres par un facteur 3 tout en conservant une précision équivalente, voire supérieure !

Model	#Param	SQuAD 2.0	MNLI-m	SST-2	QNLI	CoLA
BERT-Base	109M	76.8	84.5	93.2	91.7	58.9
MiniLM-L12xH384	33M	81.7	85.7	93.0	91.5	58.5
MiniLM-L6xH384	22M	75.6	83.3	91.5	90.5	47.5

Tableau de comparaison des modèles MiniLM et BERT-Base

MiniLMv2 : Des améliorations pour plus de puissance

En 2021, une nouvelle version, MiniLMv2, a été publiée avec plusieurs améliorations significatives :

1. Une distillation d'attention plus avancée

MiniLMv2 généralise la technique de distillation de MiniLMv1 en utilisant un transfert de relations d'auto-attention. Cette approche supprime la contrainte sur le nombre de têtes d'attention du modèle élève, rendant le processus plus flexible.

2. Une meilleure sélection des couches du modèle professeur

Dans MiniLMv1, la dernière couche du professeur était systématiquement transférée à l'élève. MiniLMv2 optimise ce processus en sélectionnant des couches spécifiques selon le modèle de départ :

- Pour BERT-large, la 21^{e} couche est utilisée.
- Pour RoBERTa-large et XML-R-large, la 19^{e} couche est utilisée.
- Pour les modèles plus petits, la dernière couche est conservée.

Cette approche permet de mieux capturer les informations importantes des modèles professeurs.

3. Une expérimentation sur plusieurs types de relations d'auto-attention

Les chercheurs ont testé le transfert de plusieurs relations d'auto-attention entre le professeur et l'élève (Q-K, K-Q, Q-V, V-Q, K-V et V-K).

Cependant, introduire trop de relations augmente considérablement le coût de calcul. Pour un équilibre optimal entre performance et vitesse, MiniLMv2 ne transfère que les relations suivantes :

- Q-Q (Questions-Questions)
- K-K (Clés-Clés)
- V-V (Valeurs-Valeurs)

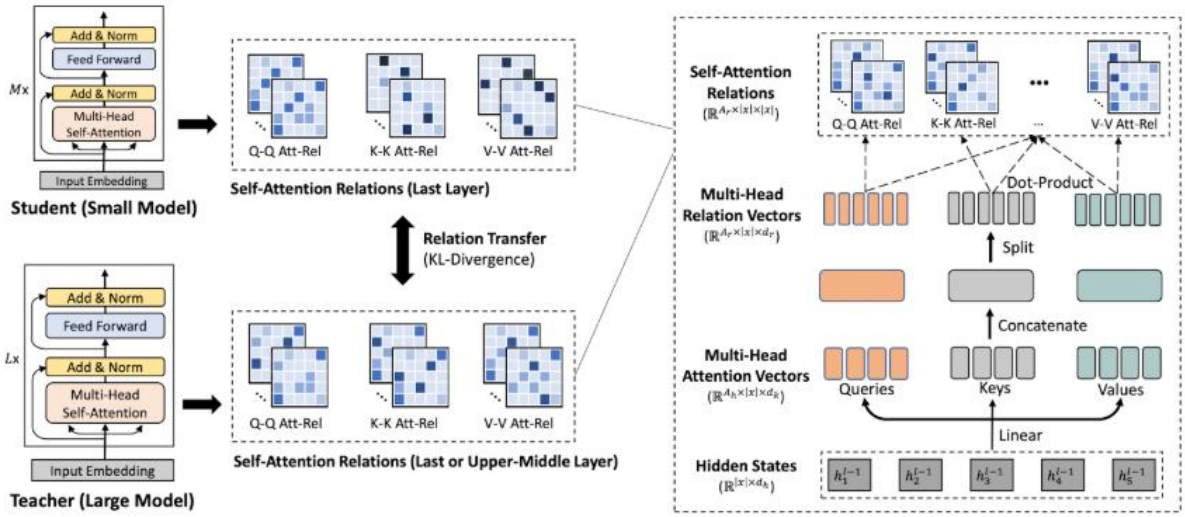

Schéma illustrant la distillation multi-têtes dans MiniLMv2

4. Des performances améliorées

MiniLMv2 offre des gains de vitesse et de précision par rapport à MiniLMv1 et aux autres modèles distillés.

Model	Teacher Model	Speedup	#Param	MNLI-m (Acc)	SQuAD 2.0 (F1)
L6xH768 MiniLMv2	RoBERTa-Large	2.0x	81M	87.0	81.6
L12xH384 MiniLMv2	RoBERTa-Large	2.7x	41M	86.9	82.3
L6xH384 MiniLMv2	RoBERTa-Large	5.3x	30M	84.4	76.4
L6xH768 MiniLMv2	BERT-Large Uncased	2.0x	66M	85.0	77.7
L6xH384 MiniLMv2	BERT-Large Uncased	5.3x	22M	83.0	74.3
L6xH768 MiniLMv2	BERT-Base Uncased	2.0x	66M	84.2	76.3
L6xH384 MiniLMv2	BERT-Base Uncased	5.3x	22M	82.8	72.9

Tableau comparatif des performances et de la rapidité des modèles MiniLMv2

Pourquoi utiliser all-MiniLM-L6-v2 pour la classification de produits ?

MiniLM et MiniLMv2 permettent d'obtenir des modèles légers, rapides et précis, idéaux pour les applications en intelligence artificielle embarquée, en analyse de texte en temps réel et en traitement du langage naturel sur des ressources limitées.

- MiniLM est 3 fois plus compact que BERT avec une précision quasi identique.
- MiniLMv2 améliore encore l'efficacité grâce à des techniques de distillation avancées.
- Une alternative idéale pour ceux qui cherchent un modèle puissant mais rapide et léger.

4. Description de la méthode

Le processus de modélisation se décompose comme suit :

Préparation des données :

Le jeu de données est d'abord nettoyé et filtré pour ne conserver que les colonnes essentielles à la tâche de classification : l'ID unique du produit, le nom, la catégorie et la description. La catégorie principale du produit est extraite de l'arbre hiérarchique des catégories et nettoyée pour faciliter le traitement. Un prétraitement des descriptions (tokenisation, nettoyage de texte) est réalisé pour les rendre compatibles avec les modèles de NLP.

Extraction des embeddings :

Le modèle all-MiniLM-L6-v2 est chargé via la bibliothèque sentence-transformers. Les descriptions des produits sont dès lors encodées en embeddings, chacune étant traitée par lots de 10 éléments dans le but d'optimiser la gestion de la mémoire. Grâce à cette manœuvre, chaque produit reçoit un vecteur de caractéristiques, représentant les informations sémantiques de sa description dans un espace vectoriel dense.

Pour comparaison, le même processus est effectué pour le modèle bert-base-uncased afin de comparer la qualité des embeddings produits par les deux modèles.

Classification des produits :

Les embeddings extraits sont utilisés comme caractéristiques d'entrée dans un modèle de régression logistique, afin de prédire la catégorie principale du produit. La classification est effectuée par entraînement du modèle sur un sous-ensemble des données (80%) et évaluation sur un autre sous-ensemble (20%).

Évaluation des performances :

Les performances des modèles sont évaluées à l'aide de plusieurs métriques :

- Accuracy : La précision globale du modèle.
- Classification Report : Un rapport de classification détaillant la précision, le rappel et le score F1 pour chaque catégorie est généré.
- Clustering avec KMeans et ARI : Le clustering des embeddings via KMeans permet d'observer la séparation des catégories dans l'espace des embeddings. Le Adjusted Rand Index (ARI) est calculé pour évaluer la qualité de l'association entre les clusters et les catégories réelles.

Visualisation des résultats avec t-SNE :

Pour mieux comprendre la distribution des embeddings dans l'espace vectoriel, une réduction de dimension est effectuée à l'aide de t-SNE, suivi d'une analyse visuelle des résultats pour illustrer la distribution des embeddings des produits dans un espace bidimensionnel. Cette réduction de dimension permet de visualiser comment les produits sont regroupés en fonction de leurs descriptions et de la similarité de leurs embeddings.

Les résultats sont ensuite comparés avec un clustering KMeans pour observer la correspondance entre les clusters et les catégories réelles. Un score de l'Adjusted Rand Index (ARI) est calculé pour évaluer la qualité de l'association.

5. Résultats obtenus

Les résultats du POC montrent les performances respectives des deux modèles :

Prédiction de la catégorie des produits avec régression logistique :

Rapport de classification pour bert-base-uncased :

Catégorie	Precision	Recall	F1-score	Support
Baby Care	0.88	0.78	0.82	27
Beauty and Personal Care	1.00	1.00	1.00	21
Computers	0.97	0.97	0.97	38
Home Decor & Festive Needs	0.85	0.93	0.89	30
Home Furnishing	0.91	0.89	0.90	35
Kitchen & Dining	0.93	0.96	0.94	26
Watches	1.00	1.00	1.00	33
Accuracy	0.93			210

Rapport de classification pour all-MiniLM-L6-v2 :

Catégorie	Precision	Recall	F1-score	Support
Baby Care	1.00	0.70	0.83	27
Beauty and Personal Care	0.88	1.00	0.93	21
Computers	0.97	1.00	0.99	38
Home Decor & Festive Needs	0.85	0.97	0.91	30
Home Furnishing	0.94	0.97	0.96	35
Kitchen & Dining	1.00	0.96	0.98	26
Watches	1.00	1.00	1.00	33
Accuracy	0.95			210

Le modèle all-MiniLM-L6-v2, avec une performance globale élevée, montre une très bonne capacité à distinguer les catégories, notamment avec une accuracy de 94.76 %, supérieure à celle de bert-base-uncased (93 %).

Classification par catégories :

De plus Le modèle a bien performé pour toutes les catégories, avec des scores F1 supérieurs à 0.90 pour presque toutes les catégories. Les catégories avec des scores F1 les plus élevés incluent Computers, Kitchen & Dining, et Watches, qui ont des performances quasi parfaites.

Ainsi, le modèle de régression logistique a bien appris les distinctions entre les différentes catégories, même dans des catégories déséquilibrées. Certaines catégories, comme Beauty and Personal Care, présentent un score F1 élevé, ce qui suggère que le modèle parvient à différencier efficacement ces classes.

Visualisation avec t-SNE et clustering KMeans :

Score ARI (Adjusted Rand Index) :

bert-base-uncased : 0.3095

all-MiniLM-L6-v2 : 0.7119

Le clustering des embeddings via KMeans a montré que les produits étaient en grande partie bien séparés en clusters distincts, et que ces clusters avaient une correspondance notable avec les catégories réelles. Cependant, certains chevauchements de produits persistent, notamment pour les catégories les plus proches sur le plan sémantique (ex. Baby Care et Home Decor & Festive Needs).

Pour autant, le modèle all-MiniLM-L6-v2 a montré une meilleure séparation des produits dans l'espace des embeddings, avec un score ARI significativement plus élevé que celui de bert-base-uncased, ce qui suggère une meilleure cohérence dans le clustering des produits.

6. Interprétation des résultats

Les résultats obtenus confirment que all-MiniLM-L6-v2, couplé avec un classificateur simple comme la régression logistique, constitue une méthode efficace pour la classification de produits à partir de leurs descriptions. Avec une accuracy de 94.76%, %, dépassant légèrement bert-base-uncased (93%), le modèle montre une très bonne performance. Le modèle montre également de bonnes capacités de séparation des catégories, avec des scores F1 supérieurs à 0.90 pour la majorité des catégories. En particulier, des catégories comme Computers, Kitchen & Dining, et Watches ont montré des résultats quasi parfaits.

L'Adjusted Rand Index (ARI) de 0.7119 suggère une bonne séparation des produits dans l'espace des embeddings, ce qui est corroboré par les résultats du clustering KMeans. Bien que quelques chevauchements existent entre certaines catégories (par exemple, Baby Care et Home Decor & Festive Needs), le modèle parvient à bien capturer la structure des données.

De plus, l'ARI élevé de all-MiniLM-L6-v2 indique que le modèle parvient à bien structurer les relations entre les catégories de produits, même lorsque certaines sont sous-représentées. Cela signifie que le modèle est capable de gérer efficacement le déséquilibre des classes en maintenant une forte capacité discriminante. Un ARI élevé montre que les produits appartenant à la même catégorie réelle sont bien regroupés dans des clusters cohérents, ce qui renforce la robustesse du modèle et son adéquation à la tâche de classification.

Ainsi, vis-à-vis de bert-base-uncased, qui a montré des résultats solides mais légèrement inférieurs, le modèle all-MiniLM-L6-v2 le surpasse.

7. Conclusion et perspectives

Ce POC démontre l'efficacité du modèle all-MiniLM-L6-v2 pour la classification de produits dans un contexte e-commerce. Les résultats montrent des performances solides, surpassant celles de bert-base-uncased en termes d'accuracy et de séparation des catégories. Le modèle est prometteur pour des applications réelles en production. Cependant, des pistes d'amélioration existent pour perfectionner le modèle, notamment :

- Affinage des embeddings : L'intégration de la description complète du produit ou d'autres attributs (par exemple, la marque) pourrait améliorer les résultats, surtout pour les catégories plus difficiles à séparer.
- Optimisation de l'architecture : L'utilisation de modèles plus complexes ou l'ajustement des hyperparamètres du modèle pourrait potentiellement améliorer encore la précision dans les catégories plus complexes.
- Augmentation des données : Enrichir le jeu de données avec des données supplémentaires davantage détaillées pourrait également augmenter la diversité des exemples et améliorer la robustesse du modèle.

En somme, ce POC valide l'utilisation de all-MiniLM-L6-v2 pour la classification de produits dans des environnements de e-commerce, tout en montrant qu'il existe encore des pistes pour perfectionner ce modèle pour des applications en production.

Ressources bibliographiques :

- Wang et al., "**MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers**"
    - **Arxiv** : https://arxiv.org/abs/2002.10957
    - **Paperswithcode** : https://paperswithcode.com/paper/200210957
    - **ConnectedPapers** : https://www.connectedpapers.com/search?q=MiniLM
- **SentenceTransformers Documentation** : https://www.sbert.net/
- **Hugging Face Model Hub - all-MiniLM-L6-v2 :**

  https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
- **Building a Recommendation System with Hugging Face Transformers :**

  https://www.kdnuggets.com/building-a-recommendation-system-with-hugging-face-transformers
- **Implementing Multi-Modal RAG Systems :**

  https://machinelearningmastery.com/implementing-multi-modal-rag-systems/