Alexandre ROGUES
Etudiant Data Scientist OpenClassRooms
Projet 8 – Réalisez un dashboard et assurez une veille technique
Juillet 2024

Note Méthodologique:
Preuve de Concept du Modèle PHI-3 Mini 4K

Table des matières

I. Dataset retenu

Le dataset retenu pour cette proof of concept est issu du Projet 6 « Classifiez automatiquement des biens de consommation ».
Ce dataset est fourni par une entreprise intitulée e-commerce marketplace.
Pour rappel le principe du projet 6 consiste à vérifier la faisabilité de labellisation automatique d'items en vente sur le site internet, à la fois à partir d'informations de descriptions textuelles des items (traitement NLP) et à partir d'illustrations visuelles des images.
L'objectif est ici de faciliter à la fois la mise en ligne d'items par les vendeurs ainsi que la recherche d'items par les clients, tout en gardant en tête la perspective de passage à l'échelle.

Ce dataset contient 1050 items classifié équitablement (150 items pour chaque) dans 7 catégories

- Home Furnishing
- Baby Care
- Watches
- Home Decor & Festive Needs
- Kitchen & Dining
- Beauty and Personal Care
- Computers

Chaque item est renseigné à propos de 15 features :
'uniq_id', 'crawl_timestamp', 'product_url', 'product_name', product_category_tree', 'pid','retail_price', 'discounted_price', 'image', 'is_FK_Advantage_product', 'description', 'product_rating', 'overall_rating', 'brand', 'product_specifications'

On ne va s'intéresser qu'à la feature description pour cette POC. Il s'agit d'un texte descriptif en langue anglaise. En voici un exemple :

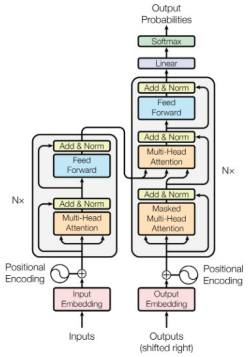
Description: Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 This curtain enhances the look of the interiors.This curtain is made from 100% high quality polyester fabric.It features an eyelet style stitch with Metal Ring.It makes the room environment romantic and loving.This curtain is ant- wrinkle and anti shrinkage and have elegant apparance.Give your home a bright and modernistic appeal with these designs. The surreal attention is sure to steal hearts. These contemporary eyelet and valance curtains slide smoothly so when you draw them apart first thing in the morning to welcome the bright sun rays you want to wish good morning to the whole world and when you draw them close in the evening, you create the most special moments of joyous beauty given by the soothing prints. Bring home the elegant curtain that softly filters light in your room so that you get the right amount of sunlight.,Specifications of Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) General Brand Elegance Designed For Door Type Eyelet Model Name Abstract Polyester Door Curtain Set Of 2 Model ID Duster25 Color Multicolor Dimensions Length 213 cm In the Box Number of Contents in Sales Package Pack of 2 Sales Package 2 Curtains Body & Design Material Polyester

II. Les concepts du modèle PHI3

A) Un modèle génératif basé sur les Transformers

Les modèles génératifs basés sur les transformers ont révolutionné le domaine du traitement du langage naturel (NLP) grâce à leur capacité à comprendre et à générer du texte de manière efficace.

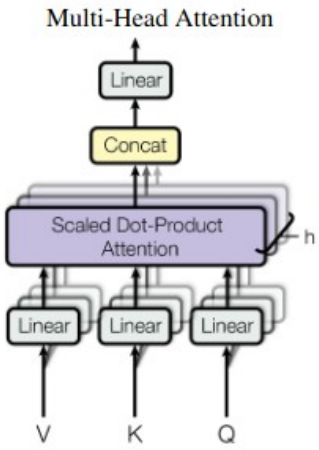
Voici les concepts clés de ces modèles :



Principe des Encoders-Decoders :

Les transformers utilisent une architecture en deux parties : l'encoder et le decoder. L'encoder transforme les séquences de texte en représentations internes denses et informatives, tandis que le decoder génère les séquences de texte de sortie à partir de ces représentations.

Principe de l'Attention (Attention Mechanism) :



Self-Attention : Permet à chaque mot d'une séquence de se concentrer sur tous les autres mots de la séquence, indépendamment de leur position, améliorant ainsi la compréhension contextuelle.

Multi-Head Attention : Utilise plusieurs "têtes" d'attention pour capturer différentes nuances contextuelles simultanément, offrant une compréhension plus riche et variée du contexte.

Ce principe a été introduit par la publication "Attention Is All You Need" ; publié en 2017, ce papier a introduit le mécanisme d'attention, transformant la façon dont les modèles NLP gèrent les dépendances de longue portée dans les séquences de texte.

Efficacité et Performance :

Les transformers équilibrent la performance et l'efficacité en utilisant des optimisations qui maximisent la capacité de traitement tout en minimisant les ressources nécessaires.

B) Phi-3 mini 4K

Phi-3 mini 4K est un modèle de langage développé par Microsoft Research, s'inscrivant dans la tendance actuelle de créer des modèles d'intelligence artificielle à la fois efficaces et compacts. Ce modèle, bien que plus petit que de nombreux autres, rivalise en performance grâce à des innovations clés.

Caractéristiques principales :

Phi-3 mini 4K est un modèle compact, mais suffisamment puissant pour être utilisé dans des applications nécessitant des ressources limitées. Entraîné sur 3,3 trillions de tokens : Le vaste ensemble de données d'entraînement, comprenant des données web filtrées et des données synthétiques, améliore la performance du modèle.

Les performances de Phi-3 mini 4K sont comparables à celles de modèles beaucoup plus grands comme Mixtral 8x7B et GPT-3.5, grâce à la qualité des données d'entraînement. Le modèle est suffisamment compact pour fonctionner sur un téléphone, rendant les applications IA plus accessibles.

Développé par une équipe de chercheurs de Microsoft, ce modèle est un exemple de l'innovation continue dans le domaine de l'IA. En plus de la version mini, il existe des versions plus grandes avec 7 milliards et 14 milliards de paramètres, offrant une gamme de modèles pour différentes applications. Une version multimodale du modèle, PHI3-Vision, intègre des capacités de traitement visuel et textuel.

C) Spécifications Techniques et Méthodes d'Entraînement :

Les données utilisées pour entraîner le modèle sont soigneusement filtrées pour assurer une haute qualité. En plus des données web filtrées, des données synthétiques sont utilisées pour enrichir l'ensemble d'entraînement, augmentant ainsi la diversité et la robustesse du modèle.

Des techniques avancées sont employées pour ajuster les hyperparamètres et régulariser le modèle. Cela inclut l'utilisation de la Maximal Update Parametrization (muP) et l'activation GEGLU pour améliorer les performances. Le modèle utilise aussi des modules d'attention blocksparse pour optimiser la vitesse d'entraînement et d'inférence, tout en maintenant une bonne performance sur les longues séquences.

Pour garantir la fiabilité et la sécurité du modèle, plusieurs mesures sont mises en place. Le modèle subit une post-formation axée sur la sécurité, incluant le fine-tuning supervisé et l'optimisation des préférences directes. Des ensembles de données RAI (Responsible AI) sont utilisés pour évaluer et améliorer la performance en matière de sécurité, et une équipe indépendante de "red-teaming" examine le modèle pour identifier et corriger les domaines à risque.

Conclusion :

Phi-3 mini 4K représente une avancée significative dans le domaine des modèles de langage, combinant une taille réduite avec des performances robustes. Sa conception vise à équilibrer l'efficacité et la performance, rendant l'IA plus accessible et durable. Les innovations dans la qualité des données d'entraînement et les optimisations internes permettent à ce modèle de rivaliser avec des modèles beaucoup plus grands, tout en étant suffisamment compact pour fonctionner sur des appareils mobiles. En somme, Phi-3 mini 4K est un exemple de l'évolution continue vers des modèles d'IA plus efficaces et performants, répondant aux besoins de diverses applications tout en minimisant les coûts de calcul et l'empreinte énergétique.

III.Modélisation et Méthodologie

Pour cette proof of concept (POC), deux méthodes principales ont été utilisées pour la classification automatique des items en vente sur le site marketplace, en utilisant uniquement les descriptions textuelles. Ces méthodes ont été mises en œuvre pour vérifier la performance du model PHI3 comparé aux autres modèles.

A. Méthode 1 : Modèle de Complétion

1. Interrogation du Modèle

◦ Utilisation du modèle de complétion PHI3 pour prédire la catégorie des items en soumettant la description textuelle complète.

2. Affinage des Réponses

◦ Ajustement des réponses du modèle (fonction de nettoyage) pour aligner les prédictions avec les sept catégories spécifiques du dataset : Home Furnishing, Baby Care, Watches, Home Decor & Festive Needs, Kitchen & Dining, Beauty and Personal Care, Computers.

3. Évaluation

◦ Création d'une matrice de confusion pour comparer les labels prédits avec les labels réels.
◦ Calcul des métriques d'évaluation classiques telles que l'accuracy, le F1-score, la précision et le rappel.

B. Méthode 2 : Embeddings et Clustering

1. Extraction des Embeddings

◦ Utilisation du modèle pour obtenir les vecteurs d'embeddings des descriptions textuelles, l’objectif étant de transformer chaque description en un vecteur de caractéristiques.

2. Réduction de Dimensionnalité

◦ Application d'une Analyse en Composantes Principales (PCA) pour réduire la dimensionnalité des embeddings.
◦ Utilisation de T-SNE pour projeter les données réduites en un espace à 2 dimensions afin de faciliter la visualisation.

3. Clustering

◦ Application de l'algorithme de clustering (K-means) pour regrouper les items en 7 clusters.
◦ Comparaison des clusters obtenus avec les catégories réelles en utilisant l'Indice de Rand Ajusté (ARI).

4. Évaluation

◦ Assignation de chaque cluster au label réel le plus fréquent en son sein pour aligner les prédictions avec les labels réels et évaluer la performance.
◦ Création d'une matrice de confusion pour comparer les clusters avec les catégories réelles.
◦ Calcul des métriques d'évaluation (ARI) pour évaluer la qualité du clustering.

C. Métriques d'Évaluation

Les principales métriques d'évaluation utilisées pour comparer les labels réels et les labels prédits incluent :

• Accuracy : Mesure la proportion de prédictions correctes parmi l'ensemble des prédictions.
• F1-score : Harmonie entre la précision et le rappel, particulièrement utile pour les datasets déséquilibrés.
• Précision (Precision) : Mesure la proportion de prédictions correctes parmi les prédictions faites pour une classe donnée.
• Rappel (Recall) : Mesure la proportion de prédictions correctes parmi les instances réelles de cette classe.
• Indice de Rand Ajusté (ARI) : Utilisé pour évaluer la similarité entre les clusters prédits et les catégories réelles, prenant en compte les accords dus au hasard.

D. Démarche d'Optimisation

Aucune démarche d'optimisation extensive n'a été réalisée au-delà de la comparaison des deux méthodes mises en œuvre. Toutefois, les points suivants ont été pris en considération :

1. Comparaison des Méthodes

◦ Analyse comparative des résultats des deux méthodes (modèle de complétion vs embeddings et clustering) pour identifier la plus performante.
◦ Évaluation basée sur les métriques d'évaluation classiques et ARI.

2. Affinage des Prédictions

◦ Ajustement manuel des prédictions du modèle de complétion pour garantir une meilleure correspondance avec les catégories cibles.
◦ Réduction de la dimensionnalité et clustering pour visualiser et ajuster les groupes de données.

IV. Synthèse des résultats

Pour cette proof of concept, quatre techniques ont été comparées pour la classification automatique des descriptions d'items : Word2Vec (W2V), BERT, Universal Sentence Encoder (USE) et Phi-3 mini 4K (utilisant deux approches différentes). Voici une analyse comparative des résultats obtenus par chacune des méthodes.

Word2Vec (W2V)

La méthode Word2Vec a produit les résultats suivants :

Précision moyenne : 0.48
Rappel moyen : 0.49
F1-score moyen : 0.47
ARI : 0.2176

L'analyse de la matrice de confusion montre que certains clusters, tels que le cluster 6 (Watches), ont une précision élevée (1.00), mais d'autres catégories, comme le cluster 1 (Beauty and Personal Care), ont été mal classifiées (précision et rappel de 0.00). Cela indique une grande variance dans la performance de cette méthode.

BERT

La méthode BERT a produit les résultats suivants :

Précision moyenne : 0.55
Rappel moyen : 0.53
F1-score moyen : 0.53
ARI : 0.3036

BERT a montré une amélioration par rapport à W2V, avec une meilleure performance globale. Cependant, certaines catégories comme Computers (précision et rappel de 0.21) ont toujours des performances médiocres, tandis que Watches a encore une précision et un rappel élevés (0.96 et 0.90 respectivement).

Universal Sentence Encoder (USE)

La méthode USE a produit les résultats suivants :

Précision moyenne : 0.66
Rappel moyen : 0.64
F1-score moyen : 0.64
ARI : 0.4339

USE a significativement amélioré les performances par rapport à W2V et BERT. Les clusters tels que Baby Care et Watches montrent des résultats très positifs avec des précisions élevées. Cependant, certaines catégories, bien que meilleures que dans les méthodes précédentes, montrent encore une certaine variance.

Phi-3 mini 4K

Méthode 1 : Complétion

Précision moyenne : 0.72
Rappel moyen : 0.67
F1-score moyen : 0.68
ARI : Non applicable

Cette méthode utilisant Phi-3 mini 4K a montré les meilleures performances parmi toutes les techniques testées. La précision et le rappel pour des catégories comme Watches et Computers sont très élevés. Toutefois, certaines catégories comme Home Decor & Festive Needs et Kitchen & Dining ont des performances relativement plus faibles.

Méthode 2 : Embeddings et Clustering

Précision moyenne : 0.68
Rappel moyen : 0.56
F1-score moyen : 0.59
ARI : 0.2668

Bien que cette méthode ait montré des performances acceptables, elle est inférieure à la méthode de complétion en termes de précision et de rappel moyens. Les catégories comme Watches continuent de montrer une haute performance, mais d'autres catégories ont des résultats variables.

Conclusion

L'analyse des résultats montre que les modèles basés sur les transformers, notamment Phi-3 mini 4K et BERT, surpassent les méthodes traditionnelles telles que Word2Vec. En particulier, la méthode de complétion utilisant Phi-3 mini 4K a démontré des performances supérieures avec une précision moyenne de 0.72 et un rappel moyen de 0.67.

V. Analyse de la feature importance globale et locale

L'analyse de l'importance des features dans un modèle de machine learning permet de comprendre l'influence relative de chaque variable d'entrée sur les prédictions du modèle. Cette analyse peut se faire à deux niveaux :

- **Importance Globale des Features :** Elle mesure l'impact de chaque feature sur l'ensemble des prédictions du modèle. Elle aide à identifier quelles variables sont les plus influentes en moyenne pour toutes les prédictions.
- **Importance Locale des Features :** Elle évalue l'influence de chaque feature pour des prédictions spécifiques. Cela permet de comprendre comment chaque variable affecte une prédiction individuelle.

Des techniques SHAP ou LIME sont souvent utilisées pour cette analyse.

Contexte de Notre POC

Dans le cadre de notre proof of concept (POC), nous avons utilisé uniquement une seule feature : **la description textuelle des items**, pour prédire leur catégorie.

Pourquoi l'Analyse de l'Importance des Features n'est pas Applicable ?

Dans ce contexte, il n'est pas pertinent de procéder à une analyse de l'importance des features puisque l'intégralité de la prédiction repose sur une seule et unique variable.

L'analyse de l'importance des features, qu'elle soit globale ou locale, nécessite une certaine variabilité dans les variables d'entrée pour comparer leur impact relatif. Avec une seule feature, il n'y a pas d'autres variables à comparer ou à mesurer en termes d'influence relative. De plus, l'étude de l'importance des features, même dans des modèles utilisant une seule feature, requiert une compréhension approfondie et des recherches bibliographiques sur les méthodes spécifiques, ce qui dépasse le cadre de notre POC actuel.

Notre modèle basé sur les transformers, Phi-3 mini 4K, utilise des techniques avancées de traitement du langage naturel (NLP) pour extraire des informations contextuelles et sémantiques de la description textuelle. L'analyse de l'importance des features serait plus pertinente dans un cadre où plusieurs types de données (comme les images, les prix, les spécifications techniques) seraient utilisés conjointement.

En conclusion, bien que l'analyse de l'importance des features soit une démarche cruciale dans de nombreux modèles de machine learning pour comprendre l'influence relative des différentes variables d'entrée, elle ne s'applique pas dans notre cas. Notre modèle utilise exclusivement la description textuelle des items pour la classification, rendant ainsi cette analyse non pertinente. Pour de futures itérations ou extensions du modèle, notamment avec l'intégration de données visuelles ou d'autres features textuelles, une telle analyse pourrait alors devenir utile et applicable.

VI. Limites et améliorations possibles

On rappelle que notre modèle se base uniquement sur les descriptions textuelles, ce qui le rend dépendant de la qualité et de la consistance de ces descriptions. Cette dépendance limite la richesse des informations disponibles pour la classification et peut affecter la performance globale du modèle. De plus, les transformers, comme Phi-3 mini 4K, sont souvent considérés comme des boîtes noires en raison de leur complexité, ce qui rend difficile l'explication des décisions prises par le modèle. La variabilité des descriptions textuelles ajoute une autre couche de complexité, car elles peuvent varier considérablement en qualité et en consistance.

Pour améliorer les performances et l'interprétabilité, nous pouvons optimiser les hyper-paramètres du modèle. Cependant, une amélioration plus significative viendrait de l'intégration de données multimodales, telles que les images des produits, les prix, et d'autres métadonnées. Ces sources d'information supplémentaires enrichiraient le modèle, permettant une classification plus précise et robuste. En outre, l'utilisation d'outils comme LIME ou SHAP pour expliquer les prédictions du modèle aiderait à rendre les résultats plus transparents et compréhensibles. Cette approche permettrait de mieux comprendre l'importance des différentes features et d'expliquer de manière plus claire pourquoi le modèle fait certaines prédictions.

Références

Abdin, M., Huynh, J., Saarikivi, O., Ade, S., Jacobs, M. A., Iter, D., Saied, A., et al. (2024). *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*. arXiv. https://arxiv.org/pdf/2404.14219

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*. arXiv. https://arxiv.org/pdf/1706.03762