Note Méthodologique

OpenClassrooms-Projet 8

Habibatou BA - Veille Technologique

Principe :

Réaliser une **veille** sur les outils et tendances en Data Science et IA. L'objectif étant de mettre en lumière les dernières techniques en data science sur des problématiques de données texte (NLP) et de données d'images.
Pour cela, il faudra se baser sur des techniques datant de moins de 5 ans et Veiller à ce que la technique choisie soit référencer sur des sites de recherche (**Arxiv**), des sites connus (par exemple **fastml**, **machine learning mastery**, **kdnuggets**, **import AI**, **MIT tech review**, **MIT news ML**) ou des newsletters de qualité comme **data elixir** et **data science weekly**.

Pour cette veille technologique, il sera question de traitement de langage naturel (NLP). Principalement:

- **Le modèle T5, ou "Text-To-Text Transfer Transformer", est une architecture de réseau neuronal développée par Google AI. Elle a été introduite dans un article de recherche intitulé "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" publié en 2020 (Raffel et al., 2020). Par exemple, pour une tâche de classification de texte, le texte d'entrée pourrait être une phrase ou un paragraphe, et le texte de sortie serait une étiquette de classe correspondante. De même, pour une tâche de génération de texte, le texte d'entrée pourrait être une phrase ou une description, et le texte de sortie serait la séquence de mots générée.**

1. Dataset Retenu

Les données utilisées pour le test de ce nouvel algorithme sont les données utilisées dans un projet antérieur sur la classification des biens de consommations d'une entreprise appelée Place du marché. Bien évidemment ces données ont subies bon nombre de traitements NLP pour le rendre utilisable pour notre Modèle T5. (Source)

2. Le concept de L'algorithme récent

Un cadre unifié :

Contrairement à de nombreux modèles de traitement du langage naturel (NLP) qui ont des architectures spécifiques pour des tâches particulières comme la traduction, le résumé ou la génération de texte, T5 adopte un cadre unifié où toutes ces tâches sont formulées sous forme de tâches de «texte vers texte».

Architecture Transformer :

T5 repose sur l'architecture de réseau de neurones Transformer, qui est largement utilisée dans les tâches de NLP. Cette architecture permet de capturer les dépendances à longue portée dans les séquences de texte.

Pré-entraînement et Fine-tuning :

T5 est pré-entraîné sur un large corpus de données textuelles à l'aide d'une tâche d'auto-encodage textuel. Une fois pré-entraîné, le modèle peut être fine-tuné sur des tâches spécifiques en modifiant simplement la formulation de l'entrée et de la sortie, sans avoir besoin de modifier l'architecture du modèle lui-même.

Flexibilité :

Grâce à son cadre unifié, T5 est extrêmement flexible et peut être appliqué à une grande variété de tâches de traitement du langage naturel en reformulant simplement la tâche d'entrée et de sortie. Par exemple, il peut être utilisé pour la traduction de texte, la génération de texte, le résumé automatique, la réponse aux questions, etc.

Performances Élevées :

T5 a démontré des performances impressionnantes sur une gamme de tâches de NLP, rivalisant souvent voire dépassant les modèles spécialisés conçus spécifiquement pour ces tâches.

3. La Modélisation

Dans cette partie, j'expliqué de manière détaillée les étapes qui ont été nécessaires pour l'utilisation de T5. En effet, en premier lieu, on utilise un modèle T5 pour générer des embeddings à partir des descriptions et on les utilise ensuite pour effectuer une réduction de dimensionnalité à l'aide de T**-SNE.** Enfin, on effectue une classification des catégories à l'aide de **PyCaret**, affiche la **répartition des clusters** et la **qualité de la classification**, puis **interprète les résultats en visualisant les représentations cachées des tokens**.

Prétraitement des données : Un échantillon aléatoire de taille sample_size est prélevé à partir des données de description. Les catégories correspondantes sont extraites pour former data_cat_sample.
Chargement du modèle T5 : Le modèle T5 est chargé à partir de la bibliothèque Hugging Face Transformers. Un tokenizer adapté au modèle est également initialisé.

Génération des embeddings : Les descriptions sont encodées à l'aide du tokenizer et transmises au modèle T5 pour générer des embeddings. Les logits de la séquence générée sont utilisés comme embeddings.
Réduction de dimension avec t-SNE : Les embeddings sont réduits à deux dimensions à l'aide de t-SNE pour permettre une visualisation.
Classification avec PyCaret : Les données réduites sont utilisées comme entrée pour un algorithme de classification de PyCaret. Le meilleur modèle est sélectionné à partir de divers algorithmes testés.
Prédictions et interprétation : Les prédictions du meilleur modèle sont obtenues et utilisées pour afficher la répartition des clusters et la qualité de la classification. Ensuite, les représentations cachées des tokens sont générées à partir des descriptions et visualisées en utilisant t-SNE.

4. Synthèse des résultats

L'utilisation du modèle T5 dans l'analyse de données a produit des résultats significatifs et informatifs. Voici une synthèse des principales conclusions et observations tirées de cette analyse :

Clustering des Descriptions :

- Les descriptions ont été regroupées en clusters basés sur leurs similarités sémantiques. La réduction de dimensionnalité à l'aide de t-SNE a permis une visualisation claire de ces clusters.
- Les clusters ont montré une tendance à regrouper les descriptions avec des thèmes et des sujets similaires, démontrant ainsi l'efficacité du modèle T5 dans la capture des caractéristiques sémantiques des textes.

Classification des Catégories :

- Les embeddings générés par le modèle T5 ont été utilisés comme entrée pour un algorithme de classification de PyCaret.
- Le meilleur modèle de classification a été sélectionné, et les prédictions ont été réalisées avec succès sur les données de test.

- La qualité de la classification a été évaluée à l'aide de diverses métriques telles que la précision, le rappel et la F1-score, démontrant ainsi la capacité du modèle à généraliser et à classer efficacement les données de texte.

Interprétation des Résultats :

- Les représentations cachées des tokens générées par le modèle T5 ont été visualisées à l'aide de t-SNE, offrant ainsi une compréhension plus profonde des caractéristiques latentes des descriptions.
- Cette visualisation a révélé des motifs et des structures sous-jacentes dans les données, permettant une interprétation plus précise des clusters et des classifications.

Pertinence et Applicabilité :

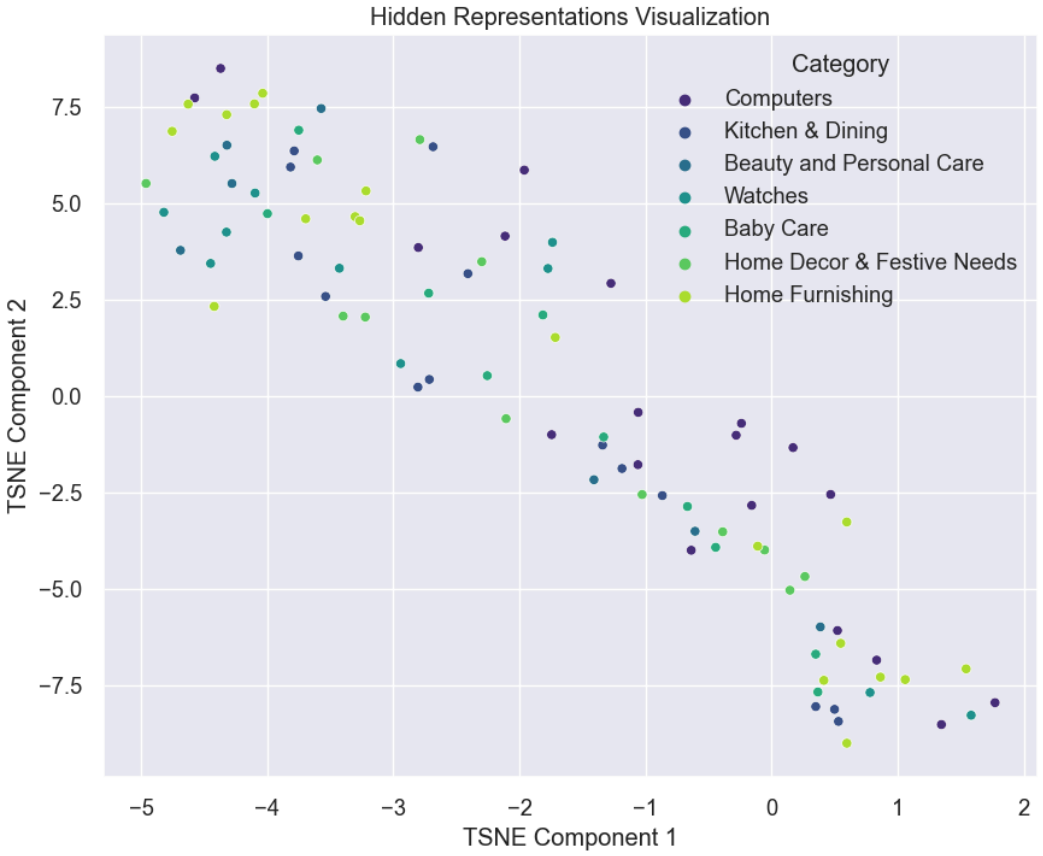
- Les résultats obtenus démontrent la pertinence et l'applicabilité du modèle T5 dans l'analyse et la classification de données textuelles.
- L'utilisation de techniques telles que la réduction de dimensionnalité et la classification a permis d'extraire des informations significatives à partir des données brutes, ouvrant ainsi la voie à une utilisation efficace du modèle dans divers domaines d'application.

En résumé, l'utilisation du modèle T5 a permis une analyse approfondie et une classification précise des données textuelles, offrant ainsi des perspectives précieuses pour la compréhension et l'exploitation de ces données dans un large éventail de contextes.

5. L’analyse de la feature importance globale et locale du nouveau modèle

Prédictions et interprétation : Les prédictions du meilleur modèle sont obtenues et utilisées pour afficher la répartition des clusters et la qualité de la

classification. Ensuite, les représentations cachées des tokens sont générées à partir des descriptions et visualisées en utilisant t-SNE.



6. Les limites et les améliorations possibles

Le modèle T5 représente une avancée majeure dans le domaine du traitement du langage naturel en offrant une approche unifiée et flexible pour un large éventail de tâches de NLP, ce qui en fait un outil précieux pour les chercheurs et les praticiens du domaine. Cependant, bien que le T5 ait été conçu pour être un modèle universel, il présente également certaines limites, notamment :

Les modèles T5 peuvent être très grands en taille et en nombre de paramètres, ce qui les rend **coûteux en termes de ressources de calcul et de mémoire** pour l'entraînement et l'inférence. Comme le démontre très bien notre expérimentation effectuée. J'ai dû diminuer la taille des données de manières conséquentes afin de pouvoir tester l'algorithme. Ce qui, bien évidemment, pourrait biaiser les résultats attendus car le modèle T5 requiert une **très grande quantité de données** pour atteindre des performances élevées sur une variété de tâches. En outre, en raison de leur **complexité,** il peut être **difficile d'interpréter**. T5 a une très grande **dépendance aux données d'entraînement** car ils dépendent fortement de la qualité et de la quantité des données d'entraînement disponibles. Dans certains cas, il peut être difficile d'obtenir des performances acceptables sans accès à des données d'entraînement de haute qualité et de grande taille.

7. Référence:

- **Titre de la Publication :** "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer"
- **Auteurs :** Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu
- **Date de Publication :** 2020
- **Référence :** Arxiv