Master 2 - Économétrie Appliquée

Natural Language Processing

Analyse et classification de documents techniques HP via NLP : cas des PC et des moniteurs

---

Dossier réalisé par :

DUPAS-BROUSSE Simon

SINEAU Angel

LE ROUX Noa

NOUHRA Elie

Année universitaire 2024-2025

SOMMAIRE

1. Introduction

Le changement climatique et la réduction de l'empreinte carbone des produits sont aujourd'hui au cœur des préoccupations dans l'industrie technologique. Face aux exigences croissantes en matière de transparence environnementale, certaines entreprises, comme **Hewlett-Packard** (HP), ont pris des initiatives notables en publiant des rapports détaillés sur l'empreinte carbone de leurs nouveaux produits[1]. Ces documents, appelés *Product Carbon Footprint Reports*, fournissent des informations précises sur les émissions de gaz à effet de serre (GES) associées à chaque produit, exprimées en équivalents CO_2 pour un horizon de 100 ans, conformément au concept de *Global Warming Potential* (GWP-100)[2]. Cette pratique permet non seulement d'informer les consommateurs soucieux de l'impact écologique de leurs achats, mais également de renforcer la responsabilité environnementale des fabricants.

Dans le cadre de ce projet, nous nous sommes intéressés à une sous-catégorie de ces documents : ceux portant sur les ordinateurs personnels (PC) et les moniteurs de la marque HP. Cette dernière est une entreprise technologique américaine fondée en 1939, qui est spécialisée et est l'un des leaders mondiaux dans la conception et la fabrication de matériel informatique, notamment des ordinateurs, des écrans et des imprimantes. L'enjeu de ce travail est de développer un modèle de classification automatique capable de distinguer, à partir du contenu textuel des rapports, s'il s'agit d'un document relatif à un PC ou à un moniteur. L'intérêt de cette démarche repose sur l'analyse de la structure linguistique et des caractéristiques techniques présentes dans les documents, avec pour objectif de mettre en évidence des motifs récurrents et des termes discriminants propres à chaque type de produit.

Pour alimenter ce modèle, nous avons d'abord constitué un corpus de documents à l'aide de techniques de **web scraping**. Cette étape a consisté à automatiser la collecte des rapports directement depuis le site de HP, en parcourant les pages contenant les *Product Carbon Footprint Reports* pour télécharger un ensemble significatif de fichiers PDF. Cette approche a permis d'agréger efficacement un volume de données suffisant

---

[1] **HP Inc.** (2023). *Product Carbon Footprint Reports*. HP Sustainable Impact. https://www.hp.com/go/carbonfootprintHP

[2] **Environmental Defense Fund.** (2022). *Global Warming Potentials (GWPs)/CO_2-equivalent (CO_2e) and the Emission Equivalency Tool*. https://www.edf.org/sites/default/files/content/emission_equivalency_tool_documentation_methodology_23062022.pdfEnvironmental Defense Fund

pour entraîner un modèle pertinent, tout en garantissant une diversité représentative des différentes familles de produits.

Une fois les données collectées, un travail de **prétraitement textuel** a été mené pour nettoyer, structurer et vectoriser les textes. Ensuite, différentes techniques de traitement du langage naturel ont été mobilisées, notamment l'utilisation de représentations vectorielles (TF-IDF), ainsi qu'un modèle de **classification supervisée** pour réaliser la tâche de prédiction.

Dans ce rapport, nous présentons tout d'abord les **données** mobilisées et les étapes de leur constitution, avant de détailler la **méthodologie** NLP mise en œuvre. Nous analyserons ensuite les **résultats** obtenus à l'aide de plusieurs métriques d'évaluation, avant de discuter des **limites** et des enseignements de cette étude. Enfin, nous proposerons quelques **perspectives d'amélioration** pour enrichir et affiner le modèle développé.

2. Présentation des données

Les données utilisées dans ce projet proviennent de rapports techniques en langue anglaise extraits directement du site officiel de HP, grâce à une procédure automatisée de web scraping. Ces documents, publiés au format PDF, décrivent de manière détaillée les caractéristiques techniques et environnementales de divers produits informatiques récents, en l'occurrence des ordinateurs personnels et des moniteurs. Ils contiennent notamment des informations sur l'empreinte carbone associée à la fabrication du produit, exprimée en équivalents CO_2.

Pour collecter ces documents, nous avons d'abord repéré les pages web listant les produits d'intérêt, puis identifié les liens pointant vers les fichiers PDF associés aux fiches techniques. À l'aide de bibliothèques Python telles que ***requests*** et ***BeautifulSoup***, nous avons automatisé la navigation sur ces pages afin d'en extraire les liens pertinents. Les fichiers PDF ont ensuite été téléchargés par script et stockés localement pour traitement.

Une fois cette collecte réalisée, les documents ont été traités à l'aide du package Python ***PyPDF2*** pour en extraire le contenu textuel brut, qui a ensuite été organisé dans un jeu de données structuré. Chaque observation correspond à un fichier PDF unique, dont on a extrait le texte, et auquel une étiquette a été attribuée selon la catégorie du produit décrit, soit « PC » pour les ordinateurs, soit « Monitors » pour les écrans.

Le corpus obtenu compte 530 documents, chacun représentant un produit distinct. La vectorisation a automatiquement intégré certains traitements standards, comme la mise en minuscules et la suppression des stopwords (mots vides).

Afin de convertir ces textes en une représentation numérique exploitable par des algorithmes de machine learning, la méthode TF-IDF (Term Frequency-Inverse Document Frequency)[3] a été utilisée. Elle permet d'associer à chaque document un vecteur de poids reflétant l'importance relative de chaque mot ou groupe de mots dans le texte, tout en tenant compte de sa fréquence dans l'ensemble du corpus. Dans ce projet, la vectorisation a été réalisée en intégrant à la fois des unigrams (mots simples) et des bigrams (groupes de deux mots consécutifs)[4], ce qui permet de mieux capter les expressions techniques spécifiques à chaque type de produit.

[3] **Capital One.** (2021). *Understanding TF-IDF for Machine Learning*. https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/Capital One

[4] **Jurafsky, D., & Martin, J. H.** (2023). *Speech and Language Processing (3rd ed. draft)*. Stanford University. https://web.stanford.edu/~jurafsky/slp3/3.pdf

Le jeu de données final se compose donc de vecteurs numériques représentant chacun des 530 documents, accompagnés de leurs étiquettes respectives, prêts à être utilisés pour l’apprentissage supervisé d’un modèle de classification.

3.Analyse et interprétation des résultats

Suite aux étapes de collecte, de prétraitement et de vectorisation du corpus des 530 rapports techniques HP en langue anglaise, représentant des ordinateurs personnels et des moniteurs, le processus d'analyse s'est concentré sur la construction et l'évaluation d'un modèle de classification automatique. L'objectif était de discriminer ces deux catégories de documents basées uniquement sur leur contenu textuel, représenté sous forme vectorielle TF-IDF intégrant unigrams et bigrams. Un modèle de régression logistique a été sélectionné pour cette tâche de classification supervisée. Le jeu de données a été partitionné en ensembles d'entraînement et de test, ce dernier représentant 20% des documents, afin d'évaluer la capacité de généralisation du modèle sur des données non utilisées durant l'apprentissage.

3.1 Évaluation de la performance du modèle

L'évaluation du modèle de régression logistique a été réalisée sur le jeu de test, qui comprenait 106 documents (50 rapports de moniteurs et 56 rapports de PC). Les performances sont résumées dans le rapport de classification suivant :

Tableau 1 : Indicateurs de performance du modèle

Classe	Précision	Rappel	Score F1	Support
monitors	0.96	1.00	0.98	50
pc	1.00	0.96	0.98	56
accuracy			**0.98**	106
macro avg	0.98	0.98	0.98	106
weighted avg	0.98	0.98	0.98	106

Ces résultats témoignent d'une **performance remarquable** du modèle de régression logistique pour la tâche de classification des rapports techniques en deux catégories.

L'**exactitude globale (accuracy)**, qui mesure la proportion de prédictions correctes sur l'ensemble du jeu de test, atteint un niveau très élevé de **0.98**, soit 98%. Cela signifie que sur les 106 documents évalués, 104 ont été correctement attribués à leur catégorie respective.

En examinant les métriques par classe, on observe une **précision** de 0.96 pour la classe *'monitors'*. Cela indique que sur tous les documents que le modèle a identifiés comme étant des rapports de moniteurs, 96% étaient effectivement des moniteurs. Le **rappel (recall)** pour cette même classe est de 1.00. C'est un résultat particulièrement fort, car il signifie que **le modèle a réussi à identifier la totalité des 50 rapports de moniteurs présents** dans le jeu de test. Pour la classe '*pc'*, la **précision** est parfaite, atteignant 1.00. Cela implique qu'**aucun document classé comme un PC par le modèle n'était en réalité un moniteur**.

Le **rappel** pour la classe '*pc*' est de 0.96, indiquant que 96% des 56 rapports de PC réels du jeu de test ont été correctement capturés par le modèle. Le **score F1**, qui représente la moyenne harmonique de la précision et du rappel, est uniformément élevé (0.98) pour les deux classes, confirmant un excellent équilibre entre la capacité du modèle à ne pas faire d'erreurs de type "faux positifs" (précision) et sa capacité à ne pas manquer les instances de la classe (rappel).

L'ensemble de ces métriques, y compris les moyennes (macro et pondérée), convergent pour attester de l'efficacité du modèle à distinguer les deux types de documents techniques en se basant sur leur contenu textuel seul.

3.2 Identification des termes clés discriminants

Au-delà de l'évaluation quantitative de la performance, il est crucial de comprendre *comment* le modèle parvient à une telle distinction. La régression logistique attribue des poids (coefficients) aux caractéristiques d'entrée (les mots ou n-grammes représentés par les valeurs TF-IDF) pour réaliser sa classification.

Les coefficients positifs élevés sont associés aux termes qui augmentent la probabilité qu'un document appartienne à la classe '*pc*', tandis que les coefficients négatifs bas sont associés aux termes qui augmentent la probabilité qu'il appartienne à la classe 'monitors'. L'analyse de ces coefficients permet d'identifier les éléments textuels les plus pertinents pour la classification.

Voici les cinq termes (unigrams ou bigrams) ayant les coefficients les plus élevés (les plus indicateurs de la classe 'pc') et les cinq termes ayant les coefficients les plus bas (les plus indicateurs de la classe 'monitors') :

Tableau 2 : Top 5 des termes les plus discriminants pour la classe 'pc'

Terme	Coefficient
desktop	2.6769
drive	1.9693
computer specifications	1.6418
computer	1.6204
pc	1.5988

Tableau 3 : Top 5 des termes les plus discriminants pour la classe 'monitors'

Terme	Coefficient
display	-3.0689
monitor	-2.0430
display specifications	-1.6737
average display	-1.6737
inches product	-1.2052

L'examen de ces tableaux valide intuitivement la pertinence des termes identifiés par le modèle comme étant les plus discriminants. Pour la classe 'pc' (Tableau 1), les termes les plus influents font directement référence au type d'appareil ("desktop", "computer", "pc") ou à des composants et spécifications techniques qui leur sont intrinsèquement liés ("drive", "computer specifications").

Ces termes sont couramment utilisés dans la description des ordinateurs personnels et de leurs capacités techniques. Inversement, pour la classe 'monitors' (Tableau 2), les termes les plus discriminants sont ceux qui décrivent les caractéristiques d'un écran, telles que le type d'affichage ("display", "monitor"), les spécifications associées ("display specifications", "average display"), ou encore des unités de mesure couramment employées pour les écrans ("inches product").

L'analyse confirme de manière explicite, comme l'indique la source, que **"desktop" est le mot le plus discriminant pour les PC**, avec le coefficient positif le plus élevé, et **"display" est le mot le plus discriminant pour les écrans**, avec le coefficient négatif le plus bas (en valeur absolue, le plus grand coefficient négatif).

Cette analyse des termes discriminants offre une explication claire et fondée sur les données de la manière dont le modèle a appris à séparer les deux catégories, en s'appuyant sur les distinctions lexicales propres à chaque type de produit technique.

4. Conclusion

Ce projet a démontré avec succès la faisabilité de développer un modèle de classification automatique basé sur le contenu textuel des rapports techniques de produits informatiques, en l'occurrence les PC et les moniteurs de la marque HP. Grâce à l'utilisation de techniques de traitement automatique des langues, notamment la vectorisation TF-IDF et la régression logistique, nous avons atteint une exactitude de classification remarquable de 98% sur un jeu de test représentatif. Cette performance atteste de la capacité du modèle à identifier avec précision les caractéristiques linguistiques et techniques distinctives de chaque catégorie de produits.

L'analyse des termes discriminants a également fourni des insights précieux sur les éléments textuels qui contribuent le plus à la distinction entre les PC et les moniteurs, confirmant ainsi la pertinence du contenu textuel pour la classification. Cette étude contribue au domaine du NLP en démontrant son applicabilité dans le contexte spécifique des documents techniques liés aux produits informatiques.

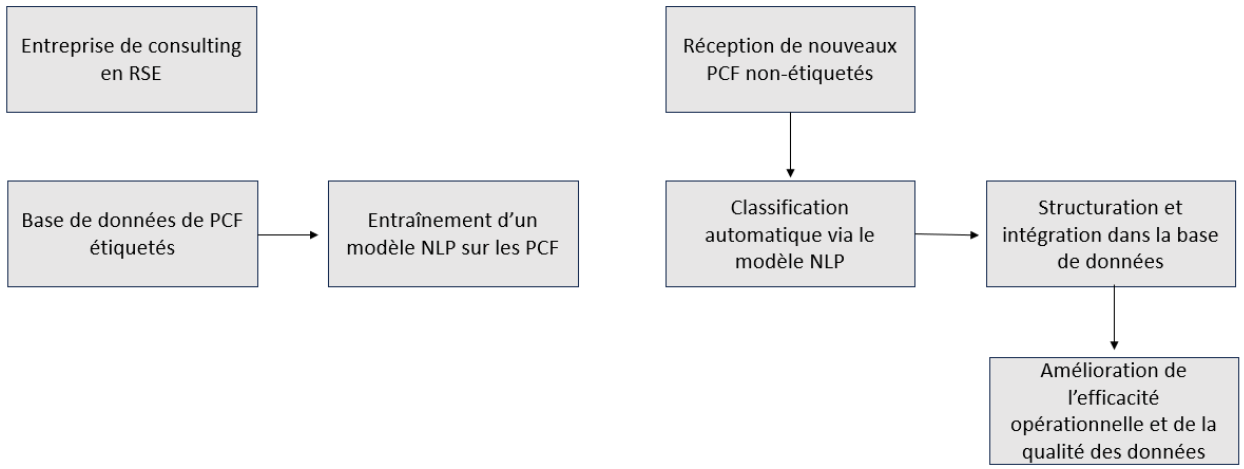
Bien que les résultats soient prometteurs, certaines limites doivent être considérées. Premièrement, l'étude est centrée sur une seule marque (HP) et deux catégories de produits (PC et moniteurs), ce qui pourrait limiter la généralisation des conclusions à d'autres marques ou catégories. Deuxièmement, l'approche adoptée repose sur des techniques de NLP qui, bien que performantes, pourraient être améliorées ou complétées par d'autres méthodes, telles que l'apprentissage profond.

Ce projet ouvre des perspectives intéressantes pour les entreprises spécialisées dans le consulting en Responsabilité Sociale des Entreprises (RSE) et les analyses environnementales. Un scénario pertinent illustrant l'intérêt de ce projet pourrait être le suivant :

Une entreprise de consulting en RSE dispose d'une vaste base de données de Product Carbon Footprint (PCF) déjà classés et étiquetés. En entraînant un modèle NLP sur cette base, comme nous l'avons fait dans ce projet, l'entreprise peut automatiser la classification de nouveaux PCF reçus, structurant ainsi efficacement ses données et automatisant entièrement le processus. Ce scénario met en lumière l'intérêt pratique de notre étude, qui peut contribuer à améliorer l'efficacité opérationnelle et la qualité des données dans le domaine du consulting en RSE.

Le diagramme ci-dessous illustre les différentes étapes de ce processus automatisé, depuis la base de données initiale jusqu’à l’intégration des nouveaux PCF classifiés.

Figure 1 : Optimisation des processus RSE grâce au NLP



En outre, l'extension de cette approche à d'autres types de documents techniques ou à d'autres secteurs industriels pourrait enrichir davantage le champ d'application du NLP et offrir des solutions innovantes pour la gestion et l'analyse de grands volumes de données textuelles. L'intégration de techniques d'apprentissage automatique plus avancées pourrait également permettre d'accroître les performances de classification et d'élargir les capacités d'analyse.

Ainsi, ce projet non seulement démontre la valeur du NLP pour la classification automatique de documents techniques mais ouvre également la voie à de futures recherches et applications dans divers domaines où la gestion efficace de l'information textuelle est cruciale.

Bibliographie

HP Inc. (2023). *Product Carbon Footprint Reports*. HP Sustainable Impact. https://www.hp.com/go/carbonfootprintHP

HP Inc. (2021). *HP Carbon Accounting Manual*. https://h20195.www2.hp.com/v2/GetDocument.aspx?docname=c08951350HP Support+6HP Support+6HP Support+6

U.S. Environmental Protection Agency (EPA). (2025). *Understanding Global Warming Potentials*. https://www.epa.gov/ghgemissions/understanding-global-warming-potentialsUS EPA

Greenhouse Gas Protocol. (2011). *Product Life Cycle Accounting and Reporting Standard*. https://ghgprotocol.org/sites/default/files/standards/Product-Life-Cycle-Accounting-Reporting-Standard_041613.pdfGHG Protocol

Environmental Defense Fund. (2022). *Global Warming Potentials (GWPs)/CO_2-equivalent (CO_2e) and the Emission Equivalency Tool*. https://www.edf.org/sites/default/files/content/emission_equivalency_tool_documentation_methodology_23062022.pdfEnvironmental Defense Fund

Built In. (2024). *TF-IDF: An Introduction*. https://builtin.com/articles/tf-idfBuilt In

Capital One. (2021). *Understanding TF-IDF for Machine Learning*. https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/Capital One

GeeksforGeeks. (2024). *Removing Stop Words with NLTK in Python*. https://www.geeksforgeeks.org/removing-stop-words-nltk-python/GeeksforGeeks

Répertoire Github contenant les codes du projet : https://github.com/simdups/NLP