Praxy.ai

Classification automatique de documents médicaux : Performance comparée de notre système hybride et de modèles de langues (LLM)

Solène Degrutère, Nadège Alavoine, Damien Forest

Résumé

Dans le contexte de l'accroissement constant du volume de documents médicaux, la classification de ces documents constitue une tâche cruciale. Nous proposons un système hybride combinant règles, pondérations et reconnaissance d'entités nommées spécifiques au domaine médical pour classifier ces documents en huit catégories distinctes. Afin d'évaluer la pertinence de notre approche, nous avons comparé les performances de notre système à celles de grands modèles de langues (LLM) sur un corpus de 168 documents médicaux. Les résultats montrent que notre approche hybride atteint un score d'exactitude de 91% sur l'ensemble des catégories, dépassant de manière significative les performances des LLM seuls sur ce type de tâches.

Abstract

Faced with an ever-increasing volume of medical documents, classification is an important task. We propose a hybrid system combining rules, weighting and named entity recognition specific to the medical domain to classify these documents into eight distinct categories. To evaluate the relevance of our approach, we compared the performance of our system with that of large language models (LLM) on a corpus of 168 medical documents. The results show that our hybrid approach achieves an accuracy score of 91% across all categories, significantly outperforming LLM.

Mots-clés : Classification de documents médicaux, méthode à base de règle, entités nommées médicales, LLM

Keywords: medical document classification, ruler-based method, medical named entities, LLM

Introduction

La gestion des documents médicaux représente un enjeu majeur pour les établissements de santé, notamment avec l'augmentation continue du volume de

documents à traiter (Cabannes *et al.*, 2023). Avec cette croissance constante il est essentiel de mettre en place un système de classification des documents médicaux afin d'assurer une gestion efficace des informations de santé.

Face à ce défi, les méthodes de classification manuelles, nécessitant un examen minutieux de chaque document, peut se révéler chronophage. S'il est envisageable d'entraîner un classificateur par apprentissage supervisé dédié à cet usage, celui-ci nécessiterait un grand volume de données difficile à collecter dans un domaine aussi sensible que la santé. Les LLM représentent une autre option, notamment grâce à leurs capacités de raisonnement *zero-shot* (Kojima *et al.*, 2023). Cependant il est important de mentionner que l'utilisation des LLM implique des coûts écologiques et économiques importants, ainsi qu'un temps d'inférence non négligeable (Bender *et al.*, 2021).

Dans ce contexte, nous proposons une solution combinant une approche à base de règle, un système de pondérations et d'extraction d'entités nommées, permettant de classifier un document médical dans une des huit catégories les plus fréquemment rencontrées. Cette solution permet de s'affranchir du besoin d'un grand nombre de documents, tout en étant plus rapide et écologique que l'usage d'un LLM. L'objectif principal de cette étude est double : d'une part, évaluer la performance de notre système hybride sur un corpus de documents médicaux, et d'autre part, comparer ces résultats avec ceux obtenus par des LLM afin de déterminer quelle approche offre le meilleur compromis.

Pour atteindre ces objectifs, nous avons adopté une démarche méthodologique structurée. Premièrement nous établirons la tâche de classification envisagée, puis nous présenterons le corpus de documents médicaux utilisé. Ensuite nous détaillerons notre approche hybride de classification et la méthode de comparaison avec les LLM, ainsi que les métriques d'évaluation retenues. Nous poursuivrons par l'analyse des résultats obtenus et conclurons par une discussion.

Méthodologie

1. Détermination de la tâche de classification

En collaboration avec des professionnels de santé, nous avons identifié les huit catégories les plus fréquentes de documents médicaux. La problématique étudiée se présente donc sous la forme d'une classification multiclasse. Pour chacune de ces catégories, une étiquette spécifique a été attribuée :

- BIOLOGY: Documents relatifs aux analyses biologiques et comptes-rendus de laboratoire de biologie.

- CERTIFICATE: Certificats médicaux et documents attestant de l'état de santé d'un individu.
- IMAGING: Documents relatifs aux comptes-rendus d'imagerie médicale.
- DISCHARGE: Lettre de sortie d'hospitalisation et document de liaison associé.
- LETTER: Courrier d'adressage ou de liaison entre professionnels de santé.
- MEDICAL_NOTE: Comptes-rendus de consultation.
- SURGERY: Comptes-rendus opératoires.
- SYNTHESIS: Volet de synthèse médicale.

Les documents médicaux ne correspondant à aucune de ces catégories sont catégorisés comme appartenant à une catégorie supplémentaire : « NOT_CLASSIFIED ».

2. Données utilisées

Notre démarche repose sur un corpus de 168 documents médicaux anonymisés, chacun correspondant potentiellement à l'une des huit catégories identifiées en collaboration avec des professionnels de santé. Ces documents ont été complétés par 95 documents médicaux (incluant notamment des modèles de documents médicaux) et non médicaux anonymisés ne pouvant pas rentrer dans ces catégories. L'ensemble du corpus couvre l'intégralité des catégories établies, avec une répartition naturelle, c'est-à-dire une distribution qui reflète fidèlement la diversité et la fréquence des cas observés en milieu hospitalier.

Ces documents au format PDF sont convertis au format texte grâce à une approche combinant l'utilisation de la librairie python PyMuPDF[1], pour l'extraction directe du contenu textuel, et d'un système de reconnaissance optique de caractères (OCR) Cette conversion permet de conserver la structure et la mise en forme des documents tout en les rendant exploitables pour nos algorithmes.

3. Approche

Notre approche se compose de deux axes principaux : une méthode hybride combinant plusieurs niveaux d'analyse et une approche par LLM.

3.1 Approche hybride

Le système que nous proposons repose sur une classification hybride qui s'articule en trois étapes complémentaires.

[1] https://pymupdf.readthedocs.io/en/latest/

3.1.1 Détection de mots-clés critiques

Cette première étape repose sur l'identification des marqueurs linguistiques spécifiques fortement discriminants. Par exemple, pour la catégorie « SURGERY » s'il y a le mot-clé *« compte rendu opératoire »* alors le document est directement catégorisé comme appartenant à la classe « SURGERY».

3.1.2 Reconnaissance pondérée de mots-clés

La deuxième phase utilise un système de pondérations basé sur des mots-clés plus généraux. Chaque terme se voit attribuer un poids en fonction de la classe à laquelle il peut appartenir.

3.1.3 Analyse d'Entités Nommées

La dernière phase exploite un modèle GLiNER (Zaratiana *et al.*, 2023). Les modèles GLiNER s'appuient sur une architecture à base de Transformers (Vaswani *et al.*, 2017) bidirectionnels et permettent de réaliser une reconnaissance d'entités nommées. Nous avons utilisé un variant de ce modèle qui soit adapté au domaine médical et à la langue française.

3.2 Approche par LLM

Pour évaluer notre approche, nous avons effectué la même tâche en utilisant deux LLM : Llama-3.1-8B-Instruct[2] (Grattafiori *et al.*, 2024), et Phi-3-mini-4k-Instruct[3] (Abdin *et al.*, 2024). Nous avons sélectionné ces modèles en raison de leur taille réduite, qui se traduit par une plus grande efficacité économique, une empreinte écologique moindre et des temps d'inférence plus courts. Par ailleurs, ils ont tous deux étés entraînés en partie sur des données en langue française, ce qui était un critère important dans le choix du modèle.

Après des expérimentations préliminaires, les hyperparamètres utilisés lors de l'inférence ont été fixés à une température de 0.1, identifiée comme la plus stable pour notre tâche de classification, ainsi qu'une limitation du nombre de *tokens* attendus en sortie à 200.

Le modèle reçoit en entrée le texte du document à classifier, précédé d'un prompt spécifique basé sur la méthode du *few-shot prompting* (Brown *et al.*, 2020) : Des exemples courts et représentatifs de textes et classifications associées ont été fournis dans le prompt. Cette approche permet au modèle de générer des prédictions plus précises pour cette tâche de classification.

[2] meta-llama/Llama-3.1-8B-Instruct · Hugging Face
[3] microsoft/Phi-3-mini-4k-instruct · Hugging Face

4. Métriques

Pour évaluer et comparer les performances des deux approches de classification, nous avons utilisé plusieurs métriques. Dans un premier temps, l'exactitude (*accuracy* en anglais) a été calculée pour mesurer la proportion globale de documents correctement classifiés. Cependant, en raison du déséquilibre des catégories au sein de notre corpus, nous avons également pris en compte les moyennes pondérées des scores de précision, de rappel et de score F1.

Résultats

Nous présentons, dans le tableau 1, une comparaison détaillée des performances obtenues par notre approche hybride et les LLM Llama-3.1-8B-Instruct et Phi-3-mini-4k-Instruct sur notre corpus de documents médicaux.

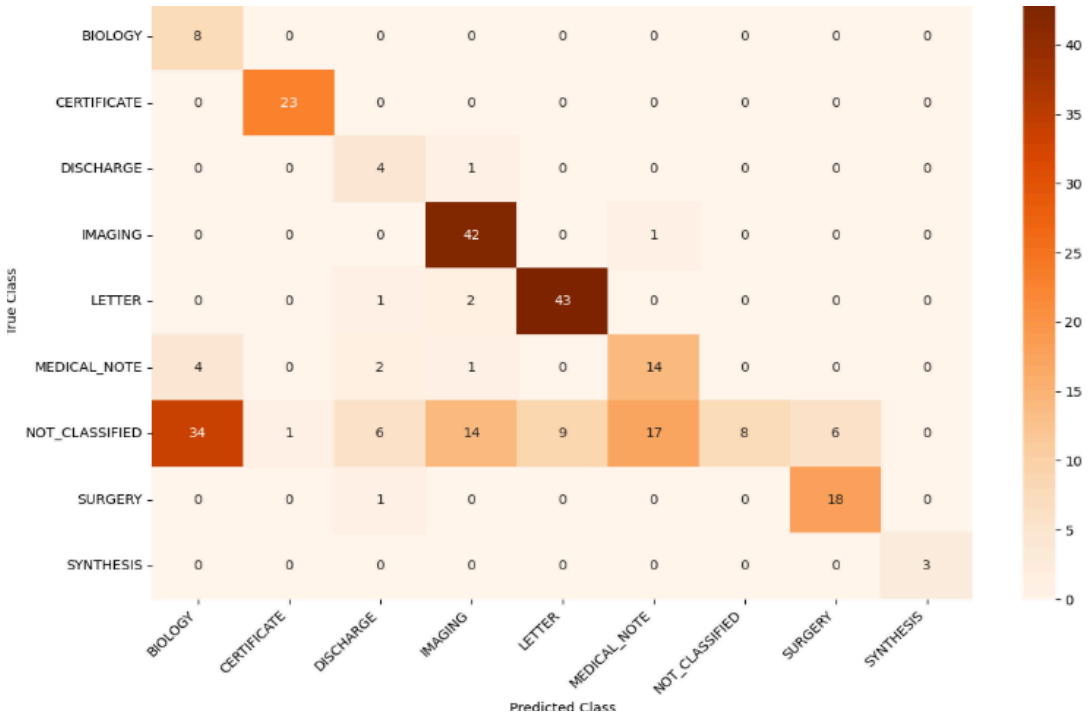
Les métriques, telles que l'exactitude, la précision, le rappel et le score F1, sont indiquées pour chaque approche, en prenant en compte ou non la catégorie « NOT_CLASSIFIED ». Étant donné la nature non-déterministe des LLM (Song *et al.*, 2024), nous avons effectué trois séries d'expérimentations pour chaque modèle sur l'ensemble des données et retenu le meilleur résultat global des trois séries pour chaque LLM.

	Approche hybride				LLM Llama-3.1-8B-Instruct				LLM Phi-3-mini-4k-Instruct			
	Ex.	Pr.	Ra.	F1	Ex.	Pr.	Ra.	F1	Ex.	Pr.	Ra.	F1
Avec catégorie NOT_CLASSIFIED	0.62	0.82	0.62	0.56	0.32	0.55	0.32	0.25	0.37	0.23	0.37	0.22
Sans catégorie NOT_CLASSIFIED	0.91	0.92	0.91	0.91	0.24	0.77	0.24	0.33	0.02	0.32	0.02	0.03

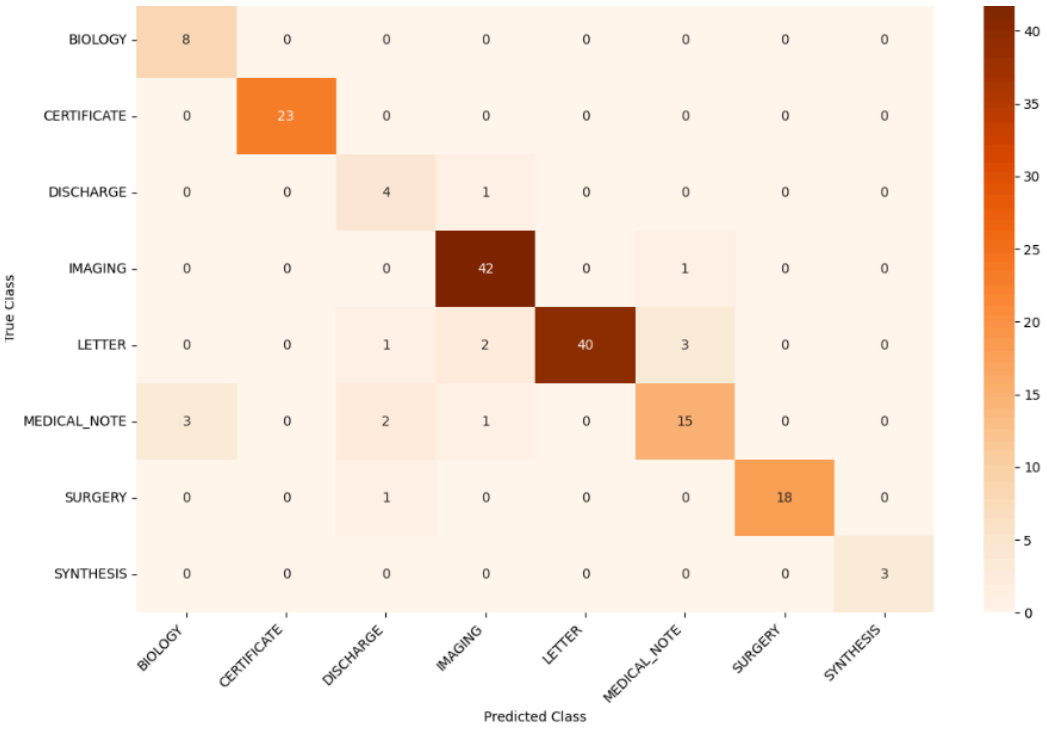
Tableau 1. Comparaison des deux approches (hybride et par LLM en *few-shot prompting*). Pour chaque approche, l'exactitude (Ex.), et la moyenne pondérée des scores de précision (Pr.), de rappel (Ra.) et F1 sont mesurées.

Les résultats obtenus montrent de meilleures performances pour notre approche hybride comparée à l'approche par LLM, tant en présence qu'en l'absence de la catégorie « NOT_CLASSIFIED ». Lorsque nous incluons cette catégorie, notre approche atteint un score d'exactitude de 62%, contre seulement 32% et 37% pour les modèles Llama-3.1-8B-Instruct et Phi-3-mini-4k-Instruct respectivement. En excluant cette catégorie, notre approche atteint 91% d'exactitude, tandis que les LLM chutent à 24% et 2%. Ces tendances sont également visibles dans les scores de précision, rappel et score F1, tant avec ou sans la catégorie « NOT_CLASSIFIED ».

Pour une analyse plus approfondie des erreurs de classification, les matrices de confusion correspondantes sont illustrées en Figures 1, 2 et 3.

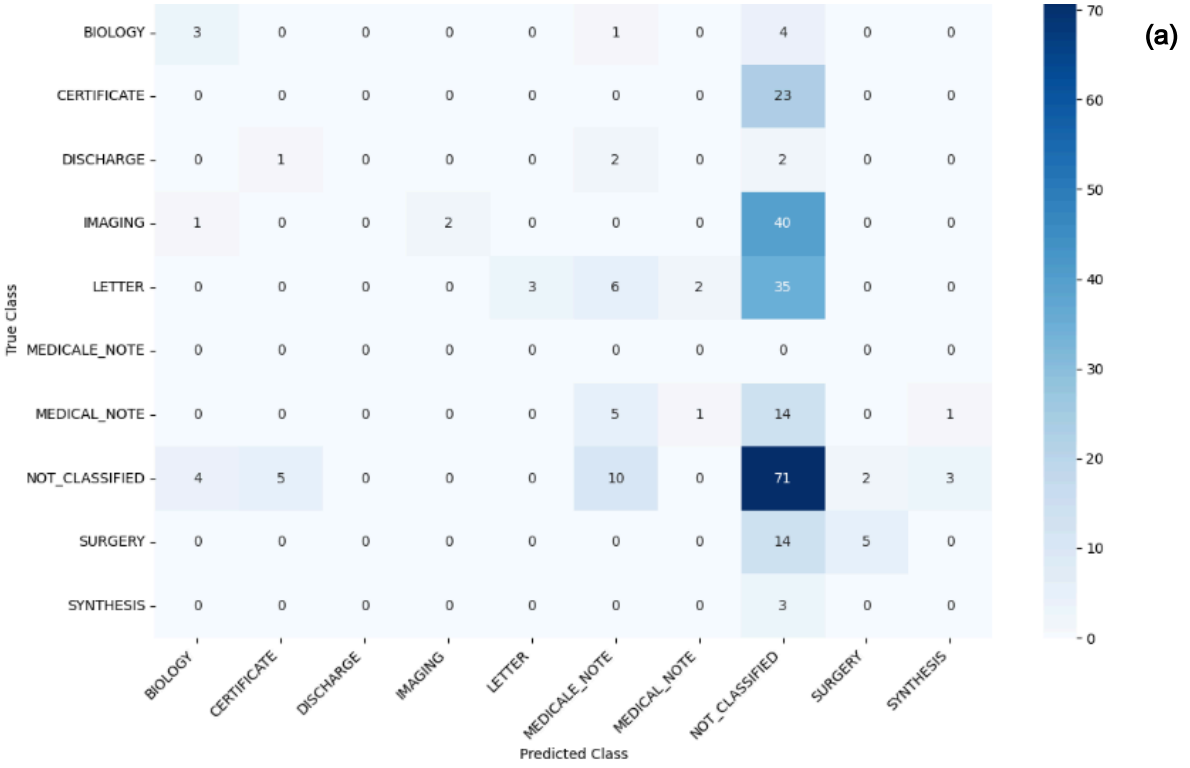


(a)



(b)

Figure 1. Matrice de confusion de la méthode hybride avec catégorie « NOT_CLASSIFIED » (a) et sans catégorie « NOT_CLASSIFIED » (b). La classe « NOT_CLASSIFIED » n'apparaît pas dans la seconde matrice car aucun document n'appartenant pas à cette catégorie n'est classifié en tant que tel.



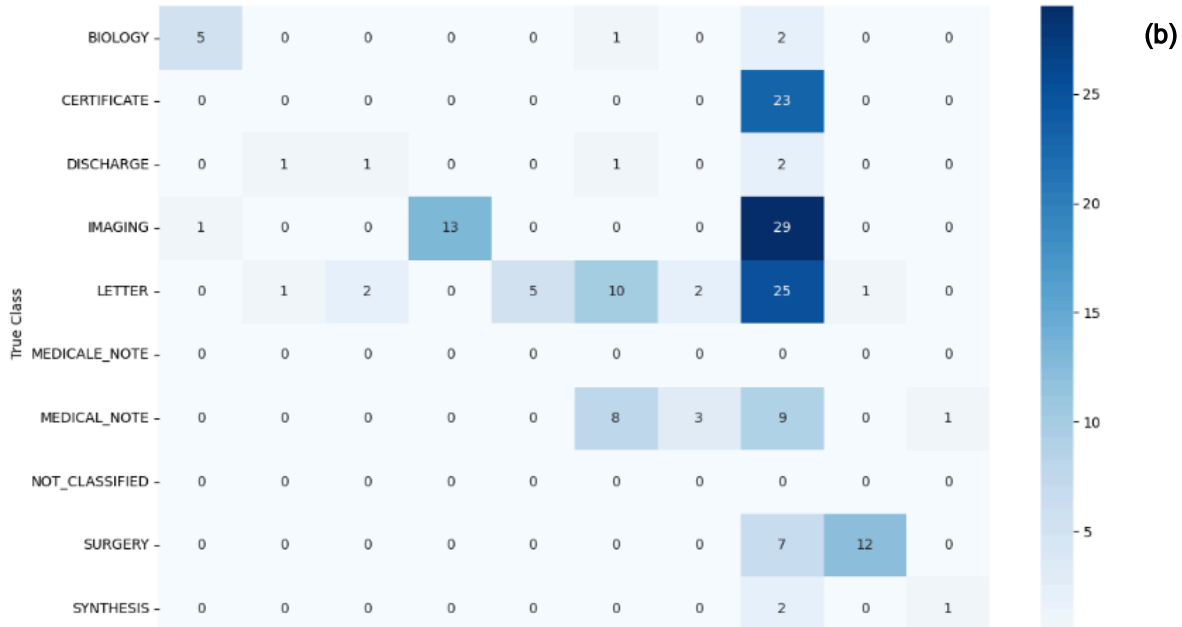
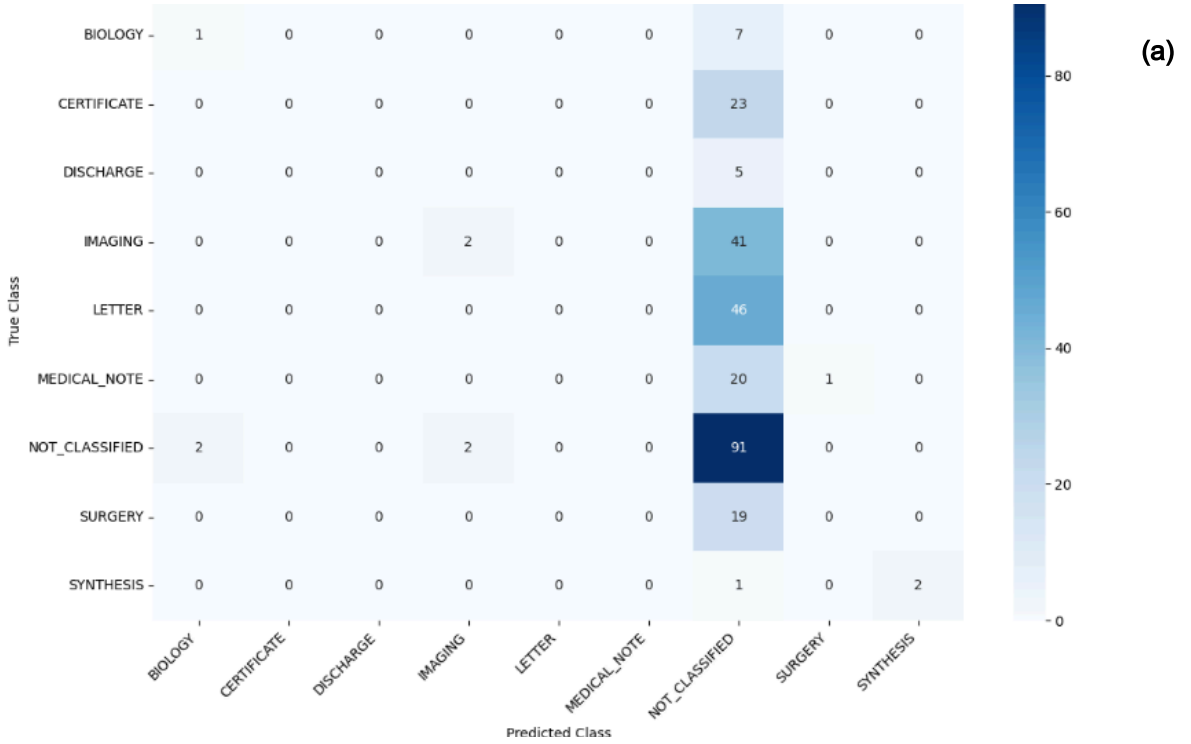


Figure 2. Matrice de confusion de la méthode LLM avec Llama-3.1-8B-Instruct avec catégorie « NOT_CLASSIFIED » (a) et sans catégorie « NOT_CLASSIFIED » (b).



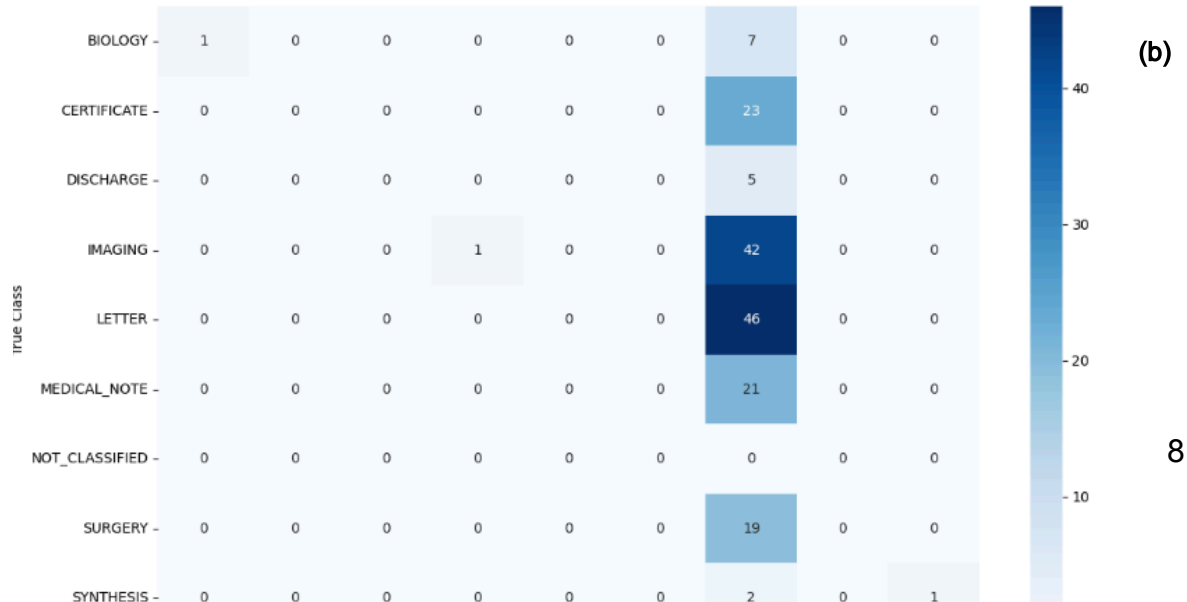


Figure 3. Matrice de confusion de la méthode LLM avec Phi-3-mini-4-instruct avec catégorie « NOT_CLASSIFIED » (a) et sans catégorie « NOT_CLASSIFIED » (b).

L'analyse des matrices de confusion révèle que les LLM ont une forte propension à classer un grand nombre de documents dans la catégorie « NOT_CLASSIFIED ». Cette tendance explique l'amélioration de leurs performances lorsque cette catégorie est prise en compte dans l'évaluation. A l'inverse, notre approche hybride obtient de meilleurs résultats lorsque cette catégorie est exclue de l'évaluation, car cette catégorie contient beaucoup de documents tels que des modèles de documents médicaux qui n'échappent pas aux règles de classification mises en place.

Les performances de notre approche hybride dans la tâche de classification des documents médicaux, même en excluant la catégorie « NOT_CLASSIFIED », peut être expliquée par plusieurs facteurs :

- Exploitation de marqueur spécifique : Notre système exploite des règles rédactionnelles fréquemment rencontrées dans les différentes catégories de documents médicaux.
- Système de pondération dépendant de la classe: Cette pondération permet d'ajuster la sensibilité de notre système selon les différentes catégories
- Extraction d'entités nommées: Le module d'extraction d'entités nommées capture de manière efficace la terminologie médicale.

En revanche, bien que les LLM possèdent de solides capacités de compréhension du langage, ils n'ont pas été spécifiquement optimisés pour le traitement de documents médicaux structurés. Cette différence explique l'écart de performances particulièrement notable sur les documents appartenant à des catégories établies.

Discussions

Les résultats obtenus confirment l'efficacité de notre approche hybride. En intégrant des connaissances spécifiques au domaine médical, notre système parvient à classifier efficacement les documents. À l'inverse, les performances très limitées des LLM, notamment sur les documents appartenant à des classes connues, mettent en évidence les défis auxquels sont confrontés les modèles de langues généralistes lorsqu'ils sont appliqués à des tâches spécialisées.

Notre approche, bien qu'efficace, présente encore des faiblesses. Cependant, l'un de ses avantages réside dans sa capacité d'adaptation. En effet, selon les retours des utilisateurs nous pouvons ajuster les règles de classification ainsi que la pondération mise en place. Cette flexibilité permet d'améliorer continuellement les performances de notre approche, un avantage significatif par rapport aux approches par LLM.

Conclusion

Notre étude comparative démontre l'efficacité de notre approche hybride combinant une méthode à base de règles, un système de pondération et l'extraction d'entités nommées pour la classification automatique de documents médicaux. Face à la complexité du langage médical, cette approche dépasse de manière significative les performances obtenues par des LLM généralistes. Au-delà de ses performances, notre approche hybride présente d'autres avantages. En effet, notre approche est plus rapide, plus économique et plus écologique que des LLM de petites tailles. De plus, elle ne requiert pas l'accès à d'importantes quantités de données d'entraînement, un atout à prendre en compte, notamment dans le domaine médical où les données annotées sont souvent limitées.

Ces aspects soulignent alors que les approches hybrides, combinant l'expertise humaine et les techniques d'apprentissage automatique, offrent une solution équilibrée pour relever le défi de la classification de documents.

Bibliographie

Abdin, M. *et al.* (2024). Phi-3 Technical Report: A highly capable language model locally on your phone. ArXiv.

Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?🦜. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).

Brown T. *et al.* (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (pp. 1877–1901).

Cabannes X., Debout C., Andarelli J.M., Dupont M. (2023). Les évolutions des dossiers médicaux des établissements de santé. Journal de droit de la santé et de l'assurance maladie, 2023, 17.

Grattafiori A. *et al.* (2024). The Llama 3 Herd of Models. ArXiv.

Kojima T., Gu S.S., Reid M., Matsuo Y., Iwasawa Y. (2023). Large language models are zero-shot reasoners. In Proceedings of the 36th International Conference on Neural Information Processing Systems.

Song Y., Wang G., Li S., Lin B. (2024). The Good, The Bad, and The Greedy: Evaluation of LLMs Should Not Ignore Non-Determinism. ArXiv.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. (2017). Attention is All you Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 6000-6010).

Zaratiana, U., Tomeh, N., Holat, P., Charnois, T. (2023). GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 5364-5376).