# Classification multilabel de concepts médicaux pour l'identification du profil clinique du patient

Christel Gérardin, Pascal Vaillant, Perceval Wajsbürt, Clément Gilavert, Ali Bellamine, Emmanuelle Kempf, Xavier Tannier

*Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles*
*Lille, France, 28 juin au 2 juillet 2021*

# Classification multilabel de concepts médicaux pour l'identification du profil clinique du patient

Christel Gérardin[1, 3] Pascal Vaillant[2, 4] Perceval Wajsbürt[2, 5] Clément Gilavert[6]
Ali Bellamine[1] Emmanuelle Kempf[1, 7] Xavier Tannier[2, 5]

(1) Assistance Publique – Hôpitaux de Paris, prenom.nom@aphp.fr
(2) Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en eSanté (LIMICS)
(3) Institut Pierre Louis d'Epidémiologie et de Santé Publique, Sorbonne Université, Inserm, 27 rue Chaligny, 75012 PARIS, christel.ducroz-gerardin@iplesp.upmc.fr
(4) Université Sorbonne Paris Nord, F-93000, Bobigny, France, vaillant@univ-paris13.fr
(5) Sorbonne Université, prenom.nom@sorbonne-universite.fr
(6) CG Conception, c.gilavert@cg-conception.fr
(7) Département d'oncologie médicale, hôpital Henri Mondor et Albert Chenevier, Créteil, France

## RÉSUMÉ

La première tâche du Défi fouille de textes 2021 a consisté à extraire automatiquement, à partir de cas cliniques, les phénotypes pathologiques des patients regroupés par tête de chapitre du MeSH-maladie. La solution présentée est celle d'un classifieur multilabel basé sur un transformer. Deux transformers ont été utilisés : le camembert-large classique (run 1) et le camembert-large *fine-tuné* (run 2) sur des articles biomédicaux français en accès libre. Nous avons également proposé un modèle « bout-en-bout », avec une première phase d'extraction d'entités nommées également basée sur un transformer de type camembert-large et un classifieur de genre sur un modèle Adaboost. Nous obtenons un très bon rappel et une précision correcte, pour une F1-mesure autour de 0,77 pour les trois runs. La performance du modèle « bout-en-bout » est similaire aux autres méthodes.

## ABSTRACT

**Multilabel classification of medical concepts for patient's clinical profile identification**

This year, the first task of the French Text Mining Challenge consisted in automatically extracting the pathological phenotypes of patients from clinical texts, grouped by head's chapter of the MeSH, disease-section. Benefiting from the annotations of previous years. The solution presented is a multilabel classifier based on a transformer. Two transformers were used : the classic Camembert-large (run 1) and a Camembert-large fine tuned on French biomedical articles in free access. We have also proposed an "end-to-end" model, with a first phase of named entity recognition also based on a transformer and a gender information extracted via an Adaboost model. The results obtained are as follows : run 1 : recall = 0.874, precision = 0.696 and F1-measure = 0.775, run 2 : F1-Measure = 0.771 and run 3 : 0.770. The performance of the end-to-end model is comparable to that of other methods.

MOTS-CLÉS : classification multilabel, Transformer, extraction d'entités nommées, concepts médicaux.

KEYWORDS: multi-label classification, Transformer, named entity recognition, medical concepts.

Atelier DÉfi Fouille de Textes (DEFT), pages 21–30.

# 1 Introduction et données utilisées

L'extraction automatisée des caractéristiques cliniques des patients à partir des comptes rendus médicaux est devenue un enjeu majeur en données médicales depuis l'apparition du dossier patient informatisé. Dans ce contexte, le défi fouille de textes 2021 (Grouin *et al.*, 2021), a organisé une épreuve d'extraction d'entités nommées à partir de cas cliniques.

Les données d'entraînements fournies par le DEFT 2021 regroupent un ensemble de 167 cas-cliniques comprenant les annotations des années précédentes (DEFT 2019 et DEFT 2020), en particulier les entités de type *signe ou symptôme* et *pathologie* avec leurs caractéristiques (négation, hypothèse, lien avec une personne autre que le ou la patiente). L'objectif de la tâche est de réaliser un phénotypage pour chaque cas : c'est-à-dire de déterminer le profil clinique du cas par l'extraction des caractéristiques pathologiques, décrites par tête de chapitre du MeSH, section [C]-Maladies [1].

La liste des intitulés descriptifs est la suivante : Infections bactériennes et mycoses, Maladies virales, Maladies parasitaires, Tumeurs, Maladies ostéomusculaires, Maladies de l'appareil digestif, Maladies du système stomatognathique, Maladies de l'appareil respiratoire, Maladies oto-rhino-laryngologiques, Maladies du système nerveux, Maladies de l'œil, Maladies urogénitales de l'homme, Maladies de l'appareil urogénital féminin et complications de la grossesse, Maladies cardiovasculaires, Hémopathies et maladies lymphatiques, Malformations et maladies congénitales, héréditaires et néonatales, Maladies de la peau et du tissu conjonctif, Maladies métaboliques et nutritionnelles, Maladies endocriniennes, Maladies du système immunitaire, États, signes et symptômes pathologiques, Troubles dus à des produits chimiques, Plaies et blessures. À ces descripteurs complets correspondent respectivement les labels suivants : *infections, virales, parasitaires, tumeur, osteomusculaires, digestif, stomatognathique, respiratoire, ORL, nerveux, oeil, homme, femme, cardiovasculaires, hemopathies, genetique, peau, nutritionnelles, endocriniennes, immunitaire, etatsosy, chimiques, blessures*.

La Figure 1 présente la répartition des labels dans le jeu de données d'entraînement, à titre indicatif. Le label *etatsosy* apparaît dans 141 textes tandis que *stomatognathique* n'est présent que dans 3 textes. La Figure 2 présente le nombre de labels par document, avec une médiane à 3.

Du fait de cette répartition hétérogène et du faible volume du jeu d'entraînement, nous avons ajouté au jeu d'apprentissage l'ensemble des termes du MeSH français, section [C]-Maladies.

# 2 Description du système

La Figure 3 décrit l'architecture générale du système que nous proposons.

Une analyse du jeu de données d'entraînement nous montre que ce sont les entités de *pathologie* et *signe ou symptôme* (*sosy*) qui donnent lieu à des classifications MeSH-maladies. Les autres entités peuvent donc être ignorées, et parmi les entités de ces deux types, deux cas doivent être ignorés également :

— les entités étiquetées dans le jeu de données comme niées, hypothétiques ou associées à une autre personne que le ou la patiente ;
— les entités n'ayant pas d'attribut particulier dans le jeu de données mais correspondant à des résultats négatifs (examen normal, analyses négatives ...).

1. http://mesh.inserm.fr/FrenchMesh/index.html

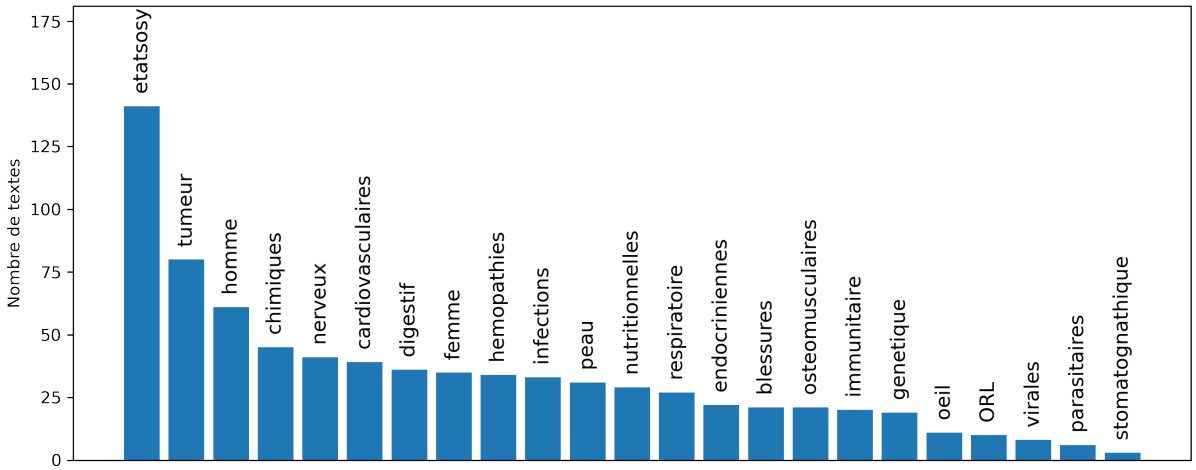


FIGURE 1 – Nombre de textes étiquetés par chaque label, dans le jeu d'entraînement

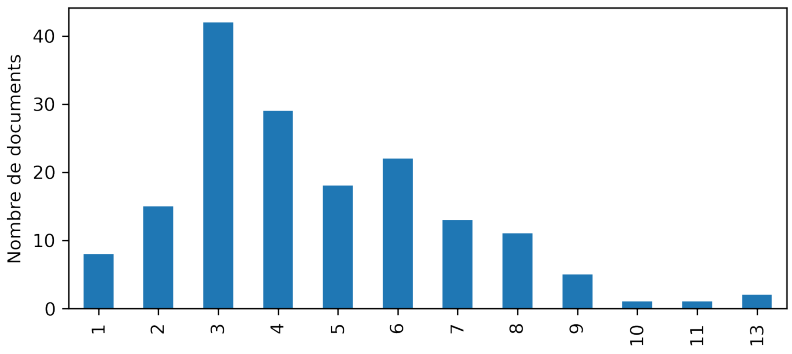


FIGURE 2 – Nombre de labels par document dans le jeu d'entraînement

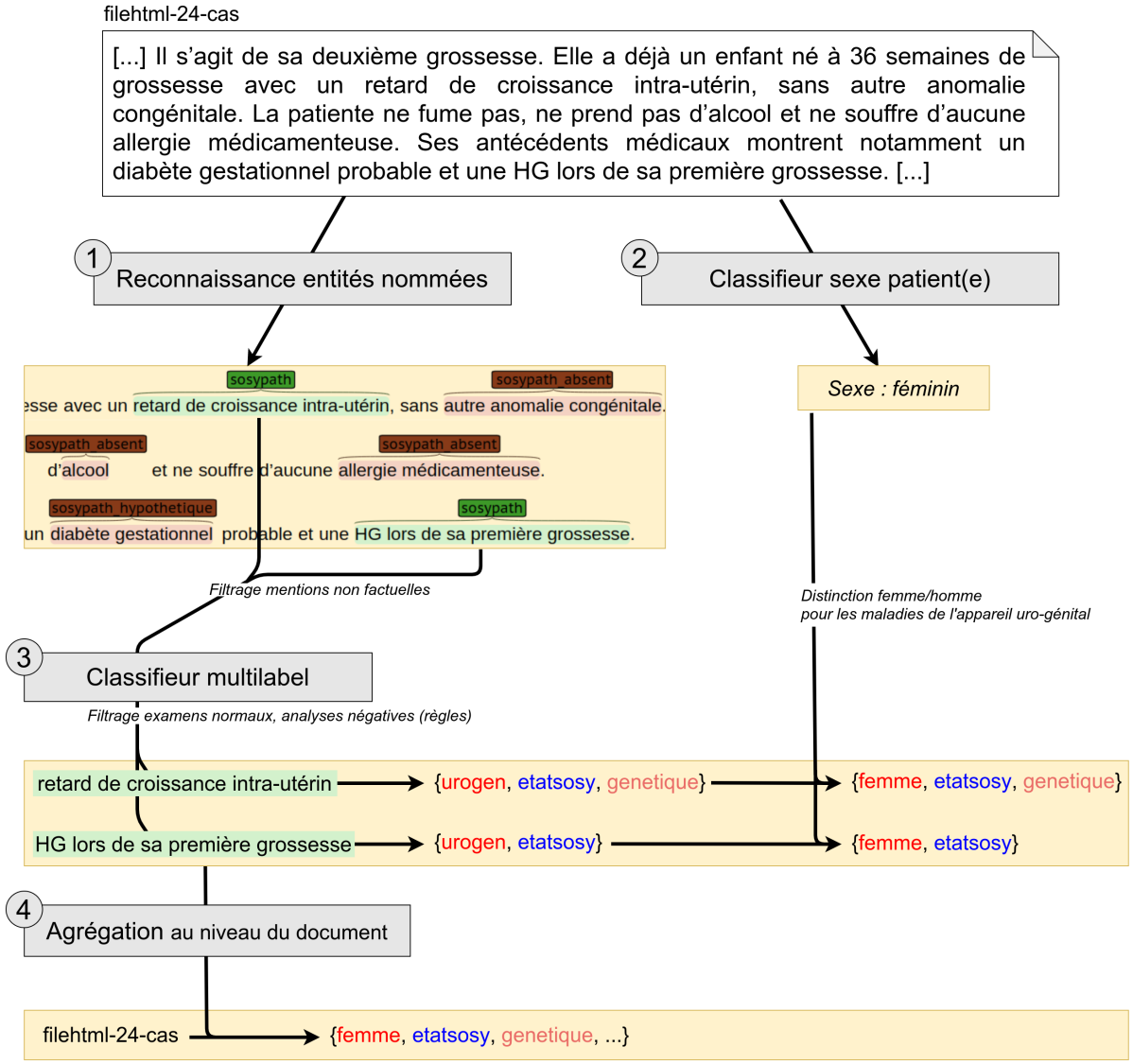


FIGURE 3 – Architecture générale du système

Les entités de type *pathologie* et *sosy*, que nous utilisons pour notre chaîne de traitement, sont fournies par les organisateurs dans le jeu d'entraînement comme dans le jeu de test. Néanmoins, dans l'optique d'évaluer un système « bout-en-bout », nous testons également l'utilisation d'un outil de reconnaissance d'entités nommées (REN) pour réaliser l'extraction de ces entités (run 3). Il s'agit de l'étape 1 présentée à la Figure 3, qui est donc optionnelle. Pour ce système, nous fusionnons les entités *pathologie* et *sosy* en une seule, pour conduire aux entités à extraire : *sosy_path*, *sosy_path_absent* (i.e. nié), *sosy_path_hypothetique*, *sosy_path_non_associe* (i.e. relatif à une autre personne). La fusion se justifie, selon nous, d'une part par la proximité sémantique des deux concepts, d'autre part par l'avantage de regrouper les contextes syntaxiques liés à la la négation, l'hypothèse, la parenté, pour favoriser l'apprentissage de ces notions non triviales.

Par ailleurs, les chapitres MeSH *femme* et *homme* ne se distinguent parfois que par le sexe du ou de la patiente (par exemple, *anurie*). Il est donc nécessaire de construire un classifieur prédisant le sexe à partir du contenu du compte-rendu (étape 2 de la figure).

Une fois les termes d'intérêt extraits par le système, un classifieur a pour tâche de prédire le ou les chapitres MeSH concernés par chaque terme (étape 3). Il s'agit donc d'un classifieur multilabel et multiclasse (les 22 classes représentées dans le jeu de données, en agrégeant *femme* et *homme* en *urogen*). Nous entraînons ce classifieur sur les termes du jeu d'entraînement, mais également sur l'ensemble des termes français de la classification MeSH-maladie contenue dans chaque branche de l'arborescence, classés par tête de chapitre [2].

Enfin, nous agrégeons les informations extraites aux niveaux des termes, pour conclure sur la classification de chaque document.

Les sections qui suivent détaillent chacune de ces étapes.

## 2.1 Reconnaissance d'entités nommées

Nous présentons ici notre modèle de reconnaissance d'entités nommées, ainsi que le *fine tuning* que nous avons effectué sur le modèle camemBERT (Martin *et al.*, 2020).

### 2.1.1 Architecture du modèle

Le modèle de reconnaissance d'entités nommées utilisé pour notre système « bout-en-bout » est un évaluateur exhaustif composé d'un Transformer BERT (Devlin *et al.*, 2019) et d'un LSTM bidirectionnel (Hochreiter & Schmidhuber, 1997). La méthode employée est similaire à celle de Yu *et al.* (2020). Les extractions sont des triplets (début, fin, classe). Chaque mot du texte est d'abord divisé en *word pieces* et contextualisé par le Transformer. Les représentations des 4 dernières couches de BERT sont moyennées avec des poids appris, et les *word pieces* d'un mot sont agrégés par *max-pooling* pour construire sa représentation. Un encodage de type char-CNN (Lample *et al.*, 2016) du mot est également concaténé à la représentation précédente. Ces représentations de mots passent par un *Highway LSTM* à 3 couches (Kim *et al.*, 2017) pour obtenir la représentation finale de chaque mot $E_i$. Enfin, chaque entité possible est évaluée par un produit scalaire entre les représentations de ses bornes de début et de fin :

$$P(span(i,j,k)) = \sigma((W_k^{\text{begin}} \cdot E_i + b_k^{\text{begin}}) \cdot (W_k^{\text{begin}} \cdot E_j + b_k^{\text{end}}) + \text{bias})$$

2. http://mesh.inserm.fr/FrenchMesh/index.html

Les paramètres sont optimisés *via* un objectif de *cross entropy* avec **Adam** (Kingma & Ba, 2015). Nous utilisons un pas d'apprentissage à décroissance linéaire avec un *warmup* de 10 % et deux valeurs initiales : $4 \cdot 10^{-5}$ pour le Transformer, et $6 \cdot 10^{-4}$ pour les autres paramètres.

### 2.1.2 *Fine-tuning* du camemBERT

Le modèle Transformer utilisé pour la reconnaissance d'entités nommées (ainsi que pour le classifieur) a été au préalable *fine-tuné* sur un jeu de données en accès libres sur Europe-PMC [3] : une première extraction a été réalisée sur l'ensemble des articles. Une fois cette base de d'articles extraite, un traitement de détection de la langue a ensuite été réalisé sur les corps de texte, avec la librairie *langdetect* de Python. Une restriction a ensuite été opérée sur le type d'articles pour ne retenir que les suivants : *case-report* pour environ 1 900 textes, *research-article* (1 060 textes), *brief-report*, *review-article*, *abstract*, *letter*, *chapitre-article*, *discussion*. Le *fine-tuning* a été ensuite réalisé à partir du modèle camemBERT-large (Martin *et al.*, 2020) pour 30 époques. La perplexité calculée sur un jeu de validation de 20 % était de 2,32.

## 2.2 Classification du sexe du patient ou de la patiente

Le corpus est assez homogène ; il est constitué de documents qui sont des descriptions de cas cliniques. Dans la majorité des cas, ces documents parlent d'une unique personne. L'information sur le sexe de cette personne est dans ce type de textes une donnée déterminante, en particulier pour discriminer entre les deux étiquettes *femme* et *homme*, comme indiqué ci-dessus.

Afin d'entraîner un classifieur à déterminer le sexe, nous avons recueilli un grand nombre de variables candidates afin d'évaluer leur pertinence. Une observation des documents a tout d'abord permis de déterminer que dans un très grand nombre de cas, dans ce type de documents, l'information décrivant le patient se trouvait dans la première phrase. Nous avons donc pondéré les variables par leur distance au début du texte (suivant une pondération fonction du numéro d'ordre de la phrase dans le document, commençant à 1 pour la première phrase et décroissant linéairement jusqu'à 0,5 pour la dernière).

| **Variable observée** | **Information Mutuelle avec la classe « sexe »** |
|---|---|
| 1. genre du mot « patient(e) » | 0,5362816 |
| 2. genre des adjectifs appliqués à des humains | 0,4401735 |
| 3. sexe associé aux préfixes caractérisant des morphèmes biomédicaux (organes, maladies, procédures chirurgicales) | 0,2874693 |
| 4. genre des mots de civilité | 0,2178381 |
| 5. genre des mots désignant l'individu | 0,1874953 |
| 6. genre des pronoms personnels 3ps | 0,0771561 |
| 7. indication explicite du sexe | 0,0200703 |
| 8. genre des prénoms | 0,0185550 |
| *autres variables explorées* | $< 10^{-2}$ |

TABLE 1 – Information mutuelle de variables extraites des documents avec la classe *sexe*.

3. http://europepmc.org/

Nous avons ensuite relevé les variables qui nous paraissaient significatives lors d'un premier parcours qualitatif du corpus (voir Table 1). La variable la plus significative est (1) le **genre du mot *patient(e)***. Cette variable, lorsqu'elle est présente (moitié des documents environ), indique sans ambiguïté le sexe. Les autres variables qui contribuent significativement à déterminer le sexe sont, par ordre d'importance : (2) le **genre des adjectifs appliqués à des humains**. Il s'agit des adjectifs qualificatifs qui sont sans ambiguïté utilisés pour des personnes (*« âgé(e) »*, *« né(e) »*, *« enceinte »* ...) ; certains décrivent le contexte de consultation, et nécessitent une petite vérification du contexte (*« adressé(e) (à notre service) »*, *« présenté(e) (aux urgences) »*, *« revu(e) (en consultation) »* ...). (3) Le nombre d'occurrences de **morphèmes faisant référence à des concepts biologiques ou médicaux** spécifiques à un sexe (par exemple *péni-*, *utér-*, *testi-*, *vagin-*. La liste utilisée pour ces morphèmes a été construite en partant de la liste des termes du MeSH français, maintenue par l'INSERM, limitée aux sous-arbres situés dans la hiérarchie sous les nœuds A05 (*appareil urogénital*) pour les termes d'anatomie, C12 (*maladies urogénitales de l'homme*) et C13 (*maladies de l'appareil urogénital féminin et complications de la grossesse*) pour les termes désignant des maladies, enfin E04.950 (*procédures de chirurgie urogénitale*) pour les termes de chirurgie. La liste des termes complets a ensuite été « rabotée » pour factoriser les termes ayant des préfixes communs suffisamment caractéristiques (*stems*). (4) Le **genre des appellatifs de civilité**, sous leurs différentes formes de surface (*« M. »*, *« Mr »*, *« Monsieur »*, *« Mme »*, *« Madame »* ...). (5) Le **genre des noms communs fréquemment utilisés pour désigner un individu humain** (*femme*, *homme*, *enfant* ...).

Plus marginalement, mais encore utilement : (6) Le **genre des pronoms personnels de troisième personne du singulier** utilisés dans le texte (méthode rapide, sans détection de coréférences). (7) L'**indication explicite du sexe** (*« masculin »* ou *« féminin »*). (8) le **genre des prénoms**, déterminé d'après une liste de référence (INSEE) des prénoms les plus fréquemment donnés en France et du genre associé (en effet, dans les études de cas, le patient est souvent désigné par « *M. A.* » ou « *Mme B.* », mais aussi parfois par un prénom).

Pour la collecte de certaines variables, nous avons extrait les catégories morphosyntaxiques (POS, genre, nombre) et les dépendances syntaxiques à l'aide de la librairie *stanza* (Qi *et al.*, 2020).

Nous avons annoté manuellement la classe **sexe du patient** sur le corpus d'entraînement et nous avons entraîné un classifieur supervisé **AdaBoost** (Freund & Schapire, 1997), à partir de ces données, pour déterminer la fonction de prédiction du sexe à partir d'un document texte. Afin de valider cette approche, nous avons entraîné le classifieur sur 80 % des données d'entraînement fournies et l'avons validé sur un jeu de 20 %. Le rappel et la précision étaient tous les deux de 1, y compris sur le jeu de test final. Les cas qui échappent à cette approche de prédiction du sexe du patient sont les rares descriptions de cas cliniques qui concernent plusieurs patients à la fois.

## 2.3 Classifieur de mentions en chapitres MeSH et agrégation des résultats

Nous effectuons un filtre préalable sur les sorties de la REN ou sur les entités fournies par les organisateurs, de façon à retirer les résultats négatifs (examen normal, analyses négatives). En effet, ces éléments sont souvent annotés comme des *sosy* dans le jeu de données, mais ne doivent pas donner lieu à une annotation MeSH. Ce filtrage est assuré par des expressions régulières simples.

Le classifieur utilisé est un classifieur également composé d'un Transformer BERT (Devlin *et al.*, 2019), plus spécifiquement, un Transformer de type CamembertForSequenceClassification de la librairie Huggingface (Martin *et al.*, 2020; Wolf *et al.*, 2020) comprenant une dernière couche

linéaire en sortie. Pour permettre d'emblée une classification multi-classes, la fonction de perte est la cross-entropie binaire, sommée pour toutes les classes. Deux modèles ont été entrainés : un modèle camembert-large, avec un optimiseur Adam (Kingma & Ba, 2015), pour 50 époques avec un pas d'apprentissage à décroissance linéaire, débutant à $1 \cdot 10^{-5}$. Un jeu de validation de 20 % pour le camembert-large et de 15 % pour le modèle *fine-tuné*.

Pour la prédiction, les scores sont calculés par la fonction sigmoïde en sortie. Le seuil retenu pour la prédiction définitive est celui qui maximise la F1-mesure sur le jeu de validation.

Le deuxième modèle utilisé correspond au CamembertForSequenceClassification *fine-tuné* à partir d'un jeu de données biomédicales françaises en libre-accès sur Europe-PMC, correspondant au même modèle décrit à la section 2.1.2.

Pour finir, la classe *urogen* est séparée en *femme* ou *homme* selon le sexe du ou de la patiente, tel que prédit par l'étape 2.

Enfin, les résultats issus des étapes précédentes sont agrégés de façon triviale pour composer la sortie finale, c'est-à-dire une liste sans doublon des chapitres MeSH-maladie concernés par le compte-rendu (étape 4).

### 2.4 Configurations d'entraînement et runs soumis

Concernant les configurations, deux runs ont été soumis à partir des annotations des années précédentes et le troisième run correspond à notre système « bout-en-bout ». Pour les deux runs réalisés à partir des annotations des années précédentes, le premier est réalisé avec le CamembertForSequenceClassification entraîné à partir du camembert-large classique et le second à partir du camembert-large *fine-tuné* sur les données Europe-PMC. Le run réalisé avec notre système bout-en-bout est basé sur le camembert-large *fine-tuné* pour l'extraction et pour le classifieur. Ces configurations sont synthétisées à la Table 2.

| | Modèle REN | Extraction du genre | Modèle classifieur |
|---|---|---|---|
| Run 1 | Aucun | AdaBoost | camembert-large |
| Run 2 | Aucun | AdaBoost | camembert-large FT |
| Run 3 | REN (camembert-large FT) | AdaBoost | camembert-large FT |

TABLE 2 – Configurations des différents runs

## 3 Résultats et discussion

Les résultats des trois runs proposés sont présentés à la Table 3, respectivement aux méthodes proposées table 2. La méthode ayant conduit au meilleur score est celle du camembert-large avec annotations *sosy* et *pathologie* des années précédentes (avec suppression des termes niés, de parenté ou d'hypothèse). De manière particulièrement intéressante, notre modèle « bout-en-bout » *run* 3 parvient à des résultats très proches de ce dernier modèle, sans s'appuyer sur aucune annotation pour la phase de test, et en ayant été entraîné sur la phase REN uniquement sur les 167 compte-rendus d'entraînement. L'étape du *fine-tuning* du camembert-large, *run* 2, n'a pas permis d'améliorer le

| | Précision | Rappel | F1-Mesure |
|---|---|---|---|
| Run 1 | 0,696 | 0,874 | 0,775 |
| Run 2 | 0,677 | 0,875 | 0,771 |
| Run 3 | 0,689 | 0,872 | 0,770 |
| Médiane DEFT 2021 | | | 0,700 |
| Meilleur DEFT 2021 | 0,885 | 0,750 | 0,812 |

TABLE 3 – Résultats officiels

| | Seuils | 0,2 | 0,3 | 0,4 | 0,5 | 0,6 | 0,7 | 0,8 | 0,9 |
|---|---|---|---|---|---|---|---|---|---|
| Run 1 | Rappel | 0,921 | 0,908 | 0,906 | 0,900 | 0,893 | 0,882 | 0,874 | 0,860 |
| | Précision | 0,619 | 0,644 | 0,655 | 0,659 | 0,666 | 0,685 | 0,696 | 0,711 |
| | F1-Mesure | 0,741 | 0,753 | 0,760 | 0,761 | 0,763 | 0,771 | **0,775** | **0,778** |
| Run 2 | Rappel | 0,922 | 0,919 | 0,915 | 0,904 | 0,904 | 0,895 | 0,889 | 0,872 |
| | Précision | 0,628 | 0,640 | 0,657 | 0,664 | 0,672 | 0,677 | 0,685 | 0,716 |
| | F1-Mesure | 0,747 | 0,754 | 0,765 | 0,766 | 0,771 | **0,771** | 0,774 | **0,787** |
| Run 3 | Rappel | 0,908 | 0,904 | 0,896 | 0,887 | 0,885 | 0,872 | 0,865 | 0,858 |
| | Précision | 0,638 | 0,650 | 0,672 | 0,676 | 0,686 | 0,689 | 0,700 | 0,728 |
| | F1-Mesure | 0,750 | 0,756 | 0,768 | 0,767 | 0,773 | **0,770** | 0,774 | **0,788** |

TABLE 4 – Variation des résultats sur le jeu de test, en fonction du seuil de classification en sortie des 3 modèles. Les valeurs en gras correspondent aux résultats officiels (pour les seuils 0.8 et 0.7, respectivement pour les run 1 et 2-3), et aux meilleurs résultats.

score, ce qui s'explique probablement par le fait que le camembert-large est déjà entraîné sur un très gros volume de données, y compris très hétérogènes, de même que les cas cliniques proposés pour la phase de test. Par ailleurs, le volume de 4 000 articles était probablement insuffisant pour permettre un réel apport au modèle.

A titre d'expérimentation supplémentaire (non soumise aux organisateurs), le modèle bout-en-bout basé sur le camembert-large (Martin *et al.*, 2020), non *fine-tuné*, pour la REN et le classifieur, fourni les résultats suivants avec le script d'évaluation des organisateurs, sur les données de tests : R=0, 871 P=0, 701 F=0, 777. Tous les autres paramètres étant égaux par ailleurs au *run* 1.

Par ailleurs, le seuil en sortie de classifieur nous a paru être un paramètre-clé dans la tache du DEFT 2021. En effet, celui-ci présentait des variations significatives en fonction du volume jeu de données de validation : 20 % pour le classifieur-camemembert large et 15 % pour le modèle *fine-tuné*. Il a été calculé à 0,7 pour le modèle *fine-tuné* et à 0,8 pour le modèle large classique sur ces jeux de validation. Pour illustrer cette importance, nous avons fait varier uniquement ce seuil pour les 3 modèles décrits et calculé les résultats sur le jeu de test avec le script d'évaluation. Les résultats correspondants sont présentés table 4. On observe que les 3 modèles sont très équivalents et que le meilleur score est obtenu pour le modèle « bout-en-bout ».

Enfin, notons que la balance rappel/précision est largement en faveur du rappel dans toutes nos expériences, à l'opposé du système qui se classe premier à la campagne, ce qui pourrait rendre des expériences d'hybridation intéressantes.

# Références


DEVLIN J., CHANG M. W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies - Proceedings of the Conference*, **1**, 4171–4186.

FREUND Y. & SCHAPIRE R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, **55**(1), 119–139. DOI : https ://doi.org/10.1006/jcss.1997.1504.

GROUIN C., GRABAR N. & ILLOUZ G. (2021). Classification de cas cliniques et évaluation automatique de réponses d'étudiants : présentation de la campagne deft 2021. In *Actes de DEFT, Lille*.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long Short-Term Memory. *Neural Computation*, **9**(8), 1735–1780. DOI : 10.1162/neco.1997.9.8.1735.

KIM J., EL-KHAMY M. & LEE J. (2017). Residual LSTM : Design of a Deep Recurrent Architecture for Distant Speech Recognition. In *Interspeech 2017*, volume 2017-Augus, p. 1591–1595, ISCA : ISCA. DOI : 10.21437/Interspeech.2017-477.

KINGMA D. P. & BA J. L. (2015). Adam : A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 260–270, Stroudsburg, PA, USA : Association for Computational Linguistics. DOI : 10.18653/v1/N16-1030.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics.

QI P., ZHANG Y., ZHANG Y., BOLTON J. & MANNING C. D. (2020). Stanza : A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., SCAO T. L., GUGGER S., DRAME M., LHOEST Q. & RUSH A. M. (2020). Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics.

YU J., BOHNET B. & POESIO M. (2020). Named Entity Recognition as Dependency Parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 6470–6476, Stroudsburg, PA, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.577.