Performance en classification de données textuelles des passages aux urgences des modèles BERT pour le français

Gabrielle Chenais, Hélène Touchais, Marta Avalos, Loïck Bourdois, Philippe Revel, Cédric Gil-Jardiné, Emmanuel Lagarde

Performance en classification de données textuelles des passages aux urgences des modèles BERT pour le français

Performance of BERT models for French in the classification of textual data from emergency room visits

Gabrielle Chenais[1], Hélène Touchais[1], Marta Avalos-Fernandez[1,2], Loïck Bourdois[1], Philippe Revel[1,3], Cédric Gil-Jardiné[1,3], and Emmanuel Lagarde[1]

[1] Univ. Bordeaux, INSERM, BPH U1219, F-33000 Bordeaux, France

[2] SISTM team, INRIA BSO, F-33405, Talence, France

[3] CHU de Bordeaux, Service des urgences, F-33000, Bordeaux, France
∗ Auteur correspondant emmanuel.lagarde@u-bordeaux.fr

Résumé : Les modèles de langue contextualisés basés sur l'architecture Transformer tels que BERT (Bidirectional Encoder Representations from Transformers) ont atteint des performances remarquables dans des diverses tâches de traitement de la langue. CamemBERT et FlauBERT en sont des versions pré-entraînées pour le français. Nous avons utilisé ces deux modèles afin de classer automatiquement des notes cliniques libres issues de visites aux urgences à la suite d'un traumatisme. Leurs performances ont été comparées à la méthode TF-IDF (Term-Frequency – Inverse Document Frequency) associé au classifieur SVM (Support Vector Machine) sur 22481 notes cliniques provenant du service des urgences du CHU de Bordeaux. CamemBERT et FlauBERT ont obtenu des résultats légèrement supérieurs à ceux du couple TF-IDF/SVM pour le micro F1-score. Ces résultats encourageants permettent d'envisager l'utilisation des transformers pour automatiser le traitement des données des urgences dans le cadre de la mise en place d'un observatoire national du traumatisme en France.
Mots-clés : Traitement automatique du langage, CamemBERT, FlauBERT, TF-IDF, SVM classification supervisée multi-classe, Urgences.

Abstract : Contextualized language models based on the Transformer architecture such as BERT (Bidirectional Encoder Representations from Transformers) have achieved remarkable performances in various language processing tasks. CamemBERT and FlauBERT are pre-trained versions for French. We used these two models to automatically classify free clinical notes from emergency department visits following a trauma. Their performances were compared to the TF-IDF (Term-Frequency - Inverse Document Frequency) method associated with the SVM (Support Vector Machine) classifier on 22481 clinical notes from the emergency department of the Bordeaux University Hospital. CamemBERT and FlauBERT obtained slightly better results than the TF-IDF/SVM couple for the micro F1-score. These encouraging results allow us to consider further developments in the use of transformers in the automation of emergency department data processing in order to consider the implementation of a national observatory of trauma in France.
Keywords : Natural Langage Processing, Artificial Intelligence, CamemBERT, FlauBERT, TF-IDF, SVM, multi-class classification, Emergency.

1 Introduction

Environ un tiers des visites aux urgences sont la conséquence de traumatismes [1]. En 2017 en France, les traumatismes et blessures ont représenté 7,0(6,8- 7,3) % des décès [2]. L'informatisation des services hospitaliers, et particulièrement des urgences, offre la possibilité d'enregistrer en continu les informations relatives aux visites et de les rendre rapidement

1. Cour des comptes, accès le 3/3/2020
2. The Institute for Health Metrics and Evaluation, accès le 3/3/2020

disponibles pour la veille sanitaire. Pour autant, peu d'informations sont disponibles lorsqu'il s'agit de traumatisme : si l'on peut connaître la nature de la blessure principale, les circonstances (accident de la route, agression, suicide, etc) ne sont pas renseignées de façon standardisée. Ces informations sont disponibles dans le dossier patient informatisé, mais sous forme de texte libre. Le traitement automatique du langage semble indiqué pour extraire les informations souhaitées sur les mécanismes de traumatisme à partir des anamnèses, une tâche qui n'est pas envisageable manuellement compte-tenu du coût-horaire du personnel de santé ainsi qu'à la surcharge administrative pré-existante.

L'analyse du langage naturel a connu un tournant récent avec l'essor de l'apprentissage profond et, en particulier, de l'architecture de type Transformer. Introduite en 2017 par Google et basée sur la notion "d'attention" (Vaswani *et al.*, 2017), les transformers peuvent être pré-entraînés à partir d'un corpus de texte qui peut être très grand puisqu'il ne nécessite pas d'annotation. Cette phase non supervisée de l'apprentissage consiste alors à prédire un token du texte artificiellement masqué à partir des mots qui l'entourent (Rothe *et al.*, 2019). BERT (Bidirectional Encoder Representations from Transformers) (Devlin *et al.*, 2018) est un exemple de Transformer bidirectionnel, composé uniquement de couches d'encodeur. Pour de nombreuses tâches, dont la classification de textes, ses performances sont systématiquement supérieures aux modèles convolutionnels et auto régressifs utilisés jusqu'alors (Devlin *et al.*, 2018). Des dérivés français tels que FlauBERT (Le *et al.*, 2019) et CamemBERT (Martin *et al.*, 2019) ont été proposés en 2019. FlauBERT est un BERT français entraîné sur un corpus français très large et hétérogène. CamemBERT, quant à lui, est basé sur l'architecture RoBERTa (Robustly Optimized BERT Pretraining).

RoBERTa propose une approche avec une méthodologie d'entraînement améliorée par rapport à celle de BERT et bénéficie de puissance de calcul plus importante aboutissant à des capacités supérieures à BERT (2-20%) sur plusieurs tâches (Liu *et al.*, 2019).

Extraire des mécanismes et/ou types de traumatismes relève de la classification multi-classe. Une grande variété de techniques ont été utilisées à cette fin pour des données médicales françaises : des approches basées sur des ontologies (Cossin *et al.*, 2018), des réseaux neuronaux (Flicoteaux, 2018; Ive *et al.*, 2018), ou encore des forêts aléaloires (Metzger *et al.*, 2017). La technique TF-IDF (Terms Frequency-Inverse Document Frequency) (Spärck Jones, 2004) associée au classifieur SVM (Support Vecteur Machine) représentait jusqu'à récemment l'état de l'art dans la classification de textes biomédicaux (Zhong *et al.*, 2018; Wang *et al.*, 2017).

Dans le cadre du projet TARPON (Traitement Automatique des Résumés de Passage aux urgences dans le but de créer un Observatoire National) destiné à démontrer la faisabilité de la mise en place d'un observatoire national du traumatisme à partir des urgences, nous proposons ici de comparer les performances de plusieurs transformers pour la classification des visites aux urgences pour traumatismes à partir des anamnèses émanant du services des urgences adulte du CHU de Bordeaux. Nous avons comparé les FlauBERT et CamemBERT à la combinaison TF-IDF/SVM. La disponibilité de ces modèles étant très récente, il s'agit d'une des premières études portant sur l'utilisation des transformers sur des données cliniques françaises pour de la classification multi-classes.

2 Méthodes

Base de données. Les notes cliniques ont été extraites des dossiers médicaux électroniques des urgences adultes stockés au sein du système d'information de l'hôpital universitaire de Bordeaux, en France. Elles correspondent aux 374 900 dossiers médicaux des visites aux sein du service des urgences adultes de l'hôpital Pellegrin de 2012 à 2020. Les variables disponibles étaient : l'âge, le sexe, la date et l'heure de la venue, l'anamnèse générée par les médecins/internes ainsi que l'anamnèse écrite par les infirmiers et infirmières d'accueil et d'orientation.

Stratégie d'échantillonnage et d'annotation manuelle. 69 110 anamnèses ont été extraites au hasard afin d'être codées manuellement. Notre équipe de codeurs était composée

d'épidémiologistes spécialisés dans le traumatisme, de médecins urgentistes, d'infirmières des urgences, d'assistants de recherche et de biostatisticiens, pour un total de 16 codeurs. La phase d'annotation a duré 5 mois. Pour chaque anamnèse, un code décrivant le contenu du texte a été attribué. La grille d'annotation utilisée pour le codage a été développée pour les besoins du projet. Le code associé à chaque anamnèse comprenait 8 champs. L'objectif étant de classer les types de traumatisme, nous avons utilisé majoritairement les données du champ "Type de traumatisme ou Mode de déplacement pour l'AVP". La distribution de celui-ci étant déséquilibrée, nous avons créé une variable composite contenant 8 classes mutuellement exclusives afin d'avoir un nombre plus important d'anamnèses par classe. Par conséquent, nous avons procédé au regroupement de certains types de traumatisme (par exemple, "Chute" qui comprenait "Chute de sa propre hauteur", "Chute d'une hauteur donnée" et "Chute dans les escaliers"). Afin d'obtenir la classe "Sport", nous avons utilisé le champ "Activité". La variable composite comprenait les classes/étiquettes suivantes : "Accident d'exposition aux fluides corporels (accident d'exposition au sang, rapports sexuels non protégés à risque)" (AEF), "Accident de sport" (Acc. sport), "Agression", "Accident de la Voir Publique (AVP)", "Auto-agression", "Chute (sauf sport)", "Corps étranger dans les yeux" (CE), "Autre traumatisme".

Une analyse de sensibilité a été réalisée afin d'étudier l'impact de contenus potentiellement ambigus dans les anamnèses. L'échantillon test a donc été relu par un expert. Un contenu potentiellement ambigu quant à sa classification est défini ici comme l'accumulation de plusieurs mécanismes ou types de traumatismes et/ou une difficulté majeure à attribuer une étiquette à une note clinique donnée compte-tenu de son texte.

Modèles. Les modèles sélectionnés pour la comparaison, et disponibles gratuitement sous forme de contenu open-source, étaient le couple TF-IDF et SVM ainsi que 2 modèles de type transformer pré-entraînés sur des corpus français : CamemBERT (Martin *et al.*, 2019), FlauBERT (Le *et al.*, 2019). Le couple TF-IDF/SVM a été utilisé via le package scikit-learn (Pedregosa *et al.*, 2011) [3], la tokenisation a été effectuée à l'aide du package nltk (Bird & Klein, 2009) [4], de même les mots les plus fréquents (par exemple : "le", "lui", "ce") ont été supprimés. CamemBERT et FlauBERT ont été utilisés grâce au package Transformers de la plate-forme Hugging Face (Wolf *et al.*, 2020) [5]. Le modèle CamemBERT pré-entrainé sur le jeu de données CCNet (Wenzek *et al.*, 2019) a été choisi car CCNNet se positionne entre les deux autres jeux de données sur lesquels CamemBERT a été testé pour le pré-entraînement soit OSCAR (Suarez *et al.*, 2019), peu filtré voire bruité, et Wikipedia, totalement édité. La tokenisation a été effectuée avec SentencePiece (Wu *et al.*, 2016) pour CamemBERT et avec le Byte-Pair Encoding (BPE) pour FlauBERT (Shibata *et al.*, 1999). Les données ont été nettoyées à l'aide d'expressions régulières dans Python 3.7. La normalisation Unicode a été effectuée en UTF-8 (Universal Character Set Transformation Format - 8). Les modèles ont été entraînés sur notre station de travail avec soit une carte graphique GeForce GTX (NvidiaR©) 1080 Ti avec 11GB de VRAM soit une Titan RTX (NvidiaR©) avec 24Go de VRAM.

Phase d'apprentissage Supervisé. 80 % du jeu de données annotées a été dédié à l'apprentissage supervisé des modèles. Ces données d'apprentissage ont été divisées en un échantillon d'entraînement (n=14532) et un échantillon de validation (n=3634). La mesure utilisée pour l'analyse de la validation était l'exactitude.

CamemBERT et FlauBERT ont été entraînés à partir de 9 initialisations différentes, sur 7 epochs grâce à Pytorch [6]. Afin de calculer le micro F1-score sur l'échantillon de validation, un vote a été effectué entre les 9 prédictions pour chaque anamnèses. Ceci nous a permis de sélectionner, pour chaque transformer, une epoch donnée (celle où le micro F1-score issu du vote était le plus élevé) afin d'obtenir les prédictions sur les échantillons de données test. Les modèles ont été entraînés avec la variable composite comme source d'étiquettes.

Phase de Test. L'échantillon initial de test contenait 20 % du jeu de données annotées soit

3. https://scikit-learn.org/stable
4. https://www.nltk.org
5. https://github.com/huggingface/transformers
6. https://github.com/pytorch/pytorch

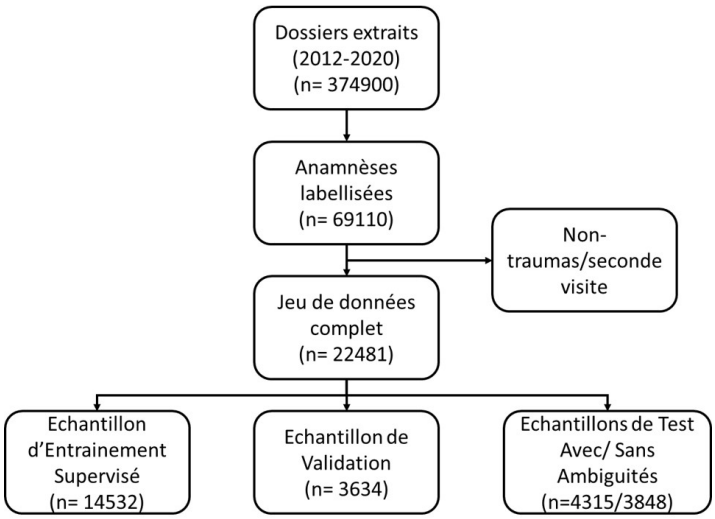

FIGURE *1 – Diagramme de flux.*

4315 anamnèses. La seconde lecture de ces anamnèses a permis d'en identifier 467 comme étant ambiguës. L'analyse comprend donc d'une part le jeu de données test avec ambiguïtés (n=4315) et d'autre part sans ambiguïtés (n=3848). La répartition des différents échantillons est donnée dans la Figure 1. Afin d'obtenir les probabilités pour chaque prédiction, une couche d'activation de softmax a été appliquée. Un modèle ayant été sélectionné avec une epoch donnée après la phase d'entraînement, un vote a été appliqué entre les 9 exécutions d'entraînement en vue d'obtenir une seule prédiction pour une anamnèse donnée.

Métriques. Les métriques permettant de mesurer la performance de chaque méthode étaient : la macro précision, l'exactitude qui est égale à la micro-précision ainsi qu'au micro-rappel et au micro F1-score et le top-2 d'exactitude.

3 Résultats

De manière générale, les micro F1-scores et les macro F1-scores étaient plus élevés avec les transformers qu'avec TF-IDF comme l'indique la Figure 2. Avec l'échantillon de données de test sans ambiguïtés CamemBERT a atteint un micro F1-score de 0,921 et FlauBERT de 0,918 tandis que celui de TF-IDF/SVM était de 0,905. La macro précision de TF-IDF était, quant à elle, plus élevée pour TF-IDF/SVM (0,903) que pour les transformers. Il semblerait donc que TF-IDF/SVM soit plus performant sur les classes les moins communes que CamemBERT et FlauBERT.

La distribution des n anamnèses par classe n'étant pas équilibrée, la micro-F1 est dans tous les cas moins élevée avec les classes peu représentées comme les accidents d'exposition aux fluides corporels ou les auto-agressions. Bien que la classe "corps étrangers dans les yeux" soit plus représentée que celles contenant moins de 100 anamnèses dans l'échantillon de test, les résultats de la micro-f1 sont moindres pour les trois modèles. L'analyse des erreurs des modèles a montré une dispersion dans le codage manuel de cette classe en particulier, certains codeurs attribuant, pour des anamnèses similaires, la mention impact (dans le dernier champ 'Type de Trauma' de la grille d'annotation, devenant 'Autre trauma' avec la variable composite) plutôt que la classe "Corps étranger dans l'oeil". Concernant l'exactitude des différentes classes, CamemBERT a des résultats légèrement plus élevés que TF-IDF et FlauBERT comme indiqué dans le Tableau 1.

Le Tableau 2 montre que la suppression des anamnèses ambiguës est accompagnée d'une augmentation des performances de tous les modèles.

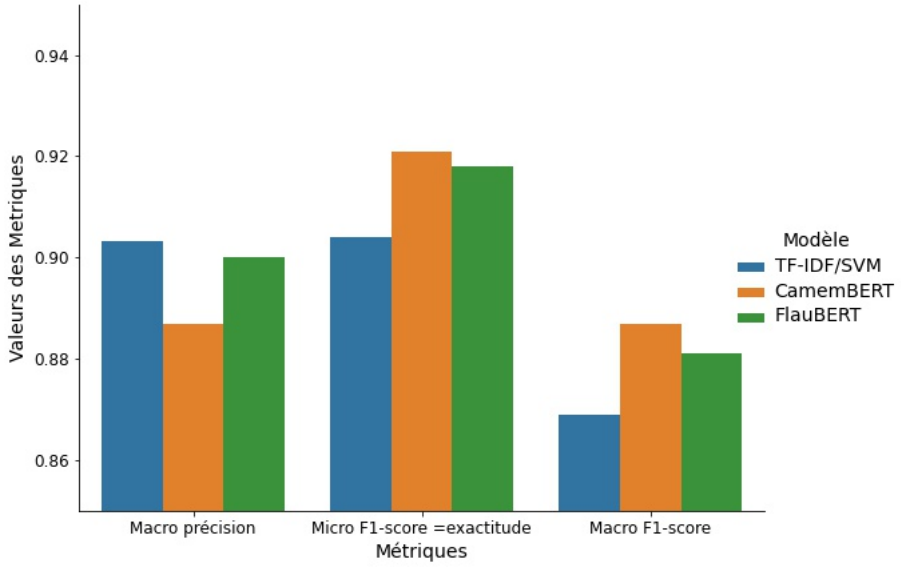

FIGURE *2 – Métriques évaluées sur le jeu de données de test "sans ambiguïtés".*

Type de trauma	*N*	TF-IDF/SVM	CamemBERT	FlauBERT
Accident d'exposition aux fluides corporels	36	0,81	0,82	0,84
Accident de la voie publique	541	0,97	0,97	0,97
Agression	474	0,93	0,93	0,94
Accident de sport	318	0,89	0,91	0,91
Auto-agression	95	0,79	0,80	0,75
Chute	1348	0,95	0,96	0,95
Corps étranger dans les yeux	177	0,80	0,85	0,83
Autre trauma	859	0,84	0,86	0,86
Total	3848			
Macro F1-score		0,869	0,887	0,881
Micro F1-score = accuracy		0,904	0,921	0,918
Top-2 accuracy		0,973	0,971	0,976

TABLE *1 – Micro F1-scores par classe sur le jeu de données de test "sans ambiguïtés", macro, micro F1-scores et exactitude top-2.*

Type de Trauma	*N*	TF-IDF/SVM	CamemBERT	FlauBERT
Accident d'exposition aux fluides corporels	41	0,83	0,84	0,84
Accident de la voie publique	498	0,90	0,91	0,92
Agression	568	0,91	0,90	0,91
Accident de sport	186	0,79	0,84	0,82
Auto-agression	1554	0,90	0,92	0,91
Chute	371	0,82	0,83	0,83
Corps étranger dans les yeux	112	0,75	0,76	0,73
Autre trauma	985	0,80	0,83	0,82
Total	4315			
Macro F1-score		0,838	0,878	0,873
Micro F1-score = exactitude		0,864	0,878	0,873
Exactitude Top-2		0,952	0,956	0,962

TABLE *2 – Micro F1-scores par classe sur le jeu de données de test "avec ambiguïtés", macro, micro F1-scores et exactitude top-2.*

4 Discussion

Transformers : des améliorations à envisager. Les transformers appliqués aux données en texte libre issus des urgences du CHU de Bordeaux ont montré des résultats légèrement supérieurs à TF-IDF sans toutefois se démarquer nettement. La supériorité des transformers sur le couple TF-IDF/SVM, en ce qui concerne les données médicales, n'est pas évident dans la littérature. Les résultats sont, de manière globale, légèrement meilleurs pour différents types de transformers. Lors de la classification de transcriptions de discours de patients en vue de prédire la maladie d'Alzheimer, (Searle *et al.*, 2020) ont montré des performances légèrement meilleures d'un modèle de transformer 'DistilBERT' (Sanh *et al.*, 2019) par rapport à TF-IDF/SVM avec une base de données dont les classes étaient équilibrées. (Hong *et al.*, 2020) ont obtenu des résultats similaires avec le Longformer (Beltagy *et al.*, 2020). L'équipe de Beltagy a appliqué des algorithmes de traitement automatique du langage pour identifier les patients présentant des troubles cognitifs et ont comparé les performances des modèles. Bien que les performances du modèle d'apprentissage profond aient été les meilleures, elles n'étaient que marginalement supérieures à celles des modèles basés sur une régression logistique avec ou sans TF-IDF. L'équipe de H.Goodrum a comparé les performances d'une régression logistique appliquée à une transformation TF-IDF et le modèle Transformer ClinicalBERT (Alsentzer *et al.*, 2019) appliqués à des documents de dossiers médicaux électroniques scannés (Goodrum *et al.*, 2020). Ce dernier modèle a atteint une précision moyenne de 0,882 pour 12 classes, tandis que la précision de TF-IDF/régression logistique était de 0,823. L'équipe allemande dirigée par A. Saadullah (Amin *et al.*, 2019) a obtenu des résultats similaires lors de l'attribution de codes CIM-10 à des résumés non techniques. Les scores f1 de TF-IDF/SVM et BioBERT (Lee *et al.*, 2020) étaient, respectivement, de 0,72 et 0,732.

Choix du pré-training et du tokenizer. Le choix d'utiliser des modèles pré-entrainés sur des corpus en français avec un tokenizer en français a vraisemblablement contribué aux performances de nos modèles. Les auteurs ayant proposé le modèle CamemBERT n'ayant pas comparé les différents modèles issus des jeux de données OSCAR, CCNet et Wikipedia sur une tâche de classification, un futur travail pourrait comparer les différents jeux sur notre base de données. Dans cette logique, il serait opportun, alors que nous n'avons utilisé que les modèles de base de CamemBERT et FlauBERT, de tester les différentes tailles de jeux de données de pré-entraînement sur une tâche de classification ainsi que les différentes tailles de modèles. En effet, l'équipe de L.Martin a démontré que le modèle CamemBERT standard (110 millions de paramètres) entraîné sur l'ensemble des 138Go de texte d'OSCAR, ne surpasse pas massivement le modèle entraîné « uniquement » sur l'échantillon de 4Go (Martin *et al.*, 2020) en étiquetage morphosyntaxique, en analyse syntaxique, en reconnaissance d'entités nommées (NER) et en reconnaissance d'implication textuelle (Natural Language Inference, NLI). Une perspective envisagée est de tester différents modèles de transformers francophones apparus depuis CamemBERT et FlauBERT comme les GPT2 (Generative Pretrained Transformer 2, OpenAI) francophones (BelGPT2, Pagnol, etc.) ou BARThez. Il serait également intéressant de procéder à une phase d'entraînement supplémentaire sur les 300000 anamnèses non-étiquetées à notre disposition avec un tokenizer spécifique aux anamnèses. De même, le traitement des notions médicales et des abréviations reste une piste d'amélioration. Le recours à des ontologies développées dans le domaine des urgences pourrait constituer une piste d'amélioration. Les transformers ont aussi récemment été testés pour l'identification et le remplacement des abréviations avec de bons résultats pour BERT (Kim *et al.*, 2020; Adams *et al.*, 2020), néanmoins il n'y a pas encore eu d'essai sur des données issues d'un mélange de langage courant et de termes médicaux en français.

Choix de la grille d'annotation Les performances des modèles deviennent acceptables lorsque nous avons exclu de notre base de test les notes cliniques dont l'interprétation nous paraissait ambiguë. La grille d'annotation créée pour le projet est ainsi en partie responsable certaines erreurs de classement dans le sens où il existe des zones de chevauchement sémantiques selon les classes. De plus, le système de codage utilisé ne permettait pas de coder plusieurs mécanismes traumatique (par exemple une collision entre deux individus , suivie

d'une chute). Cela explique que quand le top-2 est pris en compte, l'exactitude atteint la valeur de 0,976 pour FlauBERT. Pour permettre de rendre compte de ces situations, un nouveau système de codage sera utilisé pour les phases suivantes du projet, utilisant l'International Classification of External Causes of Injuries (ICECI).

5 Conclusion

Les transformers ont démontré une efficacité relative dans une tâche de classification multi-classe sur des données narratives médicales. Le choix de ce type de modèle dans le traitement automatique des résumés de passage aux urgences dans le but de créer un observatoire national doit être approfondi. La prochaine phase de notre projet consistera à utiliser un nouveau système de codage basé sur l'ICECI et à tester les transformers français avec une phase supplémentaire de pré-entraînement sur notre base de données de 300 000 anamnèses non-labellisées. L'expansion des acronymes, quant à elle, est à l'étude dans la chaîne de traitement d'automatisation. Les phases suivantes intégreront progressivement les données d'autres services d'urgences.

Remerciements

Le projet TARPON, porté par l'équipe Inserm *Injury epidemiology* et le service des urgences du CHU de Bordeaux en collaboration avec l'équipe Inria et Inserm SISTM, est lauréat du 2nd second appel à projets du Health Data Hub, Grand Défi "Amélioration des diagnostics médicaux par l'Intelligence Artificielle" et Bpifrance.

Références

ADAMS G., KETENCI M., BHAVE S., PEROTTE A. & ELHADAD N. (2020). Zero-shot clinical acronym expansion via latent meaning cells. In *Machine Learning for Health*, p. 12–40 : PMLR.

ALSENTZER E., MURPHY J. R., BOAG W., WENG W.-H., JIN D., NAUMANN T. & MCDERMOTT M. (2019). Publicly available clinical bert embeddings. *arXiv preprint arXiv :1904.03323*.

AMIN S., NEUMANN G., DUNFIELD K., VECHKAEVA A., CHAPMAN K. A. & WIXTED M. K. (2019). Mlt-dfki at clef ehealth 2019 : Multi-label classification of icd-10 codes with bert. *CLEF (Working Notes)*.

BELTAGY I., PETERS M. E. & COHAN A. (2020). Longformer : The long-document transformer. *arXiv preprint arXiv :2004.05150*.

BIRD, STEVEN E. L. & KLEIN E. (2009). nltk. In *Natural Language Processing with Python* : O'Reilly Media Inc.

COSSIN S., JOUHET V., MOUGIN F., DIALLO G. & THIESSARD F. (2018). Iam at clef ehealth 2018 : Concept annotation and coding in french death certificates. *arXiv preprint arXiv :1807.03674*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

FLICOTEAUX R. (2018). Ecstra-aphp@ clef ehealth2018-task 1 : Icd10 code extraction from death certificates. In *CLEF (Working Notes)*.

GOODRUM H., ROBERTS K. & BERNSTAM E. V. (2020). Automatic classification of scanned electronic health record documents. *International Journal of Medical Informatics*, **144**, 104302.

HONG Z., MAGDAMO C. G., SHEU Y.-H., MOHITE P., NOORI A., YE E. M., GE W., SUN H., BRENNER L., ROBBINS G. *et al.* (2020). Natural language processing to detect cognitive concerns in electronic health records using deep learning. *arXiv preprint arXiv :2011.06489*.

IVE J., VIANI N., CHANDRAN D., BITTAR A. & VELUPILLAI S. (2018). Kcl-health-nlp@ clef ehealth 2018 task 1 : Icd-10 coding of french and italian death certificates with character-level convolutional neural networks. In *CLEF (Working Notes)*.

KIM J., GONG L., KHIM J., WEISS J. C. & RAVIKUMAR P. (2020). Improved clinical abbreviation expansion via non-sense-based approaches. In *Machine Learning for Health*, p. 161–178 : PMLR.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2019). Flaubert : Unsupervised language model pre-training for french. *arXiv preprint arXiv :1912.05372*.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2020). Biobert : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

MARTIN L., MULLER B., JAVIER ORTIZ SUÁREZ P., DUPONT Y., ROMARY L., VILLEMONTE DE LA CLERGERIE E., SAGOT B. & SEDDAH D. (2020). Les modèles de langue contextuels Camembert pour le français : impact de la taille et de l'hétérogénéité des données d'entrainement. In *JEP-TALN-RECITAL 2020 - 33ème Journées d'Études sur la Parole, 27ème Conférence sur le Traitement Automatique des Langues Naturelles, 22ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, p. 54–65, Nancy, France : ATALA.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2019). Camembert : a tasty french language model. *arXiv preprint arXiv :1911.03894*.

METZGER M.-H., TVARDIK N., GICQUEL Q., BOUVRY C., POULET E. & POTINET-PAGLIAROLI V. (2017). Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts : a french pilot study. *International journal of methods in psychiatric research*, **26**(2), e1522.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

ROTHE S., NARAYAN S. & SEVERYN A. (2019). Leveraging pre-trained checkpoints for sequence generation tasks. *arXiv preprint arXiv :1907.12461*.

SANH V., DEBUT L., CHAUMOND J. & WOLF T. (2019). Distilbert, a distilled version of bert : smaller, faster, cheaper and lighter. *arXiv preprint arXiv :1910.01108*.

SEARLE T., IBRAHIM Z. & DOBSON R. (2020). Comparing natural language processing techniques for alzheimer's dementia prediction in spontaneous speech. *arXiv preprint arXiv :2006.07358*.

SHIBATA Y., KIDA T., FUKAMACHI S., TAKEDA M., SHINOHARA A., SHINOHARA T. & ARIKAWA S. (1999). *Byte Pair encoding : A text compression scheme that accelerates pattern matching*. Rapport interne, Technical Report DOI-TR-161, Department of Informatics, Kyushu University.

SPÄRCK JONES K. (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

SUAREZ P. J. O., SAGOT B. & ROMARY L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, p. 9 – 16, Mannheim : Leibniz-Institut für Deutsche Sprache.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. In *Advances in neural information processing systems*, p. 5998–6008.

WANG S., LI X., CHANG* X., YAO L., SHENG Q. Z. & LONG G. (2017). Learning multiple diagnosis codes for icu patients with local disease correlation mining. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **11**(3), 1–21.

WENZEK G., LACHAUX M.-A., CONNEAU A., CHAUDHARY V., GUZMAN F., JOULIN A. & GRAVE E. (2019). Ccnet : Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv :1911.00359*.

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). Huggingface's transformers : State-of-the-art natural language processing.

WU Y., SCHUSTER M., CHEN Z., LE Q. V., NOROUZI M., MACHEREY W., KRIKUN M., CAO Y., GAO Q., MACHEREY K. *et al.* (2016). Google's neural machine translation system : Bridging the gap between human and machine translation. *arXiv preprint arXiv :1609.08144*.

ZHONG J., GAO C. & YI X. (2018). Categorization of patient disease into icd-10 with nlp and svm for chinese electronic health record analysis. In *Proceedings of the 2018 International Conference on Artificial Intelligence and Pattern Recognition*, p. 101–106.