Extension et adaptation des modèles de langues pour la classification de corpus en santé animale

Edmond Menya*, Mathieu Roche **,***, Roberto Interdonato **,***, Dickson Owuor *

* CES Strathmore University, Nairobi, Kenya
{emenya, dowuor}@strathmore.edu
** CIRAD, F-34398 Montpellier, France
{mathieu.roche,roberto.interdonato}@cirad.fr
*** TETIS - Univ Montpellier - AgroParisTech - CIRAD - CNRS - INRAE, Montpellier, France

Résumé. Nous présentons EpidBioBERT, un classifieur de documents de biosurveillance épidémiologique. Notre modèle, entraîné sur un corpus qui contient des articles de presse sur les épidémies de maladies animales, a pour objectif de distinguer les documents pertinents et non pertinents pour une tâche d'extraction d'informations. Nous adoptons un modèle de langue biomédical pré-entraîné avec une approche de réglage fin, en nous concentrant sur les descripteurs thématiques épidémiologiques, à savoir la maladie, l'hôte, le lieu et la date. Nous expérimentons l'impact de chaque descripteur sur le classifieur dans le cadre d'études d'ablation. Nous comparons également notre approche biomédicale pré-entraînée avec un modèle de langue général.

1 Introduction

Ces dernières années, avec l'augmentation des épidémies de maladies infectieuses, l'accent a été mis sur l'intelligence épidémiologique et les systèmes intelligents de biosurveillance. Ces systèmes de surveillance épidémiologique ont gagné du terrain au sein de la communauté du traitement automatique du langage naturel (TALN). Ces systèmes sont capables de surveiller les sources numériques intergouvernementales officielles ainsi que les sources non officielles, par exemple les articles d'actualité publiés sur le Web, pour la détection précoce et le signalement des épidémies existantes, réémergentes et nouvelles (Woodall, 2001; Arsevska et al., 2018; Valentin et al., 2021). Les premiers systèmes de surveillance des maladies utilisaient généralement une approche fondée sur des indicateurs, c'est-à-dire des systèmes de règles formelles pour surveiller les sources officielles pertinentes (Paquet et al., 2006). Les systèmes de surveillance actuels, tels que ProMED, HealthMap et la plateforme d'extraction automatique d'informations sur les maladies animales à partir du Web PADI-web, utilisent une approche fondée sur les événements avec de multiples corpus et sources linguistiques différents (Woodall, 2001; Brownstein et Freifeld, 2007; Arsevska et al., 2018). PADI-web est un système de biosurveillance basé sur les événements et axé sur la surveillance des sources de dépêches en ligne pour la détection et l'alerte des maladies animales infectieuses existantes et émer-

gentes (Valentin et al., 2020, 2021). PADI-web 3.0 (Valentin et al., 2021) a récemment proposé une classification fine des phrases afin d'identifier des classes spécifiques (par exemple, épidémiologie descriptive, mesures de prévention et de contrôle, conséquences économiques et politiques).

Bien que l'intelligence épidémiologique se soit renforcée avec l'introduction de systèmes de surveillance épidémiologique fondés sur des événements, les principaux défis sont liés à la nécessité d'avoir à disposition des données étiquetées pour l'apprentissage supervisé. L'étiquetage de ces données est relativement coûteux et prend du temps. Afin d'entraîner un classifieur de documents épidémiologiques, les experts humains doivent étiqueter manuellement les articles d'actualité non structurés comme étant pertinents pour le traitement de la surveillance des maladies. Les corpus pertinents sont les articles d'actualité qui décrivent un événement lié à l'apparition d'une maladie animale infectieuse, les corpus non pertinents sont ceux qui ne sont pas liés à l'apparition de la maladie (Arsevska et al., 2018).

Ces dernières années, les techniques d'apprentissage profond fondées sur les plongements de mots (Mikolov et al., 2013) et les modèles de langues (Devlin et al., 2019) ont considérablement progressé. Cette étude vise à développer une nouvelle approche fondée sur le plongement thématique pour la classification de corpus épidémiologiques à partir d'articles d'actualité étiquetés. Le classifieur améliore les approches actuelles basées sur les mots-clés et l'apprentissage automatique. Notre classifieur de documents épidémiologiques apprend des plongements thématiques riches pour distinguer les dépêches d'actualité pertinentes et non pertinentes pour la surveillance des maladies du système PADI-web. Les contributions de cette étude sont les suivantes :

— Nous proposons EpidBioBERT, un modèle pré-entraîné sur le modèle de langue BioBERT (Lee et al., 2019) et affiné pour apprendre un classifieur qui discrimine les dépêches pertinentes et non pertinentes pour les tâches d'intelligence épidémiologique.

— Nous montrons que le réglage fin d'un modèle de langue biomédical améliore le classifieur de corpus épidémiologiques de manière plus significative que le réglage fin d'un modèle de langue pré-entraîné à usage général tel que BERT.

— Nous expérimentons l'impact de chaque descripteur thématique dans le classifieur épidémiologique global, montrant que les descripteurs thématiques de l'hôte et de la maladie contiennent des informations cruciales sur la pertinence du corpus pour l'intelligence épidémiologique.

— Nous améliorons la surveillance épidémiologique en évitant les fausses alertes positives dues à une mauvaise classification des dépêches d'actualité qui mentionnent des pays exempts de maladies et ceux qui décrivent les conséquences d'une épidémie.

Le présent article est un résumé de l'article publié dans la conférence "Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)" (Menya et al., 2022).

2 EpidBioBERT

Le modèle proposé, EpidBioBERT, adopte une approche d'apprentissage par transfert en deux étapes, utilisant un modèle de langue biomédical pré-entraîné suivi d'un réglage fin. Ce dernier processus consiste à améliorer la classification des documents épidémiologiques

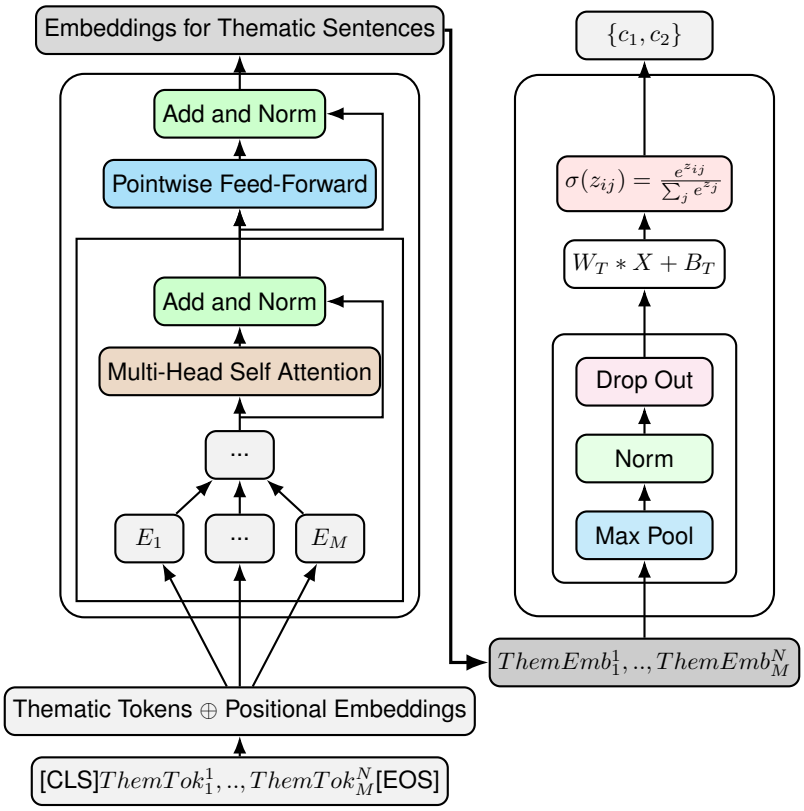

FIG. 1 – Architecture du transformeur EpidBioBERT avec des couches profondes finement ajustées sur un BioBERT pré-entraîné. [CLS]$ThemTok_1^1, .., ThemTok_M^N$[EOS] sont les N tokens de descripteurs thématiques provenant de M phrases du corpus d'entraînement annoté qui sont les entrées du modèle. [CLS] et [EOS] sont les étiquettes des tokenizers pour le début et la fin de la phrase respectivement. Une distribution de probabilité sur les classes de documents *pertinent* et *non pertinent* représentées par $\{c_1, c_2\}$ sont les étiquettes de sortie.

dans le système de surveillance des maladies PADI-web. L'architecture du modèle EpidBioBERT est fondé sur le modèle de langue biomédical BioBERT, affiné avec des données de surveillance des maladies (Fig. 1). BioBERT est une architecture qui s'appuie sur l'auto-attention (*self-attention*), pré-entraînée sur des corpus biomédicaux, atteignant des niveaux de pointe en fouille de textes biomédicaux. Trois versions pré-entraînées de BioBERT sont présentées dans (Lee et al., 2019), à savoir BioBERT(+PubMed), BioBERT(+PMC) et BioBERT(+PubMed+PMC). Ces versions diffèrent par la taille de leur architecture puisqu'elles sont pré-entraînées sur des jeux de données différents. Notre architecture est basée sur le modèle BioBERT(+PubMed) (Devlin et al., 2019; Lee et al., 2019). Dans notre approche, nous affinons d'abord l'ensemble du modèle BioBERT en dégelant (i.e., *unfreezing*) tous les poids et en utilisant le dernier état de l'optimiseur pré-entraîné pour effectuer un entraînement de

bout en bout sur le corpus PADI-web. La deuxième étape de réglage fin permet d'obtenir une classification similaire à celle des modèles de base. Nous adoptons une "loss fonction" d'entropie croisée sur les deux classes cibles. Notre modèle apprend à maximiser la probabilité des classes correctes (*pertinent/non pertinent*).

La tâche de classification de la pertinence des documents épidémiologiques prend en considération un ensemble de N dépêches notées $D = \{d_1, ..., d_N\}$. La tâche peut être définie de la manière suivante : étant données les dépêches $d_j \in D$ contenant n descripteurs thématiques épidémiologiques notés $F = \{f_1, ..., f_n\}$, notre approche produit une distribution de probabilité classant les articles selon une classe donnée, i.e. $C = \{c_1, c_2\}$ où c_1=*pertinent*, c_2=*non pertinent* pour la surveillance épidémiologique. Notre modèle apprend à maximiser la probabilité $p(c_i|d_j)$ où $c_i \in C$ et $d_j \in D$ en minimisant la fonction objective suivante :

$$L = \frac{1}{N_b} \, sum_i^{|C|} \sum_j^{|N_b|} -\{y_{ij} * \ln \sigma(z_{ij})\}$$

Dans ce contexte, b est la taille du batch défini comme hyperparamètre et y_{ij} est le vecteur d'étiquettes pour le corpus d'entraînement d_j avec les étiquettes affectées c_i. z_{ij} est la sortie de notre dernière couche linéaire telle que $z_j = W_T * X + B_T$ où W_T est la matrice de pondération des couches qui est multipliée par la matrice d'intégration des descripteurs thématiques épidémiologique X apprise avec le modèle pré-entraîné et avec l'ajout de la matrice de biais B_T. Sur la base de cette configuration, nous avons mis en œuvre EpidBioBERT avec un modèle pré-entrainé et des couches de réseau affiné, comme résumé en Figure 1. Notre proposition est évaluée dans la section suivante.

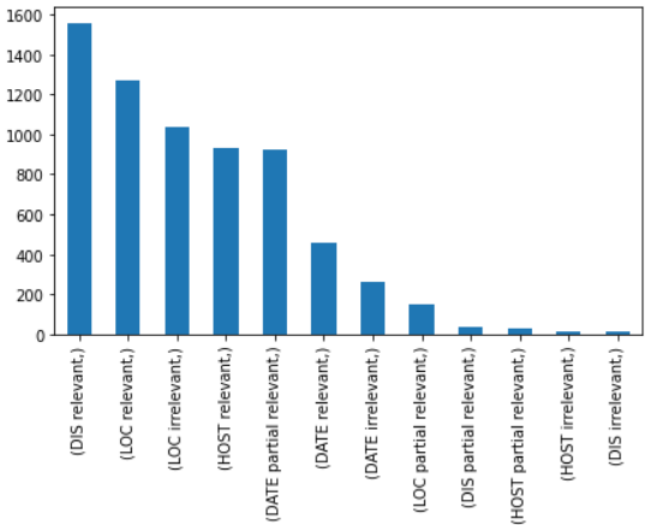

FIG. 2 – Distribution des descripteurs thématiques épidémiologiques dans PADI-web.

3 Expérimentations

3.1 Corpus

Notre corpus d'entraînement est dérivé du jeu de données PADI-web en exécutant une cascade de règles pour extraire les descripteurs épidémiologiques, leurs étiquettes annotées et les étiquettes des documents. Nous nous appuyons sur 180 dépêches (~35%) étiquetées comme pertinentes et 350 (~65%) comme non pertinentes. Les articles contiennent des entités épidémiologiques étiquetées manuellement par des experts humains. Chaque token considéré comme un candidat à une entité épidémiologique est initialement étiqueté comme suit : *location* pour le lieu du foyer, *date* pour la date de l'épidémie, *number* pour le nombre de cas signalés, *disease* pour le type de maladie rencontrée lors de l'épidémie et *host* pour l'espèce porteuse de la maladie. Nous constituons nos descripteurs thématiques épidémiologiques en sélectionnant tous ces descripteurs, à l'exception de *number*. En outre, ces entités épidémiologiques candidates sont étiquetées comme étant "correctes", "partielles" ou "incorrectes". Il existe $6K$ descripteurs thématiques avec 66% pertinentes, 20% non pertinentes et 14% partiellement pertinentes comme le montre la Fig. 2. L'ensemble des données résultant contient à la fois des documents pertinents et non pertinents composés de descripteurs épidémiologiques. Nous utilisons 60% des documents pour l'apprentissage, 20% comme corpus pour le réglage des hyperparamètres et les 20% qui restent pour l'évaluation des performances du modèle.

3.2 Méthodes concurrentes et cadre expérimental

Nous comparons EpidBioBERT avec des classifieurs d'apprentissage automatique récemment utilisés par PADI-web (Valentin et al., 2021) :

— approche par sac de mots avec des plongements de descripteurs épidémiologiques One Hot encoded (OHE) ;

— SVM boosté par un Kernel Gaussien (SVM+OHE) ;

— approche par sac de mots utilisant à la fois TF-IDF (SVM+TF-IDF) et les plongements thématiques GloVe pré-entraînés (SVM+GloVe) sur le même modèle SVM ;

— classifieur LSTM avec les plongements thématiques GloVe d'abord en gelant la couche de plongement (LSTM+GloVe$_{frozen}$), en entraînant un classifieur de bout en bout (LSTM+GloVe$_{unfrozen}$) et LSTM bidirectionnel (Bi-LSTM+GloVe$_{unfrozen}$).

En ce qui concerne le cadre expérimental, nous affinons le modèle BioBERT(+PubMed) avec une taille de plongement caché de 768, 12 Attention Heads et 12 Transformer blocks. Nous fixons un Batch de 16 et une longueur de séquence de 128 et expérimentons avec 50 époques. Pour nos couches plus fines, nous expérimentons des taux de Dropout de 0.2, 0.3 et 0.4 pour contrôler le surajustement du modèle. Nous adoptons l'optimiseur Adam avec décroissance de poids découplée (AdamW) (Loshchilov et Hutter, 2019) avec $\beta_1 = 0.9$ et $\beta_2 = 0.999$, nous définissons ϵ =1e-8 et décroissance de poids = 0.01. Nous fixons également de petits taux d'apprentissage initiaux de 1e-5 et 2e-5 avec un nombre d'époques plus élevé pour favoriser notre approche de réglage fin (Ruder, 2021). Nous évaluons et sauvegardons ensuite le meilleur modèle sur le corpus retenu.

Model	$F_1 Score$	$Precision$	$Recall$	$Accuracy$
Baselines				
SVM+OHE	0.29	**1**	0.17	70.00
SVM+TF-IDF	0.35	0.83	0.22	77.12
SVM+GloVe	0.51	0.65	0.55	65.34
LSTM+GloVe$_{frozen}$	0.84	0.84	0.85	86.13
LSTM+GloVe$_{unfrozen}$	0.85	0.85	0.85	87.12
Bi-LSTM+GloVe$_{unfrozen}$	0.86	0.89	0.85	88.11
Ours				
EpidBioBERT	**0.95**	0.97	**0.94**	**95.8**

TAB. 1 – Performance d'EpidBioBERT. Notons que le modèle fondé sur "One Hot Encoding" est nommé OHE. Les meilleurs scores sont en **gras**.

.

Thematic Feature	F_1 Score Drop	Precision Drop	Recall Drop	Accuracy Drop
Date	-4	-3	$\mathbf{-5}$	-4.8
Location	-1	-6	$+2$	-2
Host	$\mathbf{-8}$	$\mathbf{-18}$	$+2$	$\mathbf{-9.4}$
Disease	-5	-12	$+2$	-5.8

TAB. 2 – Impact de chaque descripteur sur la performance de EpidBioBERT. Les baisses de performance les plus importantes sont indiquées en **gras**.

3.3 Résultats

Les résultats de cette analyse expérimentale sont présentés dans Tab. 1. Les résultats encourageants obtenus peuvent être attribués aux architectures d'EpidBioBERT particulièrement performantes par rapport à tous les modèles de base. Tout d'abord, le modèle sous-jacent pré-entraîné BioBERT utilise l'architecture Transformer (Vaswani et al., 2017) qui dispose d'une meilleure gestion du contexte des descripteurs, la technique d'Attention surpasse les architectures LSTM. Deuxièmement, BioBERT est basé sur BERT (Devlin et al., 2019) qui est une architecture bidirectionnelle, ce qui permet de conserver les avantages de la bidirectionnalité. Troisièmement, nous affinons BioBERT(+PubMed) qui est pré-entraîné sur des données biomédicales, ce qui enrichit nos descripteurs thématiques épidémiologiques au-delà des capacités d'un modèle de langue général. Enfin, la nouvelle technique de réglage fin d'EpidBioBERT utilise des couches de réseau et des hyperparamètres qui favorisent l'apprentissage en peu de coups (*few shot learning*), ce qui compense nos données d'entraînement faibles en quantité et déséquilibrées en termes des classes.

Pour comprendre l'impact de chaque descripteur thématique dans notre modèle, nous exécutons le modèle EpidBioBERT sur différents ensembles de données, chacun ayant un descripteur thématique épidémiologique supprimé (Tab. 2). Nous nous concentrons sur quatre descripteurs thématiques, à savoir la maladie, l'hôte, le lieu et la date, et préparons quatre

ensembles d'entraînement, de validation et de test. Nous présentons les résultats de la diminution de performance en termes de F_1, précision, rappel et accuracy, par rapport aux résultats du modèle EpidBioBERT dans Tab. 1. Le descripteur thématique *Host* entraîne la plus forte baisse d'exactitude (accuracy) et F_1, respectivement de 9.4 et 8. L'entraînement du classifieur sans *Date* réduit la valeur de rappel des modèles, ce qui augmente l'erreur de classification des faux négatifs. *Location*, *Disease* et *Host* contiennent à peu près les mêmes informations pour influencer de manière égale la mesure de rappel, ce qui signifie moins de faux négatifs. Les résultats indiquent que l'hôte et la maladie sont les principaux descripteurs qui influencent notre classifieur, suivies de la date et du lieu. Ainsi l'ordre d'importance est différent (c'est-à-dire, maladie, lieu, hôte et date) de la représentation donnée en Fig. 2 fondée sur la fréquence des distributions des entités.

4 Conclusion et travaux futurs

Cet article présente un classifieur de documents épidémiologiques EpidBioBERT du système de biosurveillance des maladies animales infectieuses PADI-web. Notre contribution fondée sur BioBERT se concentre sur les descripteurs thématiques épidémiologiques des dépêches. Nos travaux ont également mis en avant que les modèles de langues biomédicaux contiennent des connaissances intrinsèques qui enrichissent et améliorent les systèmes de surveillance des maladies.

Cependant, nous avons également constaté qu'il existe peu de corpus de dépêches annotés pour cette tâche. Dans le cadre de travaux futurs, nous proposons un pipeline non supervisé qui peut prendre en compte des dépêches non étiquetées qui sont disponibles via d'autres plateformes du domaine, par exemple ProMED et HealthMap. Pour améliorer encore la classification des corpus, nous proposons d'intégrer des connaissances sémantiques, par exemple Medical Subject Headings (Mesh), pour enrichir davantage les descripteurs thématiques. Nous proposons également d'approfondir la recherche sur l'impact individuel des descripteurs thématiques dans les systèmes de surveillance des maladies.

Remerciements : Cette étude a été partiellement financée par la subvention européenne 874850 MOOD. Le contenu de cette publication relève de la seule responsabilité des auteurs et ne reflète pas nécessairement les vues de la Commission européenne.

Références

Arsevska, E., S. Valentin, J. Rabatel, J. de Goër de Hervé, S. Falala, R. Lancelot, et M. Roche (2018). Web monitoring of emerging animal infectious diseases integrated in the french animal health epidemic intelligence system. *PLOS ONE 13*, 1–25.

Brownstein, J. S. et C. Freifeld (2007). Healthmap : the development of automated real-time internet surveillance for epidemic intelligence. *Weekly releases (1997–2007) 12*(48), 3322.

Devlin, J., M. Chang, K. Lee, et K. Toutanova (2019). BERT : pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT)*, pp. 4171–4186.

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, et J. Kang (2019). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics 36*(4), 1234–1240.

Loshchilov, I. et F. Hutter (2019). Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Menya, E., M. Roche, R. Interdonato, et D. Owuor (2022). Enriching epidemiological thematic features for disease surveillance corpora classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pp. 3741–3750.

Mikolov, T., K. Chen, G. Corrado, et J. Dean (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Paquet, C., D. Coulombier, R. Kaiser, et M. Ciotti (2006). Epidemic intelligence : a new framework for strengthening disease surveillance in europe. *Eurosurveillance 11*(12), 5–6.

Ruder, S. (2021). Recent Advances in Language Model Fine-tuning. `http://ruder.io/recent-advances-lm-fine-tuning`.

Valentin, S., E. Arsevska, S. Falala, J. de Goër, R. Lancelot, A. Mercier, J. Rabatel, et M. Roche (2020). Padi-web : A multilingual event-based surveillance system for monitoring animal infectious diseases. *Computers and Electronics in Agriculture 169*, 105163.

Valentin, S., E. Arsevska, J. Rabatel, S. Falala, A. Mercier, R. Lancelot, et M. Roche (2021). Padi-web 3.0 : A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance. *One Health 13*, 100357.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, et I. Polosukhin (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008.

Woodall, J. P. (2001). Global surveillance of emerging diseases : the promed-mail perspective. *Cadernos de saude publica 17*, S147–S154.

Summary

We present EpidBioBERT, an epidemiological biomonitoring document classifier. Our model, trained on a corpus containing news articles on animal disease outbreaks, aims to distinguish relevant and irrelevant documents for an information extraction task. We adopt a pre-trained biomedical language model with a fine-tuning approach, focusing on the epidemiological thematic features, namely disease, host, location and date. We experiment with the impact of each feature on the classifier in ablation studies. We also compare our pre-trained biomedical approach with a general language model.