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-