AVIS DE PRESENTATION DE THESE EN SOUTENANCE POUR L'OBTENTION DU DIPLOME NATIONAL DE DOCTEUR

Monsieur Stephen MUTUVI

Présentera ses travaux intitulés :

« Extraction d'événements épidémiologiques dans un contexte multilingue et à faibles ressources »

Spécialité : Informatique et applications

Le 21 novembre 2022 à 14h00

Lieu :

La Rochelle Université
Pôle Communication, Multimédia et Réseaux
Amphithéâtre Michel Crépeau
44 Av. Albert Einstein
17000 LA ROCHELLE

Composition du jury :

Mme BOROS Emanuela(*Invitée*)	**Ingénieure de recherche, La Rochelle Université**
M. DOUCET Antoine	**Professeur, La Rochelle Université**
M. JATOWT Adam (*Invité*)	**Professeur, Université d'Innsbruck**
M. LEJEUNE Gaël	**Maître de conférences, Sorbonne Université**
Mme NÉVÉOL Aurèlie	**Directrice de recherche, Université Paris-Saclay**
M. ODEO Moses	**Professeur, Multimedia University of Kenya**
M. PISKORSKI Jakub	**Research associate, Polish Academy of Sciences**
M. ROCHE Mathieu	**Directeur de recherche, AgroParisTech**
M. TORRES -MORENO Juan-Manuel	**Maître de conférences, HDR, Avignon université**
Mme VILNAT Anne	**Professeure, Université Paris Saclay**

Résumé :

L'extraction d'événements épidémiques a pour but d'extraire de textes des incidents d'importance pour la santé publique, tels que des épidémies. Alors que l'extraction d'événements a fait l'objet de recherches approfondies pour les langues à fortes ressources comme l'anglais, les systèmes existants d'extraction d'événements épidémiques ne sont pas optimaux pour les contextes multilingues à faibles ressources en raison de la rareté des données d'entraînement. Tout d'abord, nous nous attaquons au problème de la rareté des données en transformant et en annotant un ensemble de données multilingues existantes au niveau des documents en un ensemble de données annotées au niveau des jetons, adapté à l'apprentissage supervisé des séquences. Ensuite, nous formulons la tâche d'extraction d'événements comme une tâche d'étiquetage de séquences et nous utilisons l'ensemble de données annotées au niveau des jetons pour entraîner des modèles supervisés d'apprentissage automatique et profond pour l'extraction d'événements épidémiques. Les résultats montrent que les modèles linguistiques pré-entraînés ont produit la meilleure performance globale dans toutes les langues évaluées. Troisièmement, nous proposons une technique d'adaptation au domaine en incluant des entités épidémiologiques (noms de maladies et lieux) dans le vocabulaire des modèles pré-entraînés. L'incorporation de ces entités a eu un impact positif sur la qualité de la tokénisation, contribuant ainsi à l'amélioration des performances du modèle. Enfin, nous évaluons l'auto-formation et observons que l'approche est légèrement plus performante que les modèles formés par apprentissage supervisé.