Étude comparative de méthodes de classification multilingue appliquées à l'épidémiologie

Stephen Mutuvi, Emanuela Boros, Antoine Doucet, Gaël Lejeune, Adam Jatowt, Moses Odeo

Étude comparative de méthodes de classification multilingue appliquées à l'épidémiologie

Stephen Mutuvi[1,2] — Emanuela Boros[2] — Antoine Doucet[2] — Gaël Lejeune[3] — Adam Jatowt[4] — Moses Odeo[1]

[1] *Multimedia University, Kenya*
[2] *Université de La Rochelle, France*
[3] *Sorbonne Université, France*
[4] *Innsbruck Universität, Autriche*

RÉSUMÉ. *Dans cet article, nous abordons la tâche de classification multilingue de textes dans le domaine épidémiologique. Nous comparons différents modèles d'apprentissage automatique et d'apprentissage profond à l'aide d'un jeu de données multilingue comprenant des articles de presse en six langues. Notre objectif est d'analyser l'influence de la famille de langue, de la structure du document et de la taille des données sur les résultats de classification. Nos résultats indiquent que les performances des modèles basés sur des modèles linguistiques dépassent de plus de 50% les baselines, parmi lesquelles un système spécialisé de surveillance épidémiologique et plusieurs modèles d'apprentissage automatique.*

ABSTRACT. *In this paper, we approach the multilingual text classification task in the context of the epidemiological field. We conduct a comparative study of different machine and deep learning text classification models using a dataset comprising news articles related to epidemic outbreaks from six languages, four low-resourced and two high-resourced, in order to analyze the influence of the nature of the language, the structure of the document, and the size of the data. Our findings indicate that the performance of the models based on fine-tuned language models exceeds by more than 50% the chosen baseline models that include a specialized epidemiological news surveillance system and several machine learning models.*

MOTS-CLÉS : *Extraction d'information, Jeux de données multilingues, Classification de texte*

KEYWORDS: *Information extraction, Multilingual datasets, Text classification*

1. Introduction

La surveillance et l'endiguement des épidémies de maladies constituent un défi permanent à l'échelle mondiale. Que ce soit auparavant avec Ebola ou aujourd'hui avec la pandémie de Covid-19, la veille sur des documents issus du Web reste un élément clé de la stratégie de santé publique. La capacité de détecter le déclenchement d'épidémies de manière précise et opportune est essentielle au déploiement de mesures d'intervention efficaces et motive la recherche de solutions multilingues visant à repérer des informations dans le plus de langues possible. La prolifération des sources de données numériques offre une voie pour la surveillance basée sur les données, dénommée *Epidemic Intelligence*. Le renseignement sur les épidémies implique la collecte, l'analyse et la diffusion d'informations clés liées aux épidémies, dans le but de détecter les épidémies et d'alerter rapidement les acteurs de la santé publique (World Health Organization, 2014). Les techniques de traitement automatique du langage naturel (TALN) ont permis d'analyser différentes données issues de sources Web, telles que les médias sociaux, les requêtes de recherche, les blogs et les articles de presse en ligne pour les incidents et / ou événements liés à la santé (Salathé *et al.*, 2013 ; Bernardo *et al.*, 2013). Compte tenu de ces défis, des approches appropriées sont nécessaires pour que la surveillance épidémiologique fondée sur les données soit couronnée de succès. Dans cet article, nous proposons une étude quantitative des modèles de classification multilingue de texte. Nous appliquons ces modèles à un jeu de données de référence comprenant des articles de presse de plusieurs langues : l'anglais, le grec, le français, le russe, le polonais et le chinois. Nous comparons différentes approches de classification appliquées à la surveillance épidémiologique, définie ici comme la détection dans des flux de presse de documents décrivant des épidémies. Le travail présenté ici est décrit plus en détail dans (Mutuvi *et al.*, 2020a).

2. Jeu de données

Nous proposons une extension du jeu de données proposé dans (Mutuvi *et al.*, 2020b) pour inclure des langues supplémentaires afin de couvrir plusieurs familles linguistiques : germanique (anglais, en), hellénique (grec, el), romane (français, fr), slave (russe, ru et polonais, pl) et enfin sino-tibétaine (chinois, zh). Les articles de presse contenus dans ce corpus ont été obtenus à partir de différentes sources d'information en ligne. La source principale est le site Web du *Program for Monitoring Emerging Disease* (PROMED) [1], qui est un programme de l'*International Society for Infectious Diseases* chargé de suivre la propagation des maladies infectieuses à travers le monde. Ceci nous a permis de collecter plus de documents pertinents que les jeux de données existants, notamment en anglais. Nous avons divisé les données, avec un total de 7574 articles, en ensembles d'entraînement, de validation et de test. La répartition par langue est présentée dans le tableau 1.

1. `https://promedmail.org/`

	Ensemble	Polonais	Chinois	Russe	Grec	Français	Anglais
Entraînement	5.074 (10,8%)	241 (7,4%)	300 (2,6%)	296 (9,5%)	253 (6,7%)	1.593 (10,9%)	2.365 (11,7%)
Validation	1.250 (10,9%)	54 (7,4%)	71 (2,8%)	60 (10%)	68 (10,2%)	388 (13,4%)	583 (12,6%)
Test	1.250 (10,5%)	46 (13%)	75 (6%)	70 (10%)	63 (4,7%)	434 (12,4%)	614 (12,8%)

Tableau 1 : Nombre de documents (dont pertinents) pour chaque langue.

3. Expériences

Les métriques que nous avons utilisées pour l'évaluation des modèles sont la précision, le rappel et la F-mesure. Notons que, s'agissant d'épidémies le rappel, et donc l'élimination des faux négatifs, est particulièrement important.

Comme modèle de base, nous utilisons DANIEL [2] (Lejeune *et al.*, 2015), un système non supervisé qui ne repose sur aucune analyse grammaticale spécifique à la langue, et qui considère notamment le texte comme une séquence de caractères, afin de limiter les pré-traitements dépendants de la langue. Nous étudions également trois modèles de classification de textes couramment utilisés comme *baselines* : régression logistique (LR), forêt d'arbres aléatoires (RF) et machine à vecteurs supports (SVM). Nous utilisons les hyper-paramètres par défaut, et la pondération TF-IDF [3].

Nous considérons enfin deux modèles plus élaborés : un réseau neuronal convolutif (CNN) et un réseau neuronal récurrent (BiLSTM) en exploitant les plongements de mots FastText (Joulin *et al.*, 2016). Nous avons choisi de réaliser également des expériences avec différentes architectures basées sur BERT (Devlin *et al.*, 2018).Nous testons enfin une approche basée sur les réseaux convolutifs de graphes (GCN) qui enrichit BERT avec des plongements de graphes (VGCN + BERT) (Lu et Nie, 2019). Les paramètres de ces modèles sont définis dans (Mutuvi *et al.*, 2020a).

4. Résultats

Les résultats compilés dans le tableau 2 montrent que parmi les *baselines* le SVM dépasse par une petite marge les LR et RF en F-mesure. De manière intéressante, le classifieur RF obtient la plus haute précision (95,70 %) de tous les modèles utilisés. On observe que les modèles d'apprentissage automatique classiques (LR, RF et SVM) donnent des résultats fortement déséquilibrés, avec une excellente précision mais un rappel assez faible. Comparé aux résultats de base fournis par DANIEL, ce modèle spécialisé donne un rappel meilleur que la précision, son rappel reste toutefois le moins élevé parmi toutes les méthodes comparées ce qui s'explique sans doute par son caractère non-supervisé : il ne tire pas bénéfice des données d'apprentissage.

Ensuite, les modèles basés sur des architectures CNN ou BiLSTM avec intégration des embeddings FASTTEXT ont des scores en F-mesure inférieurs à ceux des

2. https://github.com/NewsEye/event-detection/tree/master/event-detection-daniel
3. SCIKIT-LEARN, https://scikit-learn.org/

Modèles	Précision %	Rappel %	F-mesure %
DANIEL	33,9	60,61	43,48
Logistic Regression	93,81	68,94	79,48
Random Forest	**95,70**	67,42	79,11
Support Vector Machine	91,26	71,21	80
CNN+FastTtext	86,11	70,45	77,5
BiLSTM+FastTtext	77,44	78,03	77,74
BERT (`cased`)†	88,62	82,58	85,49
CNN+BERT (`cased`)†	88,79	71,97	79,5
BiLSTM+BERT (`cased`)†	90,20	69,70	78,63
BERT (`uncased`)†	84,67	87,88	**86,25**
CNN+BERT (`uncased`)†	82,14	87,12	84,56
BiLSTM+BERT (`uncased`)†	83,72	81,82	82,76
BERT (`cased`)	80,71	85,61	83,09
CNN+BERT (`cased`)	86,67	78,79	82,54
BiLSTM+BERT (`cased`)	75,95	**90,91**	82,76
BERT (`uncased`)	88,52	81,82	85,04
CNN+BERT (`uncased`)	86,07	79,55	82,68
BiLSTM+BERT (`uncased`)	81,51	73,48	77,29
VGCN+BERT	87,18	77,27	81,93

Tableau 2 : Évaluation des modèles analysés pour l'ensemble des langues du jeu de données, le modèle BERT est le modèle `base-multilingual`, † indique les modèles *fine-tuned*

méthodes d'apprentissage automatique classiques (LR, RF, SVM). Cela pourrait s'expliquer par le fait que les données d'apprentissage sont trop petites pour garantir une bonne précision. Dans le cas des modèles d'apprentissage profond, on peut remarquer une grande différence en terme de F-mesure des modèles basés sur BERT, par rapport à tous les autres. On peut également observer que les modèles basés sur BERT parviennent à équilibrer rappel et précision, la précision restant stable malgré l'augmentation du rappel.

5. Conclusion et perspectives

Dans cet article, nous avons proposé une étude détaillée des performances de différentes méthodes de classification appliquées à la veille épidémiologique multilingue. Les résultats que nous avons présenté suggèrent que les approches basées sur des modèles de langage raffinés et/ou des réseaux convolutifs de graphes obtiennent de très bonnes performances sur cette tâche de classification.

Remerciements

Ce travail a été financé par le programme de recherche et d'innovation Horizon 2020 de l'Union européenne au titre des subventions 770299 (NewsEye) et 825153 (Embeddia).

6. Bibliographie

Bernardo T. M., Rajic A., Young I., Robiadek K., Pham M. T., Funk J. A., « Scoping review on search queries and social media for disease surveillance : a chronology of innovation », *Journal of medical Internet research*, vol. 15, n° 7, p. e147, 2013.

Devlin J., Chang M.-W., Lee K., Toutanova K., « Bert : Pre-training of deep bidirectional transformers for language understanding », *arXiv preprint arXiv :1810.04805*, 2018.

Joulin A., Grave E., Bojanowski P., Douze M., Jégou H., Mikolov T., « FastText.zip : Compressing text classification models », *arXiv preprint arXiv :1612.03651*, 2016.

Lejeune G., Brixtel R., Doucet A., Lucas N., « Multilingual event extraction for epidemic detection », *Artificial intelligence in medicine*, vol. 65, n° 2, p. 131-143, 2015.

Lu Z., Nie J.-Y., « RALIGRAPH at HASOC 2019 : VGCN-BERT : Augmenting BERT with Graph Embedding for Offensive Language Detection », *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation*, 2019.

Mutuvi S., Boros E., Doucet A., Jatowt A., Lejeune G., Odeo M., « Multilingual Epidemiological Text Classification : A Comparative Study », *in* D. Scott, N. Bel, C. Zong (eds), *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, International Committee on Computational Linguistics, p. 6172-6183, 2020a.

Mutuvi S., Doucet A., Lejeune G., Odeo M., « A Dataset for Multi-lingual Epidemiological Event Extraction », *Proceedings of The 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 4139-4144, May, 2020b.

Salathé M., Freifeld C. C., Mekaru S. R., Tomasulo A. F., Brownstein J. S., « Influenza A (H7N9) and the importance of digital epidemiology », *The New England journal of medicine*, vol. 369, n° 5, p. 401, 2013.

World Health Organization, Early detection, assessment and response to acute public health events : implementation of early warning and response with a focus on event-based surveillance : interim version, Technical report, World Health Organization, 2014.