Neural language models for text classification in evidence-based medicine

Anonymous Author(s)
Affiliation
Address
email

Abstract

COVID-19 has brought about a significant challenge to the whole of humanity, but mainly to the medical community. Clinicians must keep updated continuously about symptoms, diagnoses, and effectiveness of emergent treatments under a never-ending flood of scientific literature. In this context, the role of evidence-based medicine (EBM) for curating the most substantial evidence to support public health and clinical practice turns especially essential but is being challenged as never before. Artificial Intelligence can have a crucial role in this situation. In this article, we report the results of an applied research project to classify scientific articles to support Epistemonikos, one of the essential foundations worldwide conducting EBM. We test several methods, and the best one, based on XLNet, improves the current approach by 93% on average F1-score, saving valuable time from physicians who volunteer to curate COVID-19 research articles manually.

1 Introduction

Evidence-based medicine (EBM) is a medical practice that aims to find all the evidence to support medical decisions. This evidence nowadays is obtained from biomedical journals, usually accessible through online databases like PubMed[3] and EMBASE[2], which provide free access to articles' abstracts and in some cases, to full articles. In the context of the COVID-19 pandemic, EBM is critical to making decisions at the individual level and public health since research articles address topics like treatments, adverse cases, and effects of public policies in medicine. The EBM foundation Epistemonikos has made essential contributions by curating and publishing updated guides of what treatments are working and not to treat COVID-19 [1]. Epistemonikos addresses EBM by a combination of software tools for data collection, storage, filtering, and retrieval, as well as by the vital labor of volunteer physicians who curate and label research articles based on quality (to include in the database), type (systematic review, randomized trial, among others) and PICO labels (patient, intervention, comparison, outcome). However, this workflow has been challenged during 2020 by increasing growth and rapidly evolving evidence of COVID-19 articles published in the latest months. Moreover, to ensure the rapid collection of the latest evidence published, pre-print repositories such as medRXiv and bioRXiv have been added to the traditional online databases.

In order to support Epistemonikos' effort to filter and curate the flood of articles related to COVID-19, we present the results of an applied AI project where we implement and evaluate a text classification system to filter and categorize research articles related to COVID-19. The current model, based on Random Forests, has an acceptable performance classifying systematic reviews (SR) but fails on classifying other document categories. In this article, we show how using BioBERT yields marginal improvements, while XLNET results in significant progress with the best performance. These results

[1]

save a considerable amount of time from volunteer physicians by pre-filtering the articles worth of manual curation and labeling for EBM.

2 Methods and results

2.1 Methods and data

We compare document classification results using random forest with a customized tokenizer made by Epistemonikos, an XLNET [6] language model representing documents using a linear layer as a classifier and the same setting with a BioBERT [1]language model. The documents' classification can be a systematic review, a primary study using a randomized controlled trial, non-randomized primary study, broad synthesis, and excluded document. The distribution of documents can be observed in the second column of Table 1. Notice that the type of document partially explains the classification models' mistakes: broad synthesis and systematic review are both kinds of surveys, while primary studies (rct and non-rct) deal with specific treatments and populations. Excluded can be of any of the other four classes, but they are not included in the official Epistemonikos dataset due to their low quality.

2.2 Results

Table 1 shows the performance of each model in terms of precision (Prec.), recall (Rec.), and f1-score (F-1) for every type of document. In general terms, we observe that XLNet obtains the top F-1 score for any category of a document, in some cases by a small margin, such as under systematic review (F-1=.97), and in other cases by a large margin, as in the classes Broad synthesis (F-1=.61), and Excluded (F-1=.78). The results indicate that the random forest and BioBERT with a linear layer have a bias towards the most dominant class, Systematic review, reporting slightly better recall (R=.99 and R=1.0)than XLNet (R=.98) in this particular type of document. However, XLNet is better than the other two models in terms of Precision upon all classes, with the only exception of Broad synthesis, where random forest ($P = .75$) performs better than XLNet ($P = .67$). However, XLNet improves ($R = .56$) upon random forest ($R = .15$) in terms of recall. It is important to note that when using the random forest implemented for Epistemonikos, a new tokenizer has to be made depending on the document categories. In the case of XLNET, it is more versatile because it is enough to train embeddings and classify them regardless of the document category. In the case of BioBERT, which has a similar operation, it does not yield consistent performance for the minority classes Broad synthesis and excluded.

Table 1: Distribution of document and results obtained for document classification of Broad Synthesis, Systematic Review, Primary Study randomized controlled trial (Primary rct), Primary Study non-randomized controlled trial (Primary non-rct), and Excluded.

		Random Forest			XLNet			BioBERT		
	# docs.	Prec.	Rec.	F-1	Prec.	Rec.	F-1	Prec.	Rec.	F-1
Broad synthesis	17,324	.75	.15	.26	.67	.56	**.61**	0	0	0
Systematic review	286,050	.93	.99	.96	.96	.98	**.97**	.85	1.0	.92
Primary rct	56,623	.25	.79	.38	.94	.85	**.89**	.71	.71	.71
Primary non-rct	35,644	.63	.40	.49	.64	.91	**.75**	.61	.90	.72
Excluded	6,096	.70	.21	.32	.82	.74	**.78**	0	0	0

3 Conclusion

In this study, we have compared three methods, one of which is currently in production at the Epistemonikos foundation, the random forest. The others are BioBERT, which, although it is based on the transformer architecture, does not achieve the results shown by XLNET. Having such reliable results can mean a big impact in times of the COVID-19 pandemic where there is an exponential growth of available literature. In future work we will incorporate explanations obtained from transformer attention mechanisms, compare them against other explnation methods like LIME[5] or SHAP[4], and conduct a user study to assess whether physicians' work is facilitated by this feature.

Broader Impact

This work seeks to decrease manual effort in the practice of evidence-based medicine, allowing physicians us to distinguish relevant documents for clinical questions. Implementing the method with the largest performance in our offline evaluation (XLNet) in production might imply an increased cost in terms of GPU needs for Epistemonikos, which is not under their current infrastructure. Adding more documents might also imply additional fine-tuning of the model, incurring in larger costs. Another aspect not addressed in this research is that of Fairness: is the current model performing better to classify certain populations being treated (e.g. white males) compared to black females? we should address this aspect actively to prevent our model from learning undesired biases already seen in several applications.

References

[1] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[2] Carol Lefebvre, Anne Eisinga, Steve McDonald, and Nina Paul. Enhancing access to reports of randomized trials published world-wide–the contribution of embase records to the cochrane central register of controlled trials (central) in the cochrane library. *Emerging Themes in Epidemiology*, 5(1):13, 2008.

[3] Wesley T Lindsey and Bernie R Olin. Pubmed searches: Overview and strategies for clinicians. *Nutrition in Clinical Practice*, 28(2):165–176, 2013.

[4] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[6] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763, 2019.