LETTER

Open Access

Semi-automating abstract screening with a natural language model pretrained on biomedical literature

Sheryl Hui-Xian Ng[1*], Teow Kiok Liang[1], Gary Yee Ang[1], Tan Woan Shin[1†] and Allyn Hum[2,3†]

Abstract

We demonstrate the performance and workload impact of incorporating a natural language model, pretrained on citations of biomedical literature, on a workflow of abstract screening for studies on prognostic factors in end-stage lung disease. The model was optimized on one-third of the abstracts, and model performance on the remaining abstracts was reported. Performance of the model, in terms of sensitivity, precision, F1 and inter-rater agreement, was moderate in comparison with other published models. However, incorporating it into the screening workflow, with the second reviewer screening only abstracts with conflicting decisions, translated into a 65% reduction in the number of abstracts screened by the second reviewer. Subsequent work will look at incorporating the pre-trained BERT model into screening workflows for other studies prospectively, as well as improving model performance.

Keywords Abstract, Classification, Semi-automation

Introduction

In recent years, there has been a growth in interest in using artificial intelligence methods in systematic reviews (SRs) [1], in particular for the stage of literature screening [2]. As the number of titles and abstracts to be screened for suitability for inclusion in a review often involves numerous hours of repetitive work, semi-automation of this stage has been suggested to deliver workload and time savings with acceptable recall and precision [3–5].

One approach targets the automated classification of studies for inclusion using prediction models. In recent work, Aum et al. developed a Bidirectional Encoder Representations from Transformer (BERT) algorithm that was pretrained on published SRs and fine-tuned on another SR, with good classification performance. The authors recommended generalizing the use of BERT-based models for this purpose, by pre-training with information from a particular clinical domain and optimizing the predictions for the individual review only at the last fine-tuning step [6]. In this letter, we demonstrate the performance and workload impact of incorporating a BERT model pretrained on citations of biomedical literature in our own abstract screening workflow.

Methods

We used abstracts retrieved from a previous literature search on prognostic factors in end-stage lung disease [7]. Bibliographic databases such as MEDLINE, Embase, PubMed, CINAHL, Cochrane Library and Web of

[†]Tan Woan Shin and Allyn Hum are joint senior authors.

*Correspondence:
Sheryl Hui-Xian Ng
sheryl_hx_ng@nhg.com.sg
[1] Health Services and Outcomes Research, National Healthcare Group, 3 Fusionopolis Link, #03-08, Singapore 138543, Singapore
[2] Department of Palliative Medicine, Tan Tock Seng Hospital, 11 Jalan Tan Tock Seng, Singapore 308433, Singapore
[3] The Palliative Care Centre for Excellence in Research and Education, Dover Park Hospice, 10 Jalan Tan Tock Seng, Singapore 308436, Singapore

Science were searched using a pre-defined search strategy and inclusion criteria (Additional file 1). A total of 21,645 abstracts were retrieved, and based on screening by reviewers, 530 (2.5%) of the studies were included in the subsequent stage, where the full text of the articles was retrieved for thorough reading.

The dataset of 21,645 abstracts consisted of the text within the abstract, excluding the title, as well as an indication of whether the abstract was classified as included by the human reviewers. For model validation, the dataset was split into a training set of 7142 abstracts (33%), and a test set of 14,503 abstracts (67%). We then used the training set to fine-tune a BERT model pretrained on citations from MEDLINE/PubMed (pBERT). A batch size of 64 was used, and convergence over 100 epochs was assessed [8].

We then applied the fine-tuned pBERT to the test set and labelled 2.5% of articles with the highest predicted probabilities of inclusion as included in the review by pBERT. Based on this set of labels, we assessed sensitivity, precision, F1 and accuracy of pBERT, as well as the proportion of conflicts and level of inter-rater agreement, which was measured by Cohen's kappa (Table 1). We also report the reduction in workload in a hypothetical scenario where pBERT performs screening as the second reviewer for 67% of the articles.

Results

Of the 14,503 abstracts in the test set, the human reviewers deemed 355 (2.5%) to be relevant and suitable for inclusion in the subsequent stage of review. Sensitivity, precision and F1 of pBERT were 37.7%, while disagreement occurred for 3.0% of all articles screened. Cohen's Kappa was 0.70, indicating moderate agreement between the reviewers and pBERT (Table 1).

In the traditional screening process, each of the two human reviewers would have to screen all 21,645 articles for relevance to the study, before reviewing any conflicts in their decisions. With pBERT incorporated into the screening workflow, both reviewers would screen the first 33% of articles, and pBERT would be fine-tuned based on this training set.

For the remaining 14,503 articles, the first reviewer (R1) would screen all the articles in accordance with the traditional workflow. pBERT would then replace the second human reviewer (R2) in identifying studies for inclusion, while R2 would only step in to review articles with conflicting decisions. In this scenario, R1 and pBERT would have agreed on decisions for 14,061 articles, leaving 442 articles (3% of 14,503) for R2 to review. Hence, R2 would have to review only 7,584 articles (7,142 + 442) or 35% of the original 21,645 articles.

Discussion

We applied a BERT model pretrained on biomedical literature to our data, with moderate model performance. Having a sensitivity of 37.7% entails that pBERT can only be used as an assistant alongside a human reviewer to increase the efficiency of screening, as opposed to being a standalone tool for automation of screening. While pBERT did not perform as well in terms of traditional metrics compared to recent models [6, 9], our dataset did have a lower inclusion rate of 2.5%, compared to 11% and 19% in both studies, impacting the predictive ability of the model.

Nonetheless, despite the constrained performance of pBERT, we were able to demonstrate that incorporating pBERT in our workflow would have reduced the workload of a second human reviewer to a third of the initial volume. Our results suggest that involving predictive tools to screen out irrelevant articles, which often comprise the bulk of the abstracts, can improve efficiency of screening processes in comparison to traditional approaches. However, while there is interest to fully automate the task of screening without human intervention, we emphasize that the role of a human reviewer remains pertinent to ensure all potentially relevant articles are included in the study [10].

Table 1 List of performance measures assessed

Measure	Definition	Estimate
Recall/sensitivity	$\frac{\text{Number of abstracts included by human reviewer and pBERT}}{\text{Number of abstracts included by human reviewer}}$	37.7%
Precision/positive predictive value	$\frac{\text{Number of abstracts included by human reviewer and pBERT}}{\text{Number of abstracts included by pBERT}}$	37.7%
F1	$2 \times \frac{precision \times recall}{precision + recall}$	37.7%
Accuracy	$\frac{\text{Number of abstracts included by human reviewer and pBERT}}{\text{Total number of abstracts screened}}$	70.2%
Disagreement	$\frac{\text{Number of abstracts with different decisions by human reviewer and pBERT}}{\text{Total number of abstracts screened}}$	3.0%

Conclusion

For semi-automation of screening of literature on prognostic factors in end-stage lung disease, we used a BERT model trained on biomedical literature to identify abstracts that were relevant to the topic and demonstrated a substantial reduction in screening workload. Subsequent work will look at integrating the current version of pBERT into screening workflows for other studies prospectively, as well as incorporating other ensemble methods to develop models with improved sensitivity to identify abstracts of relevance to the research question.

Abbreviations

AI	Artificial intelligence
BERT	Bidirectional Encoder Representations from Transformer
pBERT	PubMed BERT
R1	Reviewer 1
R2	Reviewer 2
SR	Systematic reviews

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13643-023-02353-8.

Additional file 1. Search criteria and strategy. **Appendix 1.** Study eligibility criteria. **Appendix 2.** Search strategy.

Acknowledgements

The authors acknowledge the contributions by the Medical Library team from the Lee Kong Chian School of Medicine, National Technological University, Singapore, for creating and optimizing the database search. The authors also acknowledge the support on this project rendered by our colleague, Eric Chua, Senior Executive, from the Health Services and Outcomes Research Department, National Healthcare Group, Singapore.

Authors' contributions

SHXN drafted the manuscript. TKL analysed the data. GYA, TWS, and AH made major revisions to the draft. TWS and AH acquired the data.

Funding

This work was supported by the National Medical Research Council, Singapore (grant number HSRGEoL18may-0003). The funder had no involvement in any aspect of this study or decision to publish.

Availability of data and materials

Data arising from the review may be made available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Ethics approval was granted by the review board of the National Healthcare Group as part of a larger study on prognosticating end-stage organ failure (Domain-Specific Review Board Study Reference No. 2019/00032). Consent to participate is not applicable for this study.

Consent for publication

Consent for publication is not applicable for this study.

Competing interests

The authors declare that they have no competing interests.

Received: 11 August 2023 Accepted: 13 September 2023
Published online: 23 September 2023

References

1. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019;8(1):163.
2. Blaizot A, Veettil SK, Saidoung P, Moreno-Garcia CF, Wiratunga N, Aceves-Martins M, et al. Using artificial intelligence methods for systematic review in health sciences: a systematic review. Res Synth Methods. 2022;13(3):353–62.
3. Gates A, Guitard S, Pillay J, Elliott SA, Dyson MP, Newton AS, et al. Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. Syst Rev. 2019;8(278).
4. Gates A, Gates M, DaRosa D, Elliott SA, Pillay J, Rahman S, et al. Decoding semi-automated title-abstract screening: findings from a convenience sample of reviews. Syst Rev. 2020;9(272).
5. Feng Y, Liang S, Zhang Y, Chen S, Wang Q, Huang T, et al. Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis. J Am Med Inform Assoc. 2022;29(8):1425–32.
6. Aum S, Choe S. srBERT: automatic article classification model for systematic review using BERT. Syst Rev. 2021;10(285).
7. Ng SHX, Chai GT, Gunapal PPG, Kaur P, Yip WF, Chiam ZY, et al. Prognostic factors of mortality in non-COPD chronic lung disease: a scoping review. J Palliat Med. 2023. https://doi.org/10.1089/jpm.2023.0263.
8. TensorFlow Hub. TF2.0 Saved Model (v2). 2023 (Available from: https://tfhub.dev/google/experts/bert/pubmed/2).
9. Qin X, Liu J, Wang Y, Liu Y, Deng K, Ma Y, et al. Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews. J Clin Epidemiol. 2021;133:121–9.
10. Popoff E, Besada M, Jansen JP, Cope S, Kanters S. Aligning text mining and machine learning algorithms with best practices for study selection in systematic literature reviews. Syst Rev. 2020;9(293).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.