Protocol

A Deep Learning Approach to Refine the Identification of High-Quality Clinical Research Articles From the Biomedical Literature: Protocol for Algorithm Development and Validation

Wael Abdelkader[1], MSc, MD; Tamara Navarro[1], MLiS; Rick Parrish[1], Dip T; Chris Cotoi[1], BEng, EMBA; Federico Germini[1,2], MSc, MD; Lori-Ann Linkins[2], MSc, MD; Alfonso Iorio[1,2], MD, PhD; R Brian Haynes[1,2], MD, PhD; Sophia Ananiadou[3,4], BA, DEA, PhD; Lingyang Chu[5], BSc, PhD; Cynthia Lokker[1], MSc, PhD

[1]Health Information Research Unit, Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

[2]Department of Medicine, McMaster University, Hamilton, ON, Canada

[3]Department of Computer Science, University of Manchester, Manchester, United Kingdom

[4]The Alan Turing Institute, London, United Kingdom

[5]Department of Computing and Software, Faculty of Engineering, McMaster University, Hamilton, ON, Canada

Corresponding Author:
Cynthia Lokker, MSc, PhD
Health Information Research Unit
Department of Health Research Methods, Evidence, and Impact
McMaster University
1280 Main St W, CRL-137
Hamilton, ON, L8S 4K1
Canada
Phone: 1 9055259140 ext 22208
Email: lokkerc@mcmaster.ca

Abstract

Background: A barrier to practicing evidence-based medicine is the rapidly increasing body of biomedical literature. Use of method terms to limit the search can help reduce the burden of screening articles for clinical relevance; however, such terms are limited by their partial dependence on indexing terms and usually produce low precision, especially when high sensitivity is required. Machine learning has been applied to the identification of high-quality literature with the potential to achieve high precision without sacrificing sensitivity. The use of artificial intelligence has shown promise to improve the efficiency of identifying sound evidence.

Objective: The primary objective of this research is to derive and validate deep learning machine models using iterations of Bidirectional Encoder Representations from Transformers (BERT) to retrieve high-quality, high-relevance evidence for clinical consideration from the biomedical literature.

Methods: Using the HuggingFace Transformers library, we will experiment with variations of BERT models, including BERT, BioBERT, BlueBERT, and PubMedBERT, to determine which have the best performance in article identification based on quality criteria. Our experiments will utilize a large data set of over 150,000 PubMed citations from 2012 to 2020 that have been manually labeled based on their methodological rigor for clinical use. We will evaluate and report on the performance of the classifiers in categorizing articles based on their likelihood of meeting quality criteria. We will report fine-tuning hyperparameters for each model, as well as their performance metrics, including recall (sensitivity), specificity, precision, accuracy, F-score, the number of articles that need to be read before finding one that is positive (meets criteria), and classification probability scores.

Results: Initial model development is underway, with further development planned for early 2022. Performance testing is expected to star in February 2022. Results will be published in 2022.

Conclusions: The experiments will aim to improve the precision of retrieving high-quality articles by applying a machine learning classifier to PubMed searching.

International Registered Report Identifier (IRRID): DERR1-10.2196/29398

(JMIR Res Protoc 2021;10(11):e29398) doi: 10.2196/29398

KEYWORDS

bioinformatics; machine learning; evidence-based medicine; literature retrieval; medical informatics; natural language processing; NLP; biomedical; literature; literature surveillance; model development

Introduction

Background

The biomedical literature grows exponentially every year. According to the latest National Library of Medicine statistical report, more than 1.5 million new citations were indexed in PubMed in 2020 alone [1]. This high volume of literature is fed by the publication of at least 1 new article every 26 seconds [2] and 95 clinical trials per day [3]. Nevertheless, only 1% of published clinical studies meet the criteria for high scientific quality for use in health care decisions [4], driving the need for efficient and accurate approaches to identify clinical studies that have been conducted with methodological rigor.

Methodological search filters like PubMed Clinical Queries [5] are considered the cornerstone for information retrieval in evidence-based practice [6]. These and other search filters have been developed using a diagnostic testing procedure [7] to optimize sensitivity or specificity, or the best balance between the two, for such clinical study categories as treatment, diagnosis, prognosis, etiology, and clinical prediction guides [8]. Some search filters are limited by their partial reliance on MeSH (Medical Subject Headings) indexing terms, as it can take up to a year for articles to be indexed in MEDLINE [9]. Despite having highly sensitive search filters, with an aim to optimize specificity—essentially returning the most likely relevant articles while reducing the need to assess off-target articles—they also return large numbers of articles that are not on target.

Most search filters have been developed using databases of articles that have been tagged for clinical category and methodological rigor and using a diagnostic test approach to detect true and false, positive (on-target) and negative (off-target) articles [10]. The development of such a gold standard is costly and time-consuming, requiring highly trained staff. The Clinical Hedges database, developed at McMaster University, has been used as the gold standard for new search strategy development [11-14].

Machine learning is a subset of artificial intelligence referring to the application of computational methods to improve performance or achieve precise predictions via experience. Experience, in this context, is the information given to the machine for analysis [15]. Machine learning applications in the biomedical literature have been explored by many researchers over the years. In 2007, Yoo and colleagues [16] applied a novel machine learning approach for document clustering and text summarization, producing a textual summary of the information by automatically extracting the most relevant text content from a document cluster. Machine learning has also been applied to the ranking of biomedical publications. In 2009, the MedlineRanker webserver ranked citations in a data set based on their relevance to a given topic [17]. Other applications of machine learning include accurately predicting the citation count of a given article at the time of its publication to determine its scientific impact using a support vector machine (SVM) context–based classifier [18] and automating the systematic review screening process to decrease the screening workload [19]. For example, Miwa and colleagues [20] used an SVM pool–based active machine learning model to classify articles as relevant for inclusion in a systematic review. Miwa et al's [20] experiment used “certainty” as a criterion for article selection, which is effective in dealing with imbalanced data sets. An improvement in topic detection was proposed by Hashimoto et al [21] as they used a neural network model based on paragraph vectors capturing semantic similarities between words and documents. Paragraph vectors can accurately determine semantic relatedness between textual units of varying lengths, that is, words, phrases, and longer sequences (eg, sentences, paragraphs, and documents). Methods that consider factors such as word order within the text yield superior performance [21].

Recent advancement in natural language processing (NLP) is attributed to the development of pretrained language models (PTLMs) [22]. PTLMs transfer learning from training on one data set to the performance of different NLP functions on a new data set [22,23]. PTLMs provide more stable predictions and better model generalization [24]. PTLMs are applied using one of two main strategies: feature-based or fine-tuning models [23]. Feature-based approaches use task-specific architectures that include the pretrained representations as additional features, such as Embedding from Language Models (ELMo) [22,25]. Fine-tuning approaches attempt to pretrain the language model using general-domain text, then fine-tune the model on the target data and target task [22]. Fine-tuning language models are considered the mainstream for PTLM adaption [22]. Examples of fine-tuning approaches are Universal Language Model Fine-Tuning [26] and Bidirectional Encoder Representations from Transformers (BERT) [23]. The bidirectional approach used in BERT improves its performance in understanding text context over other PTLMs, making it the state-of-the-art model. BERT can be used for multiple NLP tasks, including text summarization, retrieval, question answers, named entity recognition, and document classification [27].

Over the past two decades, machine learning has been applied to classify the biomedical literature based on methodological rigor and evidence quality. For such classification tasks, supervised machine learning approaches in which the training data is labeled based on a selected high-quality standard are most commonly used [3,28-32]. The first reported experiments to classify biomedical literature based on quality by Aphinyanaphongs and colleagues relied on the American College of Physicians Journal Club as their high-quality training standard and used a supervised SVM as a classifier [28-30]. The most recent study to classify high-quality articles was conducted by Afzal and colleagues [31] and applied an artificial neural network using data gathered from the Cochrane Library as their high-quality standard. By using supervised approaches, the model development was informed by decisions made by the researchers.

Objectives

The primary objective of this research is to derive and validate deep learning models using variations of BERT to retrieve high-quality, high-relevance evidence for clinical consideration from the biomedical literature; models will be trained using a large, tagged database of high-quality, high-relevance clinical articles.

Methods

Quality Standard Derivation

At McMaster University, the Health Information Research Unit (HiRU) has an established reputation for retrieval, appraisal, classification, organization, and dissemination of health-related research. Through the Knowledge Refinery, the unit daily screens research studies from over 120 clinical journals and identifies those that meet methodological rigor for original studies, systematic reviews, pooled original studies, and evidence-based guidelines within the categories of treatment, primary prevention, diagnosis, harm from medical interventions, economics, prognosis, clinical prediction, and quality improvement [33]. The steps in the process include the initial filtering of all journal articles using highly sensitive search filters (>99%) developed by HiRU to identify articles that fit the categories named above. This filtered subset is then manually reviewed by skilled research associates and a clinical editor. In this project, “rigor” is defined as meeting all the methodological criteria explicitly described on the HiRU website and in Multimedia Appendix 1 [34]. The process of selecting clinically relevant articles is further described by Haynes et al [35], and the high reliability of the critical appraisal step has been documented with a kappa value of over 80% for all categories of articles [36].

The Premium Literature Service (PLUS) process is based on scientific principles for critical appraisal of the medical literature to support evidence-based medicine, combined with multiple ratings of clinical relevance by a worldwide network of practicing health care professionals. Segments of the database have been used many times to test various machine learning approaches, including deep learning [3]. A vast community of >4000 clinicians then rate methodologically rigorous articles for clinical relevance and newsworthiness [37]. The resulting PLUS database contains a distillation of the most reliable and relevant published clinical research [38].

Data Set

The data used is the Critical Appraisal Process (CAP) data set, which consists of the titles and abstracts of 155,679 articles published between 2012 to 2020, identified by means of their PubMed identifier and manually labeled by research associates as those that “fulfilled” methodological rigor criteria (n=30,035) or “failed” to meet methodological rigor criteria (n=125,644). The data set will be randomly split into 80% for training, 10% for validation, and 10% for testing. Along with being unbalanced, the CAP data set is large and computationally challenging for deep learning model development. To overcome this limitation, we will first convert the data set into multiple balanced subsets, then independently train one model per each of the balanced subsets and use ensembling techniques [39-42] to combine the independently trained models into a better model with more robust performance.

Machine Learning Experiment

Using Python (Python Software Foundation), we will build our models using the HuggingFace Transformers library [43]. HuggingFace is an open-source NLP and artificial intelligence model hub that provides accessible and implementable state-of-the-art models to the community [44]. Using PTLMs available within the HuggingFace Transformers library, we will experiment with variations of BERT models to determine which have the best performance in article classification. These will include BERT [23], BioBERT [45], BlueBERT [46], and PubMedBERT [47]. These models differ in the pretraining text domain. Pretraining a biomedical BERT model follows a mixed-domain pretraining that initializes with standard BERT pretraining using text data from BookCorpus [48] and English Wikipedia (Wikimedia Foundation) [23], followed by continuous pretraining using biomedical text. BioBERT is pretrained using PubMed abstracts and PubMed Central full-text articles [45], while BlueBERT is pretrained using PubMed text and clinical notes from MIMIC-III (Medical Information Mart for Intensive Care) [46]. PubMedBERT is pretrained using domain-specific text data from a collection of 14 million PubMed abstracts, which were downloaded in February 2020, with abstracts under 128 words removed [47]. Our selection of these models was guided by their availability within the HuggingFace repository and their reported performance in the Biomedical Language Understanding and Reasoning Benchmark leaderboard [47,49].

For the top-performing model that maintains sensitivity >98%, we plan to prospectively validate its real-world performance in the McMaster PLUS reading process. A random sample of incoming articles that are classified as failed articles will be allocated to research staff blinded to the model determination.

To evaluate the performance of machine learning models, we will report the sensitivity (recall), specificity, accuracy, precision, the number of articles that need to be read before finding one that is positive, and F-score (harmonic mean of recall and precision metrics [50]) (Table 1). We will report the probability score threshold, with corresponding 95% CIs, for each model. The machine learning models return a probability score for each article that represents the probability that the article is of high quality, and ranges from 0 (does not meet criteria) to 1 (meets criteria). For a given article, the probability will vary depending on the composition of the model. To prospectively validate the performance of the best model, we will report the same diagnostic characteristics for prospective validation of the model. Fine-tuning hyperparameter settings (number of epochs, learning rate, batch size, and number of random seeds) of the selected models for validation will be reported.

Table 1. Definitions and formulas pertaining to performance metrics.

Measure	Definition	Formula
Recall (sensitivity)	The proportion of correctly identified positive articles fulfilling criteria among those predicted to be positive	TP[a]/(TP+FN[b])
Specificity	The proportion of articles correctly identified as not meeting criteria among those predicted as negative	TN[c]/(TN+FP[d])
Precision	The proportion of correctly identified positives among all classified positives	TP/(TP+FP)
F-measure	Harmonic mean of precision and recall	2 × ([precision × recall]/[precision + recall])
Accuracy	The number of correctly predicted documents out of all classified documents	(TP+TN)/(TP+FP+FN+TN)
Number needed to read	The number of articles that need to be read before finding one that is positive (meets criteria)	1/precision

[a] TP: true positive.

[b] FN: false negative.

[c] TN: true negative.

[d] FP: false positive.

Results

Initial model development is underway, with further development planned for early 2022. Performance testing is expected to start in February 2022. Results will be published in 2022.

Discussion

BERT is considered the state-of-the-art model for NLP. To our knowledge, this is the first experiment to investigate the use of PTLMs in the identification of high-quality articles from the biomedical literature [51]. Our study leverages a large data set of over 150,000 citations that have been manually tagged by experienced research associates, making it one of the few reliable sources for training machine learning models to identify high-quality clinical literature [50]. Our application and analysis of BERT models may provide a better performing automation model suitable for incorporation in literature surveillance processes at HiRU and elsewhere.

Artificial intelligence and machine learning applications are complex and known for their black-box nature, providing predictions without enough explanation [52]. Besides the accurate prediction and the decrease in workload, trust in algorithmic decisions is essential, especially in medicine and health care research [53]. To overcome the lack of transparency, interpreting machine learning models and their decision-making process has become a growing focus among academic and industrial machine learning experts [54]. Next steps include interpreting the decisions made by the model. This would allow us to understand the justification behind model decision-making [55].

Conflicts of Interest

None declared.

Multimedia Appendix 1

Inclusion criteria for articles meeting methodological rigor.
[DOCX File , 24 KB-Multimedia Appendix 1]

References

1. MEDLINE PubMed Production Statistics. National Library of Medicine. 2020. URL: https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html [accessed 2021-07-31]
2. Garba S, Ahmed A, Mai A, Makama G, Odigie V. Proliferations of scientific medical journals: a burden or a blessing. Oman Med J 2010 Oct;25(4):311-314 [FREE Full text] [doi: 10.5001/omj.2010.89] [Medline: 22043365]
3. Del Fiol G, Michelson M, Iorio A, Cotoi C, Haynes RB. A Deep Learning Method to Automatically Identify Reports of Scientifically Rigorous Clinical Research from the Biomedical Literature: Comparative Analytic Study. J Med Internet Res 2018 Jun 25;20(6):e10281 [FREE Full text] [doi: 10.2196/10281] [Medline: 29941415]
4. Haynes R. Where's the meat in clinical journals? ACP Journal Club 1993;119(3):A22. [doi: 10.7326/ACPJC-1993-119-3-A22]
5. Wilczynski NL, McKibbon KA, Walter SD, Garg AX, Haynes RB. MEDLINE clinical queries are robust when searching in recent publishing years. J Am Med Inform Assoc 2013;20(2):363-368 [FREE Full text] [doi: 10.1136/amiajnl-2012-001075] [Medline: 23019242]

6. Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. J Am Med Inform Assoc 2002;9(6):653-658 [FREE Full text] [doi: 10.1197/jamia.m1124] [Medline: 12386115]
7. Wilczynski NL, Morgan D, Haynes RB, Hedges Team. An overview of the design and methods for retrieving high-quality studies for clinical care. BMC Med Inform Decis Mak 2005 Jun 21;5:20 [FREE Full text] [doi: 10.1186/1472-6947-5-20] [Medline: 15969765]
8. Hedges. Health Information Research Unit - McMaster University. 2016. URL: https://hiru.mcmaster.ca/hiru/HIRU_Hedges_home.aspx [accessed 2021-11-18]
9. Irwin AN, Rackham D. Comparison of the time-to-indexing in PubMed between biomedical journals according to impact factor, discipline, and focus. Res Social Adm Pharm 2017;13(2):389-393. [doi: 10.1016/j.sapharm.2016.04.006] [Medline: 27215603]
10. Wong SS, Wilczynski NL, Haynes RB, Ramkissoonsingh R, Hedges Team. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. AMIA Annu Symp Proc 2003:728-732 [FREE Full text] [Medline: 14728269]
11. Geersing G, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KGM, et al. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS One 2012;7(2):e32844 [FREE Full text] [doi: 10.1371/journal.pone.0032844] [Medline: 22393453]
12. Frazier JJ, Stein CD, Tseytlin E, Bekhuis T. Building a gold standard to construct search filters: a case study with biomarkers for oral cancer. J Med Libr Assoc 2015 Jan;103(1):22-30 [FREE Full text] [doi: 10.3163/1536-5050.103.1.005] [Medline: 25552941]
13. Keogh C, Wallace E, O'Brien KK, Murphy PJ, Teljeur C, McGrath B, et al. Optimized retrieval of primary care clinical prediction rules from MEDLINE to establish a Web-based register. J Clin Epidemiol 2011 Aug;64(8):848-860. [doi: 10.1016/j.jclinepi.2010.11.011] [Medline: 21411285]
14. Lee E, Dobbins M, Decorby K, McRae L, Tirilis D, Husson H. An optimal search filter for retrieving systematic reviews and meta-analyses. BMC Med Res Methodol 2012 Apr 18;12:51 [FREE Full text] [doi: 10.1186/1471-2288-12-51] [Medline: 22512835]
15. Mohri M, Rostamizadeh A, Talwalkar A. Foundations of Machine Learning, 2nd ed. Cambridge, MA: MIT Press; 2018.
16. Yoo I, Hu X, Song I. A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method. BMC Bioinformatics 2007 Nov 27;8 Suppl 9:S4 [FREE Full text] [doi: 10.1186/1471-2105-8-S9-S4] [Medline: 18047705]
17. Fontaine J, Barbosa-Silva A, Schaefer M, Huska M, Muro E, Andrade-Navarro M. MedlineRanker: flexible ranking of biomedical literature. Nucleic Acids Res 2009 Jul;37(Web Server issue):W141-W146 [FREE Full text] [doi: 10.1093/nar/gkp353] [Medline: 19429696]
18. Fu LD, Aliferis CF. Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. Scientometrics 2010 Feb 3;85(1):257-270. [doi: 10.1007/s11192-010-0160-5]
19. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev 2015 Jan 14;4(1):5 [FREE Full text] [doi: 10.1186/2046-4053-4-5] [Medline: 25588314]
20. Miwa M, Thomas J, O'Mara-Eves A, Ananiadou S. Reducing systematic review workload through certainty-based screening. J Biomed Inform 2014 Oct;51:242-253 [FREE Full text] [doi: 10.1016/j.jbi.2014.06.005] [Medline: 24954015]
21. Hashimoto K, Kontonatsios G, Miwa M, Ananiadou S. Topic detection using paragraph vectors to support active learning in systematic reviews. J Biomed Inform 2016 Aug;62:59-65 [FREE Full text] [doi: 10.1016/j.jbi.2016.06.001] [Medline: 27293211]
22. Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X. Pre-trained models for natural language processing: A survey. Sci China Technol Sci 2020 Sep 15;63(10):1872-1897. [doi: 10.1007/s11431-020-1647-3]
23. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online Oct 1, 2018 [FREE Full text]
24. Dai AM, Le QV. Semi-supervised Sequence Learning. arXiv Preprint posted online Nov 4, 2015 [FREE Full text]
25. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. arXiv Preprint posted online Feb 14, 2018 [FREE Full text] [doi: 10.18653/v1/n18-1202]
26. Howard J, Ruder S. Universal Language Model Fine-tuning for Text Classification. arXiv Preprint posted online Jan 18, 2018 [FREE Full text] [doi: 10.18653/v1/p18-1031]
27. Adhikari A, Ram A, Tang R, Lin J. DocBERT: BERT for Document Classification. arXiv Preprint posted online Apr 17, 2019 [FREE Full text]
28. Aphinyanaphongs Y, Aliferis CF. Text categorization models for retrieval of high quality articles in internal medicine. AMIA Annu Symp Proc 2003:31-35 [FREE Full text] [Medline: 14728128]
29. Aphinyanaphongs Y, Aliferis C. Prospective validation of text categorization filters for identifying high-quality, content-specific articles in MEDLINE. AMIA Annu Symp Proc 2006:6-10 [FREE Full text] [Medline: 17238292]

30. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis C. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inform Assoc 2005;12(2):207-216 [FREE Full text] [doi: 10.1197/jamia.M1641] [Medline: 15561789]
31. Afzal M, Park BJ, Hussain M, Lee S. Deep Learning Based Biomedical Literature Classification Using Criteria of Scientific Rigor. Electronics 2020 Aug 05;9(8):1253. [doi: 10.3390/electronics9081253]
32. Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards automatic recognition of scientifically rigorous clinical research evidence. J Am Med Inform Assoc 2009;16(1):25-31 [FREE Full text] [doi: 10.1197/jamia.M2996] [Medline: 18952929]
33. McMaster Health Knowledge Refinery. Health Information Research Unit - McMaster University. 2016. URL: https://hiru.mcmaster.ca/hiru/HIRU_McMaster_HKR.aspx [accessed 2021-05-31]
34. Inclusion Criteria. Health Information Research Unit - McMaster University. 2019. URL: https://hiru.mcmaster.ca/hiru/InclusionCriteria.html [accessed 2021-05-31]
35. Haynes RB, Cotoi C, Holland J, Walters L, Wilczynski N, Jedraszewski D, McMaster Premium Literature Service (PLUS) Project. Second-order peer review of the medical literature for clinical practitioners. JAMA 2006 Apr 19;295(15):1801-1808. [doi: 10.1001/jama.295.15.1801] [Medline: 16622142]
36. Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB. Assessment of methodologic search filters in MEDLINE. Proc Annu Symp Comput Appl Med Care 1993:601-605 [FREE Full text] [Medline: 8130545]
37. McMaster Online Rating of Evidence (MORETM). McMaster Premium Literature Service (PLUS) Project, McMaster University. 2018. URL: http://hiru.mcmaster.ca/more/ [accessed 2021-05-31]
38. McMaster Premium LiteratUre Service (PLUS). Health Information Research Unit - McMaster University. 2016. URL: https://hiru.mcmaster.ca/hiru/HIRU_McMaster_PLUS_projects.aspx [accessed 2021-05-31]
39. Ganaie M, Hu M, Tanveer M, Suganthan P. Ensemble deep learning: A review. arXiv Preprint posted online Apr 6, 2021 [FREE Full text]
40. Araque O, Corcuera-Platas I, Sánchez-Rada JF, Iglesias CA. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Systems with Applications 2017 Jul;77:236-246. [doi: 10.1016/j.eswa.2017.02.002]
41. Xu Y, Qiu X, Zhou L, Huang X. Improving BERT Fine-Tuning via Self-Ensemble and Self-Distillation. arXiv Preprint posted online Feb 24, 2020 [FREE Full text]
42. Xu C, Barth S, Solis Z. Applying Ensembling Methods to BERT to Boost Model Performance. Stanford University. URL: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/default/15775971.pdf [accessed 2021-11-18]
43. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-Art Natural Language Processing. 2020 Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; Oct 2020; Online p. 38-45 URL: https://huggingface.co/transformers/ [doi: 10.18653/v1/2020.emnlp-demos.6]
44. HuggingFace. URL: https://huggingface.co/ [accessed 2021-11-16]
45. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
46. Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. arXiv Preprint posted online Jun 13, 2019 [FREE Full text] [doi: 10.18653/v1/w19-5006]
47. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. arXiv Preprint posted online Jul 30, 2020 [FREE Full text] [doi: 10.1145/3458754]
48. Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading book. 2015 Presented at: 2015 IEEE International Conference on Computer Vision (ICCV); December 7-13, 2015; Santiago, Chile p. 19-27 URL: https://www.computer.org/csdl/proceedings-article/iccv/2015/8391a019/12OmNro0HYa [doi: 10.1109/ICCV.2015.11]
49. BLURB Leaderboard. Microsoft. 2020. URL: https://microsoft.github.io/BLURB/index.html#page-top [accessed 2021-11-16]
50. Bekkar M, Djema H, Alitouche T. Evaluation measures for models assessment over imbalanced data sets. J Info Eng Appl 2013;3(10):27-38 [FREE Full text]
51. Abdelkader W, Navarro T, Parrish R, Cotoi C, Germini F, Iorio A, et al. Machine Learning Approaches to Retrieve High-Quality, Clinically Relevant Evidence From the Biomedical Literature: Systematic Review. JMIR Med Inform 2021 Sep 09;9(9):e30401 [FREE Full text] [doi: 10.2196/30401] [Medline: 34499041]
52. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ 2019 Mar 12;364:l886. [doi: 10.1136/bmj.l886] [Medline: 30862612]
53. Hoeren T, Niehoff M. Artificial Intelligence in Medical Diagnoses and the Right to Explanation. European Data Protection Law Review 2018;4(3):308-319. [doi: 10.21552/edpl/2018/3/9]
54. Du M, Liu N, Hu X. Techniques for Interpretable Machine Learning. arXiv Preprint posted online Jul 31, 2018 [FREE Full text] [doi: 10.1145/3359786]
55. Chakraborty S, Tomsett R, Raghavendra R, Harborne D, Alzantot M, Cerutti F, et al. Interpretability of deep learning models: A survey of results. 2017 Presented at: IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced &

Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI); Aug 4-7, 2017; San Francisco, CA. [doi: 10.1109/uic-atc.2017.8397411]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
CAP: Critical Appraisal Process
HiRU: Health Information Research Unit
MeSH: Medical Subject Headings
MIMIC-III: Medical Information Mart for Intensive Care
NLP: natural language processing
PLUS: Premium Literature Service
PTLM: pretrained language model
SVM: support vector machine

Edited by G Eysenbach; submitted 07.04.21; peer-reviewed by M Michelson, P Berchialla, A Thompson; comments to author 25.05.21; revised version received 24.08.21; accepted 17.09.21; published 29.11.21

Please cite as:
Abdelkader W, Navarro T, Parrish R, Cotoi C, Germini F, Linkins LA, Iorio A, Haynes RB, Ananiadou S, Chu L, Lokker C
A Deep Learning Approach to Refine the Identification of High-Quality Clinical Research Articles From the Biomedical Literature: Protocol for Algorithm Development and Validation
JMIR Res Protoc 2021;10(11):e29398
URL: https://www.researchprotocols.org/2021/11/e29398
doi: 10.2196/29398
PMID: