Journal of Clinical Epidemiology 133 (2021) 121–129

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews

Xuan Qin[a], Jiali Liu[a], Yuning Wang[a], Yanmei Liu[a], Ke Deng[a], Yu Ma[a], Kang Zou[a], Ling Li[a,*], Xin Sun[a,b,*]

[a] *Chinese Evidence-based Medicine Center, Cochrane China Center and National Clinical Research Center for Geriatrics, West China Hospital, Sichuan University, Chengdu 610041, Sichuan, China*

[b] *Evidence-based Medicine Research Center, School of Basic Science, Jiangxi University of Traditional Chinese Medicine, Nanchang 330004, Jiangxi, China*

Accepted 14 January 2021; Available online 21 January 2021

Abstract

Background and Objective: To examine whether the use of natural language processing (NLP) technology is effective in assisting rapid title and abstract screening when updating a systematic review.

Study Design: Using the searched literature from a published systematic review, we trained and tested an NLP model that enables rapid title and abstract screening when updating a systematic review. The model was a light gradient boosting machine (LightGBM), an ensemble learning classifier which integrates four pretrained Bidirectional Encoder Representations from Transformers (BERT) models. We divided the searched citations into two sets (ie, training and test sets). The model was trained using the training set and assessed for screening performance using the test set. The searched citations, whose eligibility was determined by two independent reviewers, were treated as the reference standard.

Results: The test set included 947 citations; our model included 340 citations, excluded 607 citations, and achieved 96% sensitivity, and 78% specificity. If the classifier assessment in the case study was accepted, reviewers would lose 8 of 180 eligible citations (4%), none of which were ultimately included in the systematic review after full-text consideration, while decreasing the workload by 64.1%.

Conclusion: NLP technology using the ensemble learning method may effectively assist in rapid literature screening when updating systematic reviews.

Key Words: Machine learning; Natural language processing; Systematic review update; Title and abstract screening; BERT; LightGBM

Conflict of Interest: None.

Author contributions: XQ implemented the ensemble learning LightGBM integrating four BERT models, wrote the source codes, evaluated the model performance, and drafted the manuscript. JL, YW, KD, YM, and LL helped prepare the data. YL participated in the statistical analysis design. KZ was the project administration. LL and XS provided helpful feedback and revisions to the paper.

Funding: This study was supported by the National Key R&D Program of China (grant no. 2019YFC1709804 and 2017YFC1700406), National Natural Science Foundation of China (grant no. 71904134), Sichuan Youth Science and Technology Innovation Research Team (grant no. 2020JDTD0015), China Postdoctoral Science Foundation (grant no. 2019M653444) and 1•3•5 project for disciplines of excellence, West China Hospital, Sichuan University (grant no. ZYYC08003).

Role of the Funder/Sponsors: None of the funders had any role in the design and conduct of the study; collection, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

* Corresponding author. Tel./fax: 86-028-85164187.

E-mail addresses: liling@wchscu.cn (L. Li), sunxin@wchscu.cn (X. Sun).

https://doi.org/10.1016/j.jclinepi.2021.01.010

What is new?

Key findings

- The ensemble learning method, consisting of advanced natural language processing (NLP) models, can be used to develop models that assist in automatic title and abstract screening when updating systematic reviews.
- The developed model may achieve rapid title and abstract screening, with high sensitivity.
- This approach may also reduce a substantial amount of manual work in literature screening.

What this adds to what was known?

- NLP technology may be an effective tool to develop models that assist in rapid title and abstract screening when updating systematic reviews.
- The ensemble learning model (ie, LightGBM) provided superior title and abstract screening performance, compared with single bidirectional encoder representations from transformers models.

What is the implication and what should change now?

- Advancements in NLP technology have provided an important opportunity for speeding up the generation of systematic reviews. Our study suggested that such models may effectively assist in rapid literature screening when updating systematic reviews.

1. Introduction

Systematic reviews represent important sources of information that can assist healthcare professionals and policymakers in making decisions [1]. Producing systematic reviews requires identifying evidence that is both high-quality and timely. As systematic reviews always include studies published within a defined date range, it is necessary to periodically update systematic reviews to ensure the timeliness of the research results [2,3].

One significant challenge for updating systematic reviews is the substantial workload of literature screening [4,5]. Thus, to reduce labor and workload and improve efficiency, one potential solution is to use machines speed up the literature screening process. To date, over a hundred tools, which can be grouped into three categories, have been developed to speed up parts of the review process [6]. One category is text visualization tools, such as Covidence [7], Early Review Organizing Software [8] and PICO Portal [9], which usually assist multiple partners working on the same project. The second category is the use of bag-of-words encoding or term frequency and inverse document frequency weightings methods, which estimate and rank the probability targeted literature is relevant (eg, SWIFT-Active Screener [10]). However, this approach ignores word order, thus leading to suboptimal classification. The third category is semi-automate citation screening and selection tools, which use a support vector machine (SVM) to classify documents (eg, GAPscreener [11], Abstrackr [12] and Rayyan [13]). SVM had been proven to be a successful machine learning model and widely employed to classify texts in the first decade of the 21st century [14]. Nevertheless, this approach relies heavily on arbitrarily set sample features and is unstable and labor-intensive.

In light of the complexity of natural language, the progress of natural language processing (NLP) was slow until the introduction of machine learning, which greatly promoted the development of NLP. The development of NLP has gone through three stages: rule-based, probability-based and machine learning-based. In the past decade, machine learning models based on neural networks have emerged, demonstrating strong capacity in NLP [15]. The machine learning model itself consists of two parts: model structure and model parameters. The model structure is designed by the researchers as skeleton of the model, and the model parameters are calculated by the training data. The training process is the process of calculating the parameters of the model. According to the training process, the types of machine learning are divided into three types: supervised learning, unsupervised learning and semi-supervised learning. The supervised learning refers to the process of adjusting the parameters of classifier by using a set of samples with known categories to achieve the required performance [16]. The unsupervised learning refers to adjusting the parameters by using a set of samples without known categories [16]. The semi-supervised learning refers to adjusting the parameters by using a small number of samples with known categories and a lot of samples without known categories [16]. Among these three, the supervised learning has the best accuracy [17].

In the NLP field, supervised learning with a deep neural network model may help to achieve similar or even superior performance when using the model to extract sample features (ie, variable information) while reducing the manual workload. Several deep neural network models have been developed in the NLP field, including convolution neural networks [18], recurrent neural networks [19], Long Short-Term Memory [20], *Bi*-directional Long Short-Term Memory [21], attention mechanisms [22], transformers [23], and bidirectional encoder representations from transformers (BERT) [24]. BERT, which was proposed by Google, is the latest state-of-the-art model created through pretraining and finetuning [24], which enables 11 NLP tasks including textual entailment, semantic similarity, reading comprehension, commonsense reasoning, sentiment analysis, linguistic acceptability, and multi-task benchmark. However, different pretrained sets may lead to different initial BERT parameters, which would impact the model's performance. More recently, light gradient boosting machines (LightGBM) have become a popular machine learning techniques, since a LightGBM may combine the performance of several models [24]. In addition to sav-

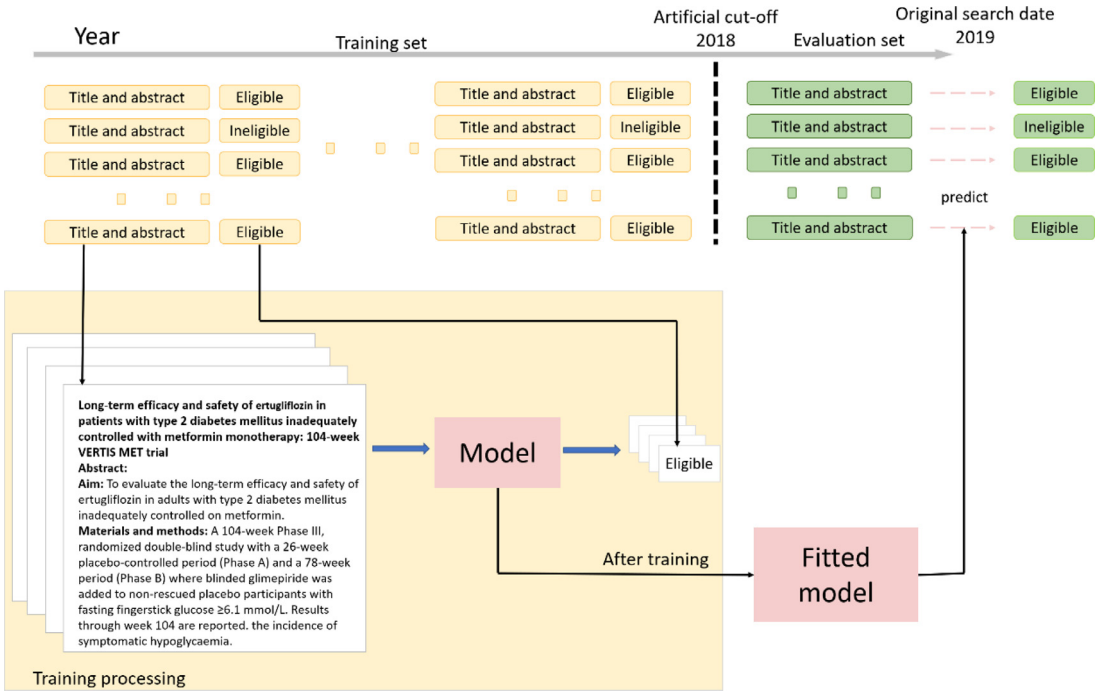

Fig. 1. The conceptual framework for developing and testing the ensemble learning model for automatic title and abstract screening when updating a systematic review.

ing time, this technique has shown far better results and performance compared with existing boosting algorithms [25].

Given the current state of development and strengths of NLP technologies, we have developed and tested an ensemble learning model using a LightGBM that integrates multiple BERT models for automatically screening titles and abstracts in the context of updating a systematic review.

2. Methods

2.1. Study design overview

The overall design of our study is summarized in Fig. 1. To develop and test the ensemble learning model, we used the searched, deduplicated literature from a recently published systematic review [26]. The eligibility of the searched literature was determined by two independent reviewers. We divided the searched literature into two sets – one that had been screened and one to be screened – that were treated as the training set and test set, respectively. Using the training set, we developed an initial model to learn how to distinguish eligible vs. ineligible literature. We then used the test set to examine model performance.

2.2. Data sources

We used the searched literature from a published systematic review of randomized controlled trials on sodium-glucose co-transporter-2 inhibitors for type 2 diabetes mellitus treatment. The literature was retrieved from PubMed, EMBASE, and the Cochrane Central Register of Controlled Trials from inception to June 2019, and included 3,858 unique citations [26]. Paired reviewers, trained in research methods, independently screened titles, abstracts, and full texts for final eligibility. Reviewers addressed disagreements through discussion or, if necessary, consultation with a third reviewer.

By setting up an artificial time cutoff of 2 years, we divided all the searched literature into screened citations and citations to be screened as the training set and test set, respectively, by time of publication. We used articles published before 2018 as the training set, and those published in 2018 and 2019 as the test set. The Cochrane handbook recommends that systematic reviews should be updated every 2 years [27].

2.3. The ensemble learning model

The ensemble learning model integrated four base classifiers and used the classification results of four individual base classifiers as features to train itself to determine the eligibility of citations (Fig. 2). The ensemble learning model was implemented using the LightGBM, and the four base classifiers were implemented using four BERT models.

The four base classifiers held the same base and uncased BERT model structure but different initial parame-

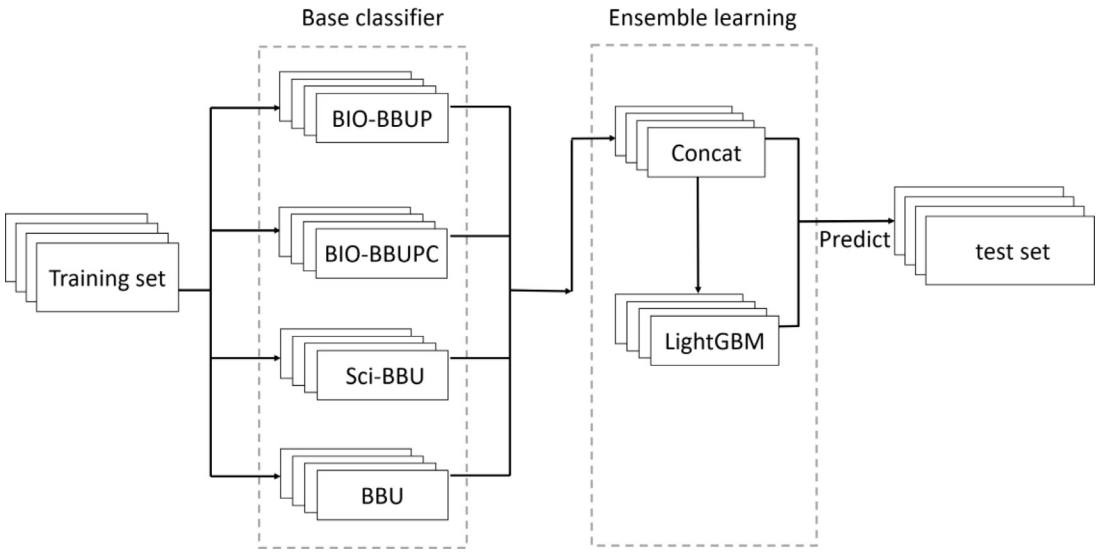

Fig. 2. The framework of the ensemble learning model. The ensemble learning model, LightGBM, contains four individual BERT models as base classifiers. The LightGBM applied classification results of the four individual BERT models on training set to train itself to identify eligible citations. The four BERT models are Sci-BBUP, Sci-BBUPC, BIO-BBU, and BBU. BERT, bidirectional encoder representations from transformers.

ters: BIO-BBUPC, BIO-BBU, Sci-BBU, and BBU. BIO-BBUPC was pretrained on PubMed abstracts and clinical notes in 2018 [28]. BIO-BBU was pretrained on PubMed abstracts in 2018 [28]. Sci-BBU was pretrained on the corpus of semanticscholar.org, which holds 1.14 million papers and 3.1 billion tokens. Sci-BBU used the full text of the papers in training, not just abstracts [29]. BBU was pretrained on Wikipedia in 2018 [24]. For each base classifier, inputs were the citations in the form of titles and abstracts, respectively. For output, each citation obtained a two-dimensional vector as the model result through each base classifier. In the BERT model, we added a classification symbol "[CLS]" symbol at the beginning of the texts. Each word in the text was transformed into three vectors through Token Embeddings, Segment Embeddings and Position Embeddings, where the value of the embedding vector was calculated by train processing rather than defined artificially. The text of each citation was transformed into a series of word embedding vectors through BERT structure to extract feature from text. The output of the highest hidden layer at the position of "[CLS]" was taken as the text-level feature. After a fully connected layer, the final classification result in form of two-dimensional vector was obtained. The BERT structure is the encoding part of the transformer [23], where each block is mainly composed of multiple self-attention, standardization (norm) and residual connection.

The input of the LightGBM (ie, ensemble learning model) was the classification results of the four base classifiers. We combined the four two-dimensional vectors into one eight-dimensional vector as the input for the integrated classifier. The output was a two-dimensional vector. We used argmax to transform each two-dimensional vector to 1-dimensional (1 or 0) as the final classification result. In the LightGBM, the input space was mapped to gradient space, through a function learned from decision tree. For each feature (each dimension of the input), the information gain of all possible split points was estimated through gradient-based one-side sampling [30] and exclusive feature bundling [30].

2.4. Model training process

There were two stages in the model training process. The first stage was trained for four based BERT classifiers, and the second stage was trained for the integrated classifier LightGBM. The LightGBM used the classification result of four base learners to train itself. During training, the classifier learned the link relationship between inputs and labels (eligible or ineligible) on the training set by itself. After training, the classifier was the capable of classifying eligible and ineligible citations without labels.

For the first training stage, to balance eligible and ineligible citations, we randomly sampled the eligible and ineligible citations at 1:1 in the training set. Then, the balanced initial screened citations were randomly divided at 4:1 as the training and development sets. In the training set, the model automatically adjusted its parameters to learn the link relationship between citations and labels (eligible or ineligible). In the development set, we optimized the model's hyperparameters.

For the second training stage, we combined the classification result of four base classifiers on screened citations as the input for training the integrated classifier. By 10-fold training and stepwise hyperparameter adjustment, we fitted the best parameters for the LightGBM.

2.5. Model test

We assessed screening performance on the test set. For the test set, the model prediction classification results of the text were compared with the original classification labels of the text to complete the model test. The performance parameters were sensitivity, specificity, missed studies, and workload saving [31]. Sensitivity is the ratio of the number of correctly labeled eligible citations to the total number of eligible citations. Specificity is the ratio of the number of correctly labeled ineligible citations to the total number of ineligible citations. Missed studies is the ratio of the number of inaccurately labeled eligible citations to the total number of eligible citations. Workload saving was considered the ratio of the number of citations predicted to be ineligible to the total number of citations. If the classifier assessment in the case study was accepted, the reviewer would only need to review the citations that were predicted to be eligible. For missed citations, we checked whether the citations were included in the final full-text classification.

2.6. Model implementation

The BERT models used were pretrained models with different initial parameters. We downloaded the configuration files for BIO-BBUPC, BIO-BBU, Sci-BBU, and BBU from GitHub. The fine-tuning process for the base BERT classifiers was implemented in Python with TensorFlow [32]. The LightGBM was implemented in lightgbm [33] and Scikit-learn [34].

2.7. Statistical analysis

To ensure the statistical accuracy and wide applicability, we used Fisher's exact test to assess the statistical significance of the differences in sensitivity and specificity between the LightGBM and each base classifier [31].

To test the model's robustness, we applied two strategies to rebuild the model data set. First, we conducted sensitivity analyses by using earlier cutoff times to separate initially screened citations and citations to screen for an update. In the sensitivity analyses, we performed a 3-year update in which citations published in 2017, 2018, and 2019 and before 2017 were the test and training sets, respectively, and a 4-year update, in which citations published in 2016, 2017, 2018, and 2019 and before 2016 were the test and training sets, respectively. Second, we randomly generated a training set, development set, and test set, which were the same size in a 2-year update.

3. Results

Table 1 presents the results of the four single models and the LightGBM for the 2-year update. In the 2-year update, the test set included 974 citations. Under the manual classification criteria, the test set was divided into 180 eligible citations that met the inclusion criteria and 767 ineligible citations that did not.

With our model classification, among the 947 citations in the test data, 340 citations were eligible and 607 citations were ineligible. Our model achieved a 96% sensitivity, and 78% specificity. If the classifier assessment in the case study was accepted, reviewers decrease their workload by 64.1% while losing only 8 of 180 eligible citations (4%), none of which were ultimately included in the systematic review after full-text consideration. The precision of our model in the test set is 49.6%, which is better than that of SWIFT-Active Screener (average 13.4%, 0.6%–41.6%) [10]. Due to the fact that no precision was directly reported in the SWIFT-Active Screener study [10], we calculated the average precision according to four indicators of the 26 projects in the paper [10], namely "Records from Search," "Included," "Obtained true recall' and 'Obtained wss@95." WSS@95 [10] is Work Saved over random Sampling at 95% recall.

For the base classifiers, Sci-BBUP included 248 citations, excluded 699 citations, and achieved a 79% sensitivity, and 86% specificity. If the classifier assessment in the case study was accepted, reviewers would decrease their workload by 73.8% and lose 38 of 180 eligible citations (21%), one of which was included in the systematic review after full-text consideration.

Sci-BBUPC included 278 citations, excluded 669 citations, and achieved 82% sensitivity, and 83% specificity. If the classifier assessment in the case study was accepted, reviewers would decrease their workload by 70.6% and lose 33 of 180 eligible citations (18%), two of which were included in the systematic review after full-text consideration.

Sci-BBU included 293 citations, excluded 654 citations, and achieved 85% 92% sensitivity, and 83% specificity. If the classifier assessment in the case study was accepted, reviewers would decrease their workload by 69.1% and lose 15 of 180 eligible citations (8%), none of which were ultimately included in the systematic review after full-text consideration.

BBU included 264 citations, excluded 683 citations, and achieved 82% sensitivity, and 85% specificity. If the classifier assessment in the case study was accepted, reviewers would decrease their workload by 72.1% and lose 33 of 180 eligible citations (18%), none of which were ultimately included in the systematic review after full-text consideration.

The LightGBM results were significantly superior to three of the four single BERT models in terms of sensitivity ($P < 0.001$), and seemed to be slightly inferior to the single BERT models in terms of specificity ($P < 0.001$), although the differences for three of the four single BERT models were not statistically significant. The results are shown in Table 1. In general, our model was found to reduce overall workload while identifying 96% of all eligible

Table 1. The comparison results of different NLP methods when updating 2 year after the initial conduct of the RCTs of SGLT2 inhibitors for treatment of T2DM

Model name	Number of citations to screen in 2-year update						accuracy	Specificity	Sensitivity	Full miss
	Total	Eligible								
		Manual	Correctly predicted	Missed	Total predicted positive	Ineligible spared to screen				
BIO-BBUP			142	38	248	699	0.85	0.86	0.79[a]	1
BIO-BBUPC			147	33	278	669	0.83	0.83	0.82[a]	2
SCI-BBU	947	180	165	15	293	654	0.85	0.83	0.92	0
BBU			147	33	264	683	0.84	0.85	0.82[a]	0
Our model (LightGBM)			172	8	347	600	0.81	0.78	0.96	0

[a] indicated that the LightGBM and Vote were significantly superior to this model in term of sensitivity ($P < 0.001$).NLP, natural language processing; RCTs, randomized controlled trials; SGLTS. sodium-glucose co-transporter-2; T2DM, type 2 diabetes mellitus.

Table 2. The comparison results of different NLP methods when updating 3 year after the initial conduct of the RCTs of SGLT2 inhibitors for treatment of T2DM

Model name	Number of citations to screen in 3-year update						accuracy	Specificity	Sensitivity	Full miss
	Total	Eligible								
		Manual	Correctly predicted	Missed	Total predicted positive	Ineligible spared to screen				
BIO-BBUP			221	49	435	1255	0.84	0.85	0.82[a]	2
BIO-BBUPC			165	105	369	1321	0.82	0.86	0.61[a]	5
SCI-BBU	1690	270	241	29	538	1152	0.81	0.79	0.89	1
BBU			195	75	433	1257	0.81	0.83	0.72[a]	3
Our model (LightGBM)			252	18	686	1004	0.73	0.69	**0.93**	1

NLP, natural language processing; RCTs, randomized controlled trials; SGLTS. sodium-glucose co-transporter-2; T2DM, type 2 diabetes mellitus.

[a] indicated that the LightGBM and Vote were significantly superior to this model in term of sensitivity ($P < 0.001$).

citations in the case study, which outperformed any of the single BERT models.

The results of the sensitivity analyses are shown in Tables 2–4. The model showed high and stable sensitivity and specificity in sensitivity analysis. It achieved 93% sensitivity, 69% specificity and decreased the workload by 59.4%, while missing 18 of 270 (6.7%), one of which was included in the case study after full-text consideration (Table 2) in 3-year update. It achieved 95% sensitivity, 69% specificity and decreased the workload by 59.4%, while missing 19 of 357 (5.3%), one of which was included in the case study after full-text consideration (Table 3) in 4-year update. With the same data size, the model performed with similar sensitivity and specificity in randomly generated test data sets, compared with the 2-year update; 97% sensitivity was achieved. Furthermore, it decreased the workload by 57.4%, while missing 6 of 180 (3.3%), none of which were ultimately included in the case study after full-text consideration. The details are shown in Table 4.

4. Discussion

4.1. Main findings and implications

We developed and evaluated an advanced NLP model, an ensemble learning model LightGBM integrating four different BERT models, for automatically screening citations when updating systematic reviews 2 years after the initial systematic review screening. Our results showed that the model achieved good and stable discriminative properties for this task. Our model missed no citations which were ultimately included after full-text consideration in the systematic review used as our case study. Contrastingly, the single BERT models, Sci-BBUP and Sci-BBUPC, missed two and one eligible citations, respectively, which were ultimately included in the systematic review after full-text consideration. Advancements in NLP technology offer an important opportunity to speed up the generation of systematic reviews, and our study suggests that such models may effectively assist in rapid literature screening when updating systematic reviews.

Table 3. The comparison results of different NLP methods when updating 4 year after the initial conduct of the RCTs of SGLT2 inhibitors for treatment of T2DM

	Number of citations to screen in 4-year update									
		Eligible								
Model name	Total	Manual	Correctly predicted	Missed	Total predicted positive	Ineligible spared to screen	accuracy	Specificity	Sensitivity	Full miss
BIO-BBUP			257	100	619	1657	0.80	0.85	0.72[a]	5
BIO-BBUPC			215	142	573	1703	0.78	0.86	0.60[a]	4
SCI-BBU	2276	357	290	67	641	1635	0.82	0.79	0.81[a]	1
BBU			317	40	885	1391	0.73	0.83	0.89	2
Our model (LightGBM)			338	19	924	1352	0.73	0.69	**0.95**	1

NLP, natural language processing; RCTs, randomized controlled trials; SGLTS. sodium-glucose co-transporter-2; T2DM, type 2 diabetes mellitus.

[a] indicated that the LightGBM and Vote were significantly superior to this model in term of sensitivity ($P < 0.001$).

Table 4. The comparison results of different NLP methods where test set were randomly generated in the same size with 2-year updates

	Number of citations in test set									
		Eligible								
Model name	Total	Manual	Correctly predicted	Missed	Total predicted positive	Ineligible spared to screen	Accuracy	Specificity	Sensitivity	Full miss
BIO-BBUP			166	14	322	625	0.82	0.80	0.92	1
BIO-BBUPC			166	14	320	627	0.82	0.80	0.92	2
SCI-BBU	947	180	171	9	414	533	0.73	0.68	0.95	1
BBU			161	19	317	630	0.82	0.80	0.89	1
Our model (LightGBM)			174	6	405	542	0.75	0.70	0.97	0

NLP, natural language processing.

4.2. Strengths

One strength of our study was that the LightGBM and BERT models we employed are novel machine learning methods popularized in recent years due to their high performance in NLP tasks. We illustrated that novel machine learning methods in NLP could achieve high sensitivity with good specificity, compared with rapid title and abstract screening, when updating systematic reviews. Compared with previous work, automatic screening using word embeddings achieved high sensitivity, but low specificity (33%), in updating living network meta-analyses [31]. Thus, the ensemble learning method consisting of advanced NLP models could indeed help reviewers screen citations with high sensitivity and good specificity when updating systematic reviews.

The second strength was that we showed the ensemble learning method consisting of advanced NLP models could maintain stable high-sensitivity results, compared with the single NLP model. Compared with the sensitivity analysis results, the best and worst sensitivity base classifiers of the four base BERT classifiers were different. However, the ensemble learning method maintained the highest sensitivity (more than 93%) and good specificity (approximately 70%).

The third strength is that we proposed all the classification results of six models, including four base classifiers and two fusion classifiers. In relation to sensitivity, we showed that an advanced machine learning fusion model could be a better method for screening citations than a single model in an actual project. However, users can still choose their own best model based on the performance of models on the evaluation set.

4.3. Limitations

Our study had a few limitations. First, there were insufficient training samples. Each systematic review project holds thousands of citations, which is not a huge number for machine learning. When we used an earlier cutoff time, the performance of each model was lower due to the training sample reduction. Second, our study was only a case study, and may not be generalizable. Nevertheless, we believe that the method is novel and effective in improving citation screening when updating systematic reviews. Hopefully, this would also generate more discussion and

interest in the use of NLP methods for facilitating the update of systematic reviews. Meanwhile, when reporting a systematic review, the information as to what stage citations are exclusion is often unclear; this fact leads to that the data are hardly available for NLP model training. We will, in our next study, make efforts by collecting more examples to further validate our method. Third, this study only focused on updating systematic reviews. Because our model parameters were fitted based on text and known text classification information. Last but not least, our model currently deals only with English literature and may not apply to other languages.

4.4. Future work

In this case study, we implemented advanced NLP models for rapid citation screening when updating systematic reviews. In the future, our model could be employed for a more updated systematic review to further explore the model's stability and generalizability. When conducting a systematic review, researchers define the eligibility criteria for literature screening according to PICOS (population or patient, intervention, control, outcome, study design) principle [35]. Thus, the clinical problem is expressed in a standardized way. However, we did not consider the PICOS principle in screening progression, although manual citation classification is based on the PICOS principle. Future studies using identification and standardization of the PICOS principle for rapid citation screening in systematic reviews could provide a highly generalizable model.

5. Conclusion

Our study showed that the LightGBM ensemble learning method consisting of advanced NLP models could indeed help systematic reviewers screen citations when updating systematic reviews. Thus, at the present stage, our study suggests that it is reasonable and feasible that the model replaces one human screeners in updating systematic reviews.

References

[1] Gupta S, Rajiah P, Middlebrooks EH, Baruah D, Carter BW, Burton KR, et al. Systematic review of the literature: best practices. Acad Radiol 2018;25(11):1481–90.
[2] Sampson M, Shojania KG, Garritty C, Horsley T, Ocampo M, Moher D. Systematic reviews can be produced and published faster. J Clin Epidemiol 2008;61(6):531–6.
[3] Oral iron-based interventions for prevention of critical outcomes in pregnancy and postnatal care: an overview and update of systematic reviews. J Evid Based Med 2019;12(2):155–66.
[4] NCBI. PubMed. Secondary PubMed 1988. Available at: https://pubmed.ncbi.nlm.nih.gov/. (Accessed July 27, 2020).
[5] McKibbon KA, Wilczynski NL, Haynes RB. Retrieving randomized controlled trials from Medline: a comparison of 38 published search filters. Health Info Libr J 2009;26(3):187–202.
[6] Michelson M, Ross M, Minton S. Ai2 leveraging machine-assistance to replicate a systematic review. Value Health 2019;22(Supplement 2):S34.
[7] Adams CE, Polzmacher S, Wolff A. Systematic reviews: work that needs to be done and not to be done. J Evid Based Med 2013;6(4):232–5.
[8] Glujovsky D, Bardach A, García Martí S, Comandé D, Ciapponi A. PRM2 EROS: a new software for early stage of systematic REVIEWS. Value Health 2011;14(7):A564.
[9] Trøseid MM. PICO portal. Available at: https://picoportalnet/.(accessed July 27, 2020).
[10] Howard BE, Phillips J, Tandon A, Maharana A, Elmore R, Mav D, et al. SWIFT-active screener: accelerated document screening through active learning and integrated recall estimation. Environ Int 2020;138:105623.
[11] Yu W, Clyne M, Dolan SM, Yesupriya A, Wulf A, Liu T, et al. GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. BMC Bioinformatics 2008;9 205-05.
[12] Gates A, Johnson C, Hartling L. Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool. Syst Rev 2018;7(1):45.
[13] Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210.
[14] Romero R, Iglesias EL, Borrajo L. A linear-RBF multikernel SVM to classify big text corpora. Biomed Res Int 2015;2015:878291.
[15] Tsuruoka Y. Deep learning and natural language processing. Brain Nerve 2019;71(1):45–55.
[16] Russell S, Norvig P. Artificial intelligence: a modern approach. Prentice Hall; 2002.
[17] Jennings NR, Wooldridge MJ. Foundations of machine learning. MIT Press; 2012.
[18] Rios A, Kavuluru R. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In: Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics. Atlanta, Georgia: Association for Computing Machinery; 2015. p. 258–67.
[19] Poon HK, Yap WS, Tee YK, et al. Hierarchical gated recurrent neural network with adversarial and virtual adversarial training on text classification. Neural Network 2019;119:299–312. doi:10.1016/j.neunet.2019.08.017.
[20] Tang D, Qin B, Feng X, Liu T. Target-dependent sentiment classification with long short term memory; 2015. arXiv preprint arXiv:151201100.
[21] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's transformers: state-of-the-art natural language processing. ArXiv 2019:arXiv: 1910.03771.
[22] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. In: Advances in neural information processing systems; 2017. p. 5998–6008.
[23] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's transformers: state-of-the-art natural language processing. arXiv preprint arXiv:191003771 2019.
[24] Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805 2018.
[25] Anghel A, Papandreou N, Parnell T, De Palma A, Pozidis H. Benchmarking and optimization of gradient boosting decision tree algorithms. arXiv preprint arXiv:180904559 2018.
[26] Liu J, Li L, Li S, Wang Y, Qin X, Deng K, et al. Sodium-glucose co-transporter-2 inhibitors and the risk of diabetic ketoacidosis in patients with type 2 diabetes: a systematic review and meta-analysis of randomized controlled trials. Diabetes Obes Metab 2020;22(9):1619–27.
[27] Tarsilla M. Cochrane handbook for systematic reviews of interventions. J Multidiscip Eval 2008;6:142–8.

[28] Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. ArXiv 2019;abs/1906.05474.
[29] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40.
[30] Meng Q. LightGBM: a highly efficient gradient boosting decision tree. In: Advances in neural information processing systems; 2018. p. 3149–57.
[31] Lerner I, Créquit P, Ravaud P, Atal I. Automatic screening using word embeddings achieved high sensitivity and workload reduction for updating living network meta-analyses. J Clin Epidemiol 2019;108:86–94.
[32] TensorFlow: large-scale machine learning on heterogeneous distributed systems. arXiv:160304467 2016.
[33] Ke G, Meng Q, Finely T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Advances in neural information processing systems; 2017. p. 3149–57.
[34] Swami A, Jain R. Scikit-learn: machine learning in python. J Machine Learn Res 2013;12(10):2825–30.
[35] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol 2009;62(10):1006–12.