Journal of Biomedical Informatics 112 (2020) 103578

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Using the contextual language model BERT for multi-criteria classification of scientific articles

Ashwin Karthik Ambalavanan, Murthy V. Devarakonda *

Arizona State University, United States

ARTICLE INFO

Keywords:
Biomedical natural language processing
Neural networks
Screening scientific articles
Text classification
Machine learning
BERT
SciBERT

ABSTRACT

Background: Finding specific scientific articles in a large collection is an important natural language processing challenge in the biomedical domain. Systematic reviews and interactive article search are the type of downstream applications that benefit from addressing this problem. The task often involves screening articles for a combination of selection criteria. While machine learning was previously used for this purpose, it is not known if different criteria should be modeled together or separately in an ensemble model. The performance impact of the modern contextual language models on the task is also not known.

Methods: We framed the problem as text classification and conducted experiments to compare ensemble architectures, where the selection criteria were mapped to the components of the ensemble. We proposed a novel cascade ensemble analogous to the step-wise screening process employed in developing the gold standard. We compared performance of the ensembles with a single integrated model, which we refer to as the individual task learner (ITL). We used SciBERT, a variant of BERT pre-trained on scientific articles, and conducted experiments using a manually annotated dataset of ~49 K MEDLINE abstracts, known as Clinical Hedges.

Results: The cascade ensemble had significantly higher precision (0.663 vs. 0.388 vs. 0.478 vs. 0.320) and F measure (0.753 vs. 0.553 vs. 0.628 vs. 0.477) than ITL and ensembles using Boolean logic and a feed-forward network. However, ITL had significantly higher recall than the other classifiers (0.965 vs. 0.872 vs. 0.917 vs. 0.944). In fixed high recall studies, ITL achieved 0.509 precision @ 0.970 recall and 0.381 precision @ 0.985 recall on a subset that was studied earlier, and 0.295 precision @ 0.985 recall on the full dataset, all of which were improvements over the previous studies.

Conclusion: Pre-trained neural contextual language models (e.g. SciBERT) performed well for screening scientific articles. Performance at high fixed recall makes the single integrated model (ITL) more suitable among the architectures considered here, for systematic reviews. However, high F measure of the cascade ensemble makes it a better approach for interactive search applications. The effectiveness of the cascade ensemble architecture suggests broader applicability beyond this task and the dataset, and the approach is analogous to query optimization in Information Retrieval and query optimization in databases.

1. Introduction

MEDLINE and EMBASE are large repositories of scientific articles that are often screened for systematic reviews and other meta-analysis. MEDLINE has about 29 million articles at this time and a million new articles are being added every year. However, very small percentage of the articles meet a specific criteria such as *scientifically sound articles* (see Table 1) – the percentage of such articles was about 1% in a carefully curated dataset [1]. Finding such a small number of studies from massive collections is a significant challenge in effectively unlocking existing knowledge in the repositories.

Systematic reviews critically depend on screening articles from the bibliographic repositories. The standard approach is to first use a bibliographic search engine such as PubMed [2] and then further screen the articles manually. PubMed also offers specialized search using *Clinical Query filters* – which are rules based on combinations of text strings, MeSH (Medical Sub Headings), and database tags that were optimized by expert librarians – to reduce the manual screening and improve the search for relevant clinical studies for systematic reviews [3–5].

* Corresponding author.
E-mail address: mvd@acm.org (M.V. Devarakonda).

https://doi.org/10.1016/j.jbi.2020.103578
Received 13 April 2020; Received in revised form 20 September 2020; Accepted 22 September 2020
Available online 13 October 2020

Table 1
Four criteria for acceptance.

Criterion Name	Description
Format	Whether it is an original study, review, case report, general and miscellaneous article.
HHC	Whether the article is of interest to human health care
Purpose	Whether the article is about etiology, prognosis, diagnosis, treatment (or prevention), costs, economics, disease related prediction, qualitative study, or something else
Rigor	Whether the study meets experiment design quality criteria specific to a purpose, e.g. random allocation to comparison groups if purpose is treatment.

Machine learning has been applied to the task. Early studies [6–8] relied on extensive feature engineering, including proprietary and time-dependent features. Two recent studies [9,10] showed that early neural networks models (CNN) can improve precision compared to the manually created search filters. However, recently developed contextual neural language models, BERT (Bidirectional Encoder Representations from Transformers) [11] or its variants [12–16], have not been studied for this task. These new neural models have established new state-of-the-art results for several general domain tasks [11] and for biomedical information extraction tasks [17,18].

Furthermore, since the logic for selecting articles often involves distinctly separate criteria (see Table 1), an interesting research question is: how to model the criteria? Is a single model or an ensemble of criterion-specific models the most optimal? While ensembles have been frequently studied in natural language processing (NLP), they were often used to aggregate predictions of different models, each implementing the same task with a different technique. Here we use ensembles to aggregate outcomes of different criteria of a given task. Our results would generalize to modeling multiple criteria or constraints of any biomedical NLP task.

Information retrieval (IR) techniques have also been used for searching articles in large repositories [19–22]. Recently, even neural networks have been used for information retrieval [23–26]. However, we mapped the task to text classification rather than to IR, as in the previous studies mentioned above, since ranking the results is not the goal in systematic reviews (a key application). Our approach can also be used in applications requiring ranking by subsequently ranking or re-ranking based on the outputs generated by our models.

We used the dataset from the Clinical Hedges project which pioneered work in this area by extracting a dataset of about 49,000 articles published in 2000 in 170 clinical journals, 161 of which were indexed in MEDLINE [27]. The articles were manually annotated on four different criteria shown in Table 1. The criteria when combined identify *scientifically sound articles* of a selected type – for example, scientifically sound original articles describing diagnostic methods (rather than treatment methods) in human healthcare. The dataset has a unique value proposition for our study in that it has been separately annotated on the four constituent criteria that make up the overall selection.

In a nutshell, the research question we explored in this study, using the Clinical Hedges dataset, was: What is the best BERT-based ensemble for classification of articles where the classification is based on a combination of criteria? The null hypothesis we tested therefore was: there is no performance difference among the ensembles on the dataset. The standard precision (P), recall (R), and F measure metrics were used for comparison. Only the combination of title and the abstract (not the full text) of each article was used. We note that while the systematic reviews process is focused on precision at very high recall (~99%), the highest achievable F measure characterizes models' performance more broadly in terms of their ability to balance the noise and the retrieval rate. Therefore, the ensembles and a single integrated model were first compared among themselves using all the performance metrics, and then the model most suitable for high-recall was used to compare performance with previous studies.

2. Methods

2.1. Dataset

The Clinical Hedges dataset consists of 49,028 unique articles retrieved from MEDLINE in 2000 (and 50,590 annotations) [27]. Each article was identified by its MEDLINE id, called PMID, title and abstract, as well as the other metadata from MEDLINE. Each article was manually rated on four criteria: (1) The "Format" of the article, which can be original study, review, case report, or general and miscellaneous studies; (2) Whether it is concerning human healthcare or not, abbreviated as "HHC"; (3) Intended "Purpose" of the paper such as treatment, diagnosis, prognosis, or economics; (4) Whether the study has scientific "Rigor" in that if it met certain study design constraints that are specific to the purpose, i.e. if the purpose is treatment then the study must allocate subjects randomly to study groups.

The annotators were instructed to rate each article on the sequence of criteria starting from criterion 1 to criterion 4. After some criteria, they were asked to stop rating an article further based on the outcome. For example, when the Format of an article was determined as "general and miscellaneous" type, raters were asked not to rate it further.

Because the articles were rated on multiple criteria, the number of positive and negative articles depended on the specific outcomes wanted from the criteria. If the task was to identify human healthcare related original studies of treatment conducted with rigor, then the positive class contained 1587 articles. The negative samples were, 49,003. Other selection criteria such as diagnosis or prognosis as the purpose was possible and would result in a different number of positive and negative samples.

Publication type is a metadata tag that MEDLINE indexers regularly assign to articles in MEDLINE. Although these tags could potentially give an insight into the relevance of the article, not all articles were assigned this tag. Also, it may take up to 6 months after article publication for MEDLINE indexers to add the publication type [10] and our analysis showed that useful tags were present for about 2000 articles of the Clinical Hedges dataset. Therefore, we could use it only as a noisy input feature (i.e. cannot always be relied on) for an article.

2.2. Models

We used uncased SciBERT, which is an uncased BERT model pre-trained on a corpus of scientific articles, as the core model in our study. The model was pre-trained on a random sample of 1.14 M papers from Semantic Scholar (semanticscholar.org). The pre-training corpus, therefore, closely resembled MEDLINE articles, and consisted of full text of papers, 18% from the computer science domain and 82% from the broad biomedical domain. The resulting corpus had 3.17B tokens, about the same as the 3.3B BERT pre-training tokens, but had only 42% words in common. Since there were different criteria for classifying an article as a scientifically sound study, we conceived four different ways the task can be modeled.

2.3. Individual Task Learner (ITL)

This is the basic form of the neural network model where a single pre-trained SciBERT model was used with a feed forward network (FFN) as the text classification head (see Fig. 1). The logits from the CLS tag of SciBERT are input to the FFN. The integrated model predicts whether the article input to it meets all the criteria for the positive class or not. Thus, a single model learns the integrated criteria for classifying an article.

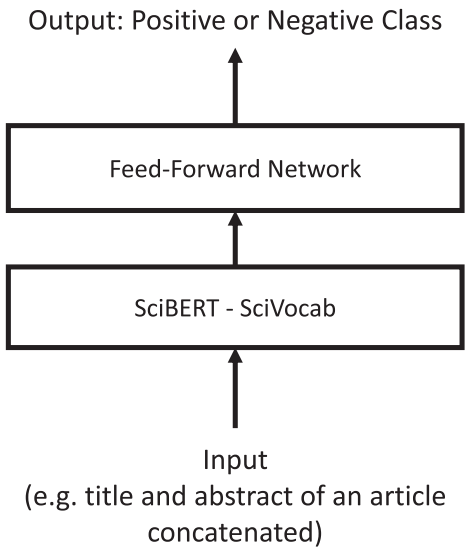

Fig. 1. Individual Task Learner (ITL): An integrated single model of all the criteria.

2.4. Cascade Learner

In this architecture, we train four different SciBERT models each having the same architecture as the ITL described above but each learning only one of the four criteria (see Fig. 2). The first model (Task-1) is trained on classifying an article based on the Format attribute alone. Similarly, the second (Task-2) was trained to classify on HHC, the third (Task-3) on Purpose, and the fourth (Task-4) on Rigor. The (sub) models are ensembled as a series of cascading blocks like a water fall. A model only sees the articles that were assigned a positive label by the previous model in the cascade. The articles that were classified as negative by the previous model are not further analyzed, and were assigned a negative label. This was done during the training as well as in the test. So, downstream models are trained and tested on only a subset of the data. Each component of the cascade ensemble was trained separately on inputs and labels corresponding to it. The intuition behind this approach is twofold. First, this is how the manual annotators labeled the data. Second, each model learns to predict a particular criterion well under the narrow conditions of its filtered input.

2.5. Boolean ensemble learner (ensemble-Boolean)

Ensembling multiple models is a well-known technique where multiple models independently analyze each input and the final prediction is made by combining the output of all models. In fact, the Cascading Learner is a novel variation of ensembled models. In this architecture, each of the 4 models were trained and tested on the entire dataset (considered in an experiment) rather than on a subset. We combine the outputs using Boolean logic (i.e. a conjunction of all outputs) according to the criteria for selecting scientifically sound articles (see Fig. 3). So, the final output label is False if any of the models classifies it as False.

2.6. Feed-forward network ensemble learner (ensemble-FFN)

The intuition for this architecture is to allow the ensemble to learn how to combine individual predictions with the help of a feed forward network (FFN) rather than simply use a Boolean conjunction. We trained four SciBERT models, one for each of the 4-tasks, as in Ensemble-Boolean. The CLS logits from each of these sub-models were concatenated together (forming a 4×768 vector) as the input to the FFN, which decides whether to give a positive (include) or negative (exclude) label to the input article (see Fig. 4). The entire model was trained together backpropagating error through the FFN and all the individual models.

3. Experiments

The goal of our experiments is to study how different ensembles and ITL perform on the task of classifying scientific articles based on the four filtering criteria using the Clinical Hedges dataset. Unlike the previous studies using the dataset, we used the gold standard labels from the dataset for both training and testing the models. Since the number of positive samples is small, we used 10-fold cross validation. As is conventional, we randomly split the positive and negative samples into 10 folds, tested the model on the i^{th} fold after training the ensemble on the remaining 9 folds, and varied i from 1 to 10. It should be noted that this process eventually tests the models on *all* the samples of the dataset. The test results were averaged across all the folds to calculate the model precision, recall, and F measure. The input to a model is a concatenation of the title and abstract of an article, in that order. As described later in this section, we experimented with varying maximum sequence length, different sampling ratios for training the models, and the effect of the Publication Type tag [28] in MEDLINE.

3.1. Dataset

The dataset we received from the Clinical Hedges project had 50,594 entries, but four of them could not be found in MEDLINE and hence we removed them, resulting in 50,590. For comparison purposes, we

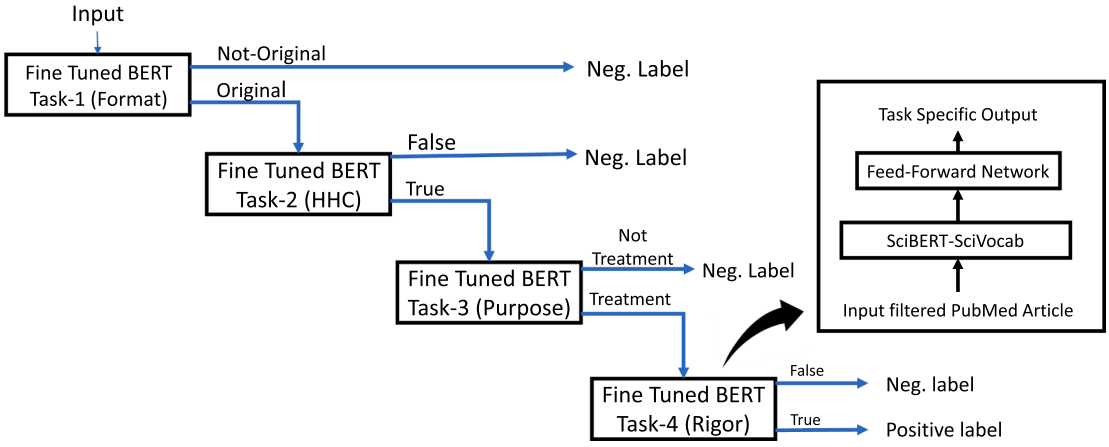

Fig. 2. Cascade Learner: A cascade ensemble of models, one for each criterion. Note that the inputs to Task i are only the positive predictions of Task i-1, i = 2, 3, or 4.

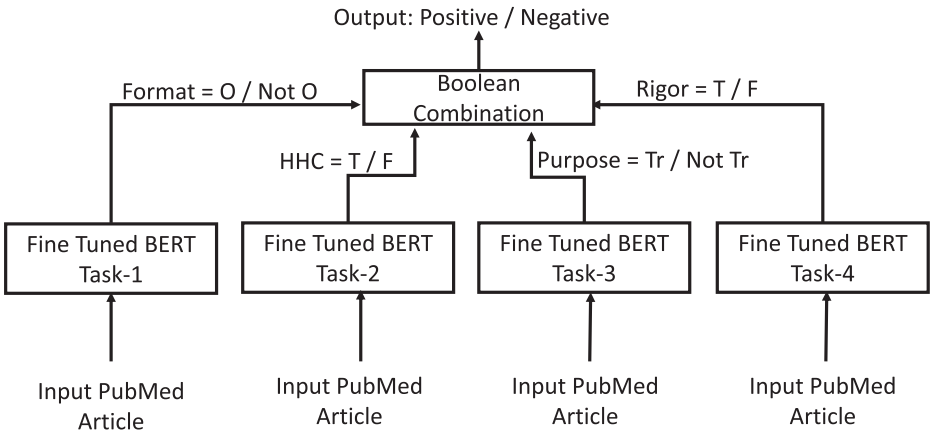

Fig. 3. Ensemble of criterion-specific models, the predictions are combined using Boolean AND operation. All models process all input samples.

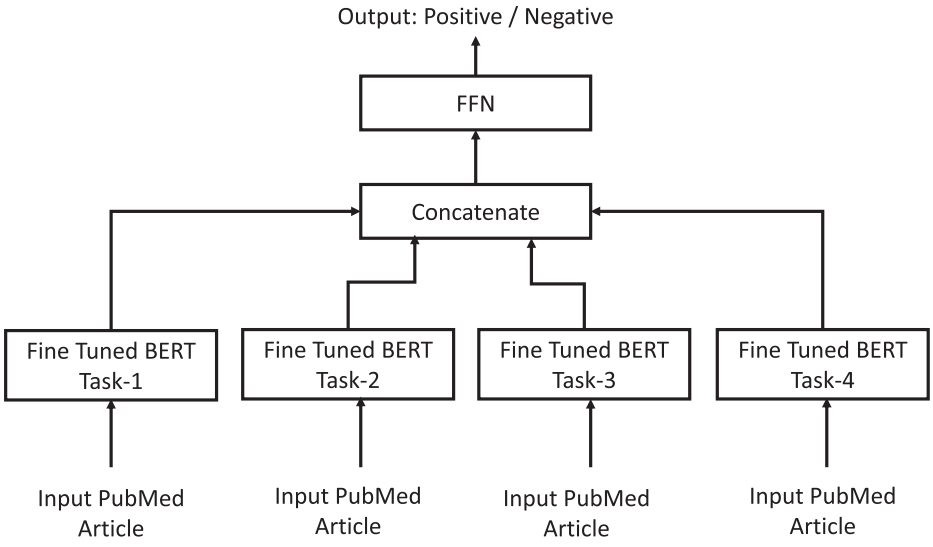

Fig. 4. Ensemble of criterion-specific models, the CLS logits are concatenated and input to a Feed Forward Network. All models process all input samples.

experimented with two subsets of the dataset that were used in the most recent previous studies - Del Fiol et al [9] and Marshall et al. [10] Marshall et al used all the samples (full dataset) from the dataset for testing, but Del Fiol et al used a subset that met the following constraint:

$$\text{Format} = \text{Original/Review/Blank}, \text{HHC} = \text{True/False}, \text{Purpose} = \text{Any}, \text{Rigor} = \text{True/False}$$

We applied this filter to create the (Del Fiol et al) subset. The result was a subset that had 28,331 articles whereas the full set had all 50,590 articles (see Table 2). Both studies used the following identical definition for positive samples:

$$\text{Format} = \text{Original}, \text{HHC} = \text{True}, \text{Purpose} = \text{Treatment and Rigor} = \text{True}$$

We used the same filter for creating the positive samples. In addition, Del Fiol et al removed articles with empty abstracts (but Marshall et al included them) resulting in a smaller set of positive samples. It should be noted that the actual numbers may be slightly different from the reports because the Clinical Hedges dataset may have changed slightly over time and also there may be additional filtering that was not documented in the previous reports. The Clinical Hedges data has been made available by the project team at McMaster, and the team should be contacted directly for accessing it.

Furthermore, the two studies used *noisy* (but larger) training datasets extracted using PubMed search queries. For the exact queries we refer the reader to the publications for details. Broadly both used articles returned by PubMed using the Clinical Queries filter tuned for precision ("Narrow" filter) in a particular time period were used as positive samples.

Table 2
Baseline studies.

Data subsets used	Positive Samples	Negative Samples	Total Samples
Subset using filters in Del Fiol et al	1553	26,778	28,331
Full set as in Marshall et al	1587	49,003	50,590*

* There are a total of 50,594 articles but 4 of them didn't have any entries in PubMed and hence were removed.

3.2. Experiments with ensembling techniques

We conducted the experiments with both subsets of the Clinical Hedges dataset. We present results for the full set only and compare P, R, and F measure obtained by the models. In all cases, we present results only for the positive class, which is the class of interest. We also studied performance on individual criteria for select ensembles to quantify which criterion is the most challenging to model accurately.

3.3. Sampling ratio experiments

The ratio of positive and negative samples is almost 1:32 in the dataset. So, we conducted experiments to study the impact of training with various sampling sizes and ratios. We randomly down-sampled negative samples and up-sampled positive samples by duplication. It should be noted that as the number of samples increased, computing resources needed for training and testing the ensembles increased substantially (multiple days of running time and GPU out of memory problems) limiting the space of alternatives explored. Furthermore, because of its architecture, changing sampling ratios in Cascade Learner makes comparison with other models and interpretation difficult and so we always used balanced 1:1 positive and negative sample ratios at all components of the ensemble in training the Cascade Learner.

3.4. Sequence length experiments

The average input sequence length in the full dataset was 178.82 words, and the maximum length was 856 words (see Table 3). The 95th, 92nd, and 69th percentiles were 420, 384, and 256 respectively. While SciBERT standard (base) model allows sequence lengths up to 512, increasing sequence length also increases the computational (GPU memory and running time) requirements. So, we used the default sequence length of 256 in the initial configuration and studied the performance impact of increasing the sequence length to 384.

3.5. Publication Type (PT) tag experiments

Some MEDLINE articles are manually assigned one or more publication type tags [28] sometime after (typically several months) the article was added to the repository. We used the publication tags as a part of the input text, i.e. we prepended the tags to title and abstract text.

3.6. Metrics

We used standard precision (P), recall (R), and the F measure for performance assessment. The F measure was used for assessing the performance characteristics of the neural architectures, and based on the results, precision at high recall for the most suitable model was used for comparison with related studies. Ten-fold cross validation provided a convenient way to calculate confidence intervals for the metrics. The arithmetic-mean of 10 folds (i.e. macro average) was reported and the (sample) standard deviation was used, along with the Student's t distribution score for 9 degrees of freedom, to calculate 95% confidence intervals. We showed confidence intervals where specific values mattered but not when we were observing general trends.

4. Results

4.1. Performance comparison of the ensembles and ITL

The first set of results (see Table 4) show performance comparison of the ensemble architectures and ITL. The Table presents precision, recall, and F measure for the four architectures. Here we used sequence length of 256 words, randomly down-sampled negative samples to the number of positive samples, and PT tag was not used. The Cascade Learner had significantly higher precision (0.663 vs. 0.388 vs. 0.478 vs. 0.320) and F measure (0.753 vs. 0.553 vs. 0.628 vs. 0.477) than ITL, Ensemble-Boolean and Ensemble-FFN. However, ITL had significantly higher recall than the other classifiers (0.965 vs. 0.872 vs. 0.917 vs. 0.944).

Table 3
Articles length.

Statistic	Sequence Length (in words)
Average Length	178.82
69th Percentile	256
92nd Percentile	384
95th Percentile	420
Maximum Length	856

Table 4
Base configuration results (Marshall dataset)- Statistical Significance.

Model	Precision	Recall	F Measure
ITL	0.388 (±0.039)	**0.965** (±0.010)	0.553 (±0.034)
Cascade Learner	**0.663** (±0.025)	0.872 (±0.019)	**0.753** (±0.016)
Ensemble-Boolean	0.478 (±0.021)	0.917 (±0.013)	0.628 (±0.016)
Ensemble-FFN	0.320 (±0.028)	0.944 (±0.014)	0.477 (±0.029)

PT = no, 256 sequence length, 1.5 K:1.5 K sampling, no PT tag.

4.2. Performance on different criteria

Next, we address the question of how well criterion-specific individual models in the ensembles performed. Table 5 shows precision, recall, and F measure values (using the full dataset) for each of the four criteria, for the Cascade Learner and the Ensemble-Boolean, which were the top two performing ensembles. Note that the inputs to the ensemble components for Cascade Learner were filtered at each stage while all the inputs were presented to the components of the Ensemble-Boolean. Performance generally reduces from Task 1 through 4 models for both architectures. Modeling a criterion was, therefore, more challenging as we go from criterion 1 through criterion 4. Specifically, the "positive" label prediction for Task 1 (format assessment) was 0.952 F for the Cascade Learner, but the positive label performance of the

Task 4 (rigor assessment) model reduced to 0.778 F, a significant drop of nearly 18% in absolute terms. This performance reduction is even worse for Ensemble-Boolean which analyzes all samples. For the latter stage criteria, the Cascade Learner performs significantly better than the Ensemble-Boolean giving it an edge overall. Once again, we see how filtering out negative samples early improves the F measure. Note that the reduction in F Measure was mainly due to the decrease in precision.

4.3. The impact of sampling ratio

We show the impact of training samples ratios in Table 6 for the ITL model. We observe that as negative samples are reduced from the baseline, recall improves at the cost of precision, and when negative samples are increased from the baseline, precision improves at the cost of recall. Not only the ratio but the number of samples seem to matter as well because when positives were roughly 10 times oversampled and

Table 5
P,R,F Values for each Step in Cascade for both Cascade / Ensemble using the full dataset.

Model	Task Number	Labels	P	R	F
Cascade Learner	Task 1- Format	Not O	0.944	0.952	0.948
		O	0.956	0.948	0.952
	Task 2- HHC	F	0.748	0.827	0.785
		T	0.956	0.931	0.944
	Task 3- Purpose	Not Tr	0.948	0.858	0.901
		Tr	0.727	0.890	0.800
	Task 4- Rigor	F	0.973	0.884	0.927
		T	0.679	0.910	**0.778**
Ensemble-Boolean	Task 1- Format	Not O	0.944	0.952	0.948
		O	0.956	0.948	0.952
	Task 2- HHC	F	0.795	0.812	0.804
		T	0.953	0.948	0.951
	Task 3- Purpose	Not Tr	0.965	0.817	0.885
		Tr	0.676	0.927	0.782
	Task 4- Rigor	F	0.986	0.773	0.867
		T	0.505	0.957	**0.661**

PT = no, 256 sequence length, 1.5 K:1.5 K sampling, no PT tag.

Table 6
Different sampling ratios (Marshall dataset), ITL, seq len = 256, PT = no.

Positive samples	Negative samples	Precision	Recall	F Measure
1.5 k	0.317 k	0.273	**0.986**	0.428
1.5 k	0.353 k	0.352	0.982	0.519
1.5 k	0.488 k	0.410	0.979	0.578
1.5 k	0.756 k	0.487	0.965	0.647
1.5 k	1.5 k	0.383	0.962	0.547
15 k	15 k	0.648	0.860	**0.739**
15 k	30 k	0.685	0.739	0.711
15 k	49 k	0.702	0.659	0.680
30 k	30 k	0.696	0.724	0.710
30 k	49 k	**0.707**	0.647	0.676
49 k	49 k	0.705	0.609	0.653

negatives were down sampled to the same number (i.e. 15 K positive and 15 K negative samples), ITL achieved its best F measure balancing the recall and precision. The substantial performance improvement for this sampling (almost 0.20 F measure) is an indication that the additional negative samples trained the ITL model better. Further research is needed to answer the question of optimal training ratio and size when data is highly imbalanced. The highest recall (0.986) was achieved with highly skewed number of positive samples.

4.4. The Impact of sequence length and PT tag

For the results reported so far, the sequence length was 256 and the PT tag was not used. We present the impact of these factors as an "ablation" study as shown in Table 7. For clarity, we discuss results for ITL and Cascade Learner (the top two performing models) only. We started with the base configuration for each model and then successively changed features. For ITL, we started with the sequence length of 256 (which is 69th percentile) and equal number of positive and negative training samples without up sampling positives (1.5 k and 1.5 k) and, increased the sequence length to 384 (92nd percentile), training samples to 15–15 K by up-sampling positives, and finally added PT tag text to the input. Similarly, for Cascade Learner, started with 256 sequence length and balanced training data at each component model, and increased sequence length to 384, and finally add PT tag text to the input.

The best recall (0.962) was achieved with 1.5 k-1.5 k baseline sampling ratio. However, 10× training samples significantly improved precision for a small loss in recall, resulting in a nearly 0.20 improvement in absolute F measure. This may not be helpful where high recall is needed (such as systematic reviews) but for other downstream applications (e.g. "search" and "news feed") that require low noise, it could be of importance. Further changes in sequence length and adding PT tag had only minor performance improvement in the F measure of ITL.

As mentioned earlier, we always used balanced positive and negative samples at each component in Cascade Learner, and as we changed sequence length and added PT tag text to the input, the performance improved by a small degree, resulting in a final precision, recall, and F measure values of 0.669, 0.891, and 0.764 respectively.

In order to provide a broader perspective, we compared our results with Del Fiol et al [9] and Marshall et al, [10] although there are some experimental differences among the three studies. Both previous studies used (different) weakly labeled separate training dataset, while we conducted a cross validation study using the "strongly" labeled data. Del Fiol et al tested on a subset of Clinical Hedges dataset. We used the ITL model for the comparison because it was clear from the earlier sections that it was better at achieving high recall.

Table 7
Ablation test (Marshall data set).

Model	Configuration	Precision	Recall	F Measure
ITL	1.5 K-1.5 k,256	0.383	**0.962**	0.547
	15 K-15 K, 256	**0.648**	0.860	0.739
	15 K-15 K, **384**	0.640	0.860	0.734
	15 K-15 K, 384, **PT tag**	0.642	0.880	**0.742**
Cascade Learner	Balanced, 256	0.660	0.880	0.754
	Balanced, **384**	0.644	0.890	0.747
	Balanced, 384, **PT tag**	**0.669**	**0.891**	**0.764**

Del Fiol et al [9] reported only one set results for a neural network (CNN) model, which achieved 0.969 recall, 0.346 precision, and 0.510 F measure. The study also reported results from earlier studies for a customized PubMed search filters, referred to as the McMaster's clinical query (CQ) filters. Performance of two kinds of filters (balanced and broad), at recall 0.970 and 0.984 respectively, were discussed. The two filters achieved 0.409 and 0.224 precision respectively.

As shown in Table 8, with recall fixed at ~0.970, the ITL classifier had higher precision (0.515 vs. 0.346 vs. 0.409) and higher F-measure (0.672 vs. 0.510 vs. 0.575) than CNN and the McMaster's CQ balanced filter respectively. In addition, with recall fixed at ~0.985, ITL had higher precision (0.381 vs. 0.224) and higher F-measure (0.550 vs. 0.365) than the McMaster's CQ balanced filter. The ITL model results for ~0.97 recall were obtained using 1.5 k-1.5 k (positive and negative) training samples, 384 sequence length, and by including PT tag in the input. The results for ~0.985 recall were obtained using 1.5 k-0.46 k (positive–negative) training samples, 256 sequence length, but not the PT tag.

The Marshall et al [10] study reported best precision results at 0.985 recall using SVM (support vector machines) in a voting ensemble with a PT tag classifier (see the study for details). As shown in Table 9, with recall fixed at ~0.98, the ITL classifier had higher precision (0.356 and 0.275 vs. 0.210) and higher F-measure (0.521 and 0.430 vs. 0.346) than the SVM model, at two slightly different operating points around 0.985 recall.

The two operating points were obtained with 1.5 k-0.353 k and 1.5 k-0.317 k training samples respectively. A linear interpolation between the two yields 0.295 precision at 0.985 recall.

5. Discussion

ITL consistently achieved better recall (in multiple feature configurations) than the ensembles including Cascade Learner. Thus, modeling all the criteria together seem to allow the model to use weak signals from one or more criteria to classify an article correctly. However, Cascade Learner achieved the highest precision and the highest F measure compared to ITL and other the ensembles, although it lost recall compared to all of them. Therefore, modeling each criterion separately appears to help the ensemble to focus on increasingly stronger signals from each criterion and hence deliver high precision. We observe two key findings here: First, ITL is most suitable for high recall applications such as systematic reviews and second, Cascade Learner is most suitable for high F measure applications. The Cascade Learner functionality and its results are analogous to query optimization in IR and in databases where rearranging the query terms is known to improve ranking (in IR) and query execution speed (in databases).

Since negatively classified samples are discarded early, Cascade Ensemble is unable to combine the weak signals from multiple criteria in later stages to achieve a high inclusion rate. This suggests future work to investigate how to give negatively classified samples another chance later in the cascade ensemble.

Results also suggest that among the individual sub-models of the ensembles, the model for the last stage, i.e. the model that tests for the scientific "rigor" requirements, was the most challenging (achieves the lowest F measure). So, determining whether a study was rigorously conducted or not is more difficult than determining if a study is an original research. Testing whether a process was followed or not, is harder even for a state-of-the-art language model such as SciBERT. Therefore, an area of future work is to improve modeling of this semantically challenging aspect of screening articles.

Another observation from our study is that since neural networks

Table 8
Del Fiol subset – comparison- Statistical Significance.

	Model	Parameters/Details	Performance		
			P	R	F
Performance at ~0.970 Recall	Del Fiol et al	CNN	0.346	0.969	0.510
	McMaster's CQ balanced filter	Customized text query	0.409	0.970	0.575
	The ITL Classifier	SciBERT, 1.5 k-1.5 k, 384, PT	**0.515** (±0.029)	0.969 (±0.006)	0.672 (±0.024)
Performance at ~0.985 Recall	McMaster's CQ broad filter	Customized text query	0.224	0.984	0.365
	The ITL Classifier	SciBERT, 1.5 k-0.46 k, 256	**0.381** (±0.021)	0.984 (±0.008)	0.550 (±0.022)

Table 9
P, R, F Values for different Proportions of Train-Test Data – Marshall Subset- 1.5 k-Varying Neg Samples, 384, PT tag.

	Model	Parameters/ Details	Performance		
			P	R	F
Performance at ~0.985 recall	Marshall et al	SVM + Voting with PT tag	0.210 (±0.010)	0.985 (±0.007)	0.346 (±0.014)
	The ITL classifier	SciBERT, 1.5 k-0.353 k, 384, PT tag	**0.356** (±0.031)	0.982 (±0.007)	0.521 (±0.034)
	The ITL classifier	SciBERT, 1.5 k-0.317 k, 384, PT tag	**0.275** (±0.017)	0.986 (±0.008)	0.430 (±0.021)

optimize for accuracy, obtaining high recall results can be a challenge. We used two techniques for obtaining high-recall operating points. One technique was adjusting the threshold on the probability distribution for the positive label (to decide if the outcome was "Yes" or "No"). Because neural models are highly non-linear, we found that the room for adjustment was rather small. The second technique was to train the model with different ratios of positive and negative samples – i.e. optimizing the model for a different sample ratios. Both are rather crude techniques (requiring much trial and error) as applied to neural networks. A better approach might be to develop optimization algorithms designed for high recall rather than for accuracy.

In our experiments we noticed that the model performance was not only dependent on the ratio of positive and negative samples in the training set, which was expected, but was also dependent on the number of samples used. As shown in Tables 6 and 7, when the samples were increased from 1.5 k to 15 k, the performance of ITL changed by large measure (independent of which direction was more helpful for a downstream application) – although the positive samples were simply duplicated 10 times. However, further increases did not change the performance very much. There appears to be an optimum number of (random) samples needed to train a model well, even if only the negative samples provide variance while positive samples are simply duplicated.

One of the surprising results was that the sequence length and PT tags did not improve the performance by any significant margin. Our conclusion was that the beginning of an abstract must contain the most useful information and the ending part had few additional clues relevant to the decision making. We also surmised that the semantic information carried in the PT tag may have been already inferred from the text by the model.

The study has a few limitations. First, we used a single dataset, although remarkable in its detail and curation, additional datasets would broaden the results. Second, unlike the previous studies, we used cross-validation on a single dataset, which may introduce certain "optimism" bias (as is the case even when a set-aside dataset is used if both the train and test sets are drawn from the same source and during the same time period). Third, our focus was on a specific *study type* selection criteria and other selection criteria, including drugs, interventions, diseases, genes, or other risk factors, may yield different results. Four, we considered only one BERT variant and other variants might yield further insights. Lastly, we identified several opportunities for future work which when considered may affect our conclusions.

6. Conclusion

We made an extensive study of how to apply the latest neural network models when the task of filtering articles involves a combination of criteria, using the Clinical Hedges dataset, a highly imbalanced dataset (1:32, positive to negative ratio) of abstracts and titles extracted from MEDLINE. We studied three ensembles, including a novel form of cascading ensemble, and the single integrated model. The overall observation was that while the cascading ensemble achieved the highest precision and the highest F-measure, the single integrated model achieved the highest recall. SciBERT was an effective core model for use in these architectures.

Results indicate that when text classification involves multiple independent criteria that a cascading ensemble of criterion-specific models is superior in terms of precision and F measure results. This is a significant result that generalizes beyond this task and the dataset, and has a similarity to query optimization in IR and in databases, and may be useful for downstream applications (such as the news feed in Facebook) that require low noise but can tolerate missing a few relevant articles.

The single integrated model however consistently achieved higher recall, suggesting that it would be a better architecture for high recall (sensitivity) applications which can tolerate slightly higher noise level. For systematic reviews this is an important characteristic where researchers are willing to review twice as many articles as they would ultimately include but do not want to miss even a few includable articles.

We identified several interesting new research questions. First among them is if we could effectively combine the high precision of the cascading ensemble with the high recall of a single integrated model. Second, if neural network optimization algorithms can be tuned to emphasize recall rather than the accuracy. Third, since we observed that the training sample size and ratio had a significant impact of the performance (especially in highly imbalanced dataset) how to determine the optimum training size and ratio without resorting to an extensive empirical analysis. The code for the models developed in this study is freely available at https://github.com/md-labs/Clinical_Hedges_BERT.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We gratefully acknowledge help and guidance from Dr. R. Brian Haynes in using the Clinical Hedges dataset and understanding related research. We also thank the anonymous referees for helping to significantly enhance the manuscript through fact checking and concrete suggestions to improve the writing style and results presentation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2020.103578.

References

[1] R.B. Haynes, Where's the meat in clinical journals, ACP J. Club 119 (3) (1993) A22–A23.

[2] N. Fiorini, K. Canese, G. Starchenko, et al., Best Match: new relevance search for PubMed, PLoS Biol. 16 (8) (2018) 1–12, https://doi.org/10.1371/journal.pbio.2005343.

[3] R.B. Haynes, N. Wilczynski, K.A. McKibbon, C.J. Walker, J.C. Sinclair, Developing optimal search strategies for detecting clinically sound studies in MEDLINE, J. Am. Med. Inf. Assoc. 1 (6) (1994) 447–458.

[4] Wilczynski NL, Morgan D, Haynes RB, Team H, An overview of the design and methods for retrieving high-quality studies for clinical care, BMC Med. Inf. Decis. Making 5 (20) (2005), https://doi.org/10.1186/1472-Received.

[5] N.L. Wilczynski, K.A. Mckibbon, S.D. Walter, A.X. Garg, R.B. Haynes, MEDLINE clinical queries are robust when searching in recent publishing years, J. Am. Med. Inf. Assoc. 20 (2) (2013) 363–368, https://doi.org/10.1136/amiajnl-2012-001075.

[6] H. Kilicoglu, D. Demner-Fushman, T.C. RindFlesch, N.L. Wilczynski, R.B. Haynes, Towards automatic recognition of scientifically rigorous clinical research evidence, J. Am. Med. Informatics Assoc. 16 (1) (2009) 25–31, https://doi.org/10.1197/jamia.M2996.

[7] Y. Aphinyanaphongs, I. Tsamardinos, A. Statnikov, D. Harding, C.F. Aliferis, Text categorization models for high-quality article retrieval in internal medicine, J. Am. Med. Inf. Assoc. 12 (2) (2005) 207–216.

[8] E.V. Bernstam, J.R. Herskovic, Y. Aphinyanaphongs, C.F. Aliferis, M.G. Sriram, W. R. Hersh, Using citation data to improve retrieval from MEDLINE, J. Am. Med. Inf. Assoc. 13 (1) (2002) 96–105.

[9] G. Del Fiol, M. Michelson, A. Iorio, C. Cotoi, A Deep Learning Method to Automatically Identify Reports of Scientifically Rigorous Clinical Research from the Biomedical Literature : Comparative Analytic Study 20 (2018) 1–12. doi: 10.2196/10281.

[10] I.J. Marshall, A. Noel-Storr, J. Kuiper, J. Thomas, B.C. Wallace, Machine learning for identifying Randomized Controlled Trials: an evaluation and practitioner's guide, Res. Synth Methods. 9 (4) (2018) 602–614, https://doi.org/10.1002/jrsm.1287.

[11] J. Devlin, K. Lee, M. Chang, ToutaKristina. BERT : Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT. Minneapolis, MN; 2019, pp. 4171–4186.

[12] Y. Liu, M. Ott, N. Goyal, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv Prepr arXiv 190711692v1. (2019) (1).

[13] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma P, Soricut R. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations, in: Proceedings of ICLR 2020, 2020, pp. 1–17.

[14] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: Proceedings Ofthe 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China; 2019, pp. 3615–3620.

[15] J. Lee, W. Yoon, S. Kim, et al. BioBERT : a pre-trained biomedical language representation model for biomedical text mining, arXiv Prepr arXiv190108746, 2019.

[16] K. Huang, J. Altosaar, R. Ranganath, Clinical bert : modeling clinical notes and predicting hospital readmission, arXiv Prepr arXiv190405342v2, 2019, pp. 1–19.

[17] Y. Si, J. Wang, H. Xu, K. Roberts, Enhancing clinical concept extraction with contextual embeddings, J. Am. Med. Inf. Assoc. 26 (11) (2019) 1297–1304, https://doi.org/10.1093/jamia/ocz096.

[18] H. Guan, M. Devarakonda, Leveraging contextual information in extracting long distance relations from clinical notes, in: AMIA Annual Symposium Proceedings. Washington, District of Columbia; 2019, 1051–1060.

[19] W.R. Hersh, Information Retrieval: A Health and Biomedical Perspective, Springer Science & Business Media, 2008.

[20] W.R. Hersh, Information retrieval for healthcare, in: C.K. Reddy, C.C. Aggarwal, (Eds.) Healthcare Data Analytics, Chapman and Hall, 2015 (Chapter 14).

[21] P.M. Marrero, S. Sánchez-cuadrado, J. Urbano, J. Morato, J. Moreiro, Information retrieval systems adapted to the biomedical domain, arXiv Prepr arXiv12036845. 2012 (March). doi:10.3145/epi.2010.may.04.

[22] B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, M. Lawley, Information retrieval as semantic inference: a Graph Inference model applied to medical search, Inf. Retr. Boston. 19 (1–2) (2016) 6–37, https://doi.org/10.1007/s10791-015-9268-9.

[23] B. Mitra, N. Craswell, Neural Models for Information Retrieval, 2017. http://arxiv.org/abs/1705.01509.

[24] C. Hauff, Machine Learning for IR. Slides. https://rure.cs.ru.nl/siks/claudia-hauff_ml-for-ir.pdf. Published 2019. Accessed September 2, 2020.

[25] Z.A. Yilmaz, S. Wang, W. Yang, H. Zhang, J. Lin, Applying BERT to document retrieval with birch, in: Proceedings Ofthe 2019 EMNLP and the 9th IJCNLP (System Demonstrations). Hong Kong, China, 2019, pp. 19–24.

[26] W. Yang, H. Zhang, J. Lin, Simple Applications of BERT for Ad Hoc Document Retrieval, arXiv Prepr arXiv 190310972v1, 2019.

[27] R.B. Haynes, Clinical Hedges - Health Information Research Unit, https://hiru.mcmaster.ca/hiru/HIRU_Hedges_home.aspx.

[28] NLM, MedLine Publication Types, Web Page, https://www.nlm.nih.gov/mesh/pubtypes.html. Accessed July 29, 2020.