How Can Explainable Artificial Intelligence Accelerate the Systematic Literature Review Process?

Seye Abogunrin[1], Siva Karthick[2], Gaugarin Oliver[2], Marie Lane[1], Andreas Witzmann[1], Samuel Kumaresan[2] | 1- F. Hoffmann-La Roche Ltd., Basel, BS, Switzerland; 2 - CapeStart, Inc, Massachusetts, United States of America

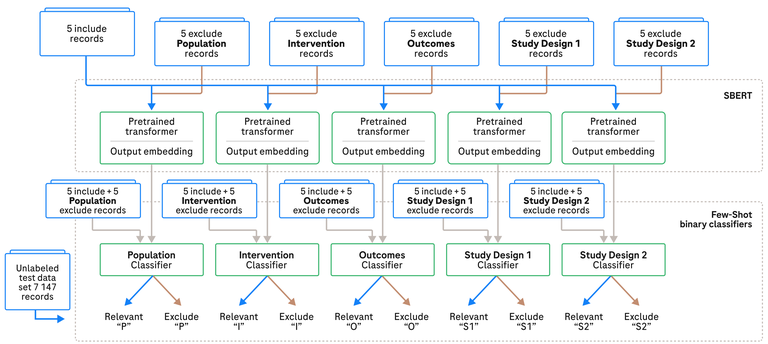
BACKGROUND

- The exponential increase in published articles makes a thorough and practical Systematic Literature Review (SLR) increasingly challenging.
- Traditional machine learning algorithms require a large amount of labeled data to train on and consequently massive computational resources to perform. This can be a limitation when dealing with tasks that have limited labeled data,
- In a prospective SLR setting, the effort to identify the training records, balanced across the different exclusion reasons, could result in a huge proportion of the data set being screened, thus negating the potential time saving benefits when applying automation to TIABS.
- Few-shot learning algorithms aim to generalize from a small set of labeled examples to classify new instances accurately.
- To help researchers conduct an SLR, we developed a machine learning (ML)-based pipeline to accelerate the title and abstract screening (TIABS) step using a transfer learning approach. We compared the results of the ML-based TIABS using human-labeled SLRs to ensure its reproducibility.

METHODS

- Human-labeled data were sourced from three systematic literature reviews (SLRs) focused on different therapeutic areas - NSCLC, CRPC and COVID-19.
  - Each record had an include or exclude status, and where excluded, a reason for exclusion was also available.
  - A subset of these data were used for training and validation purposes.
- Three separate pipelines were developed for the three SLRs.
  - The NSCLC SLR had five exclusion reasons and so its pipeline comprised five models: Population, Intervention, Outcomes, Study Design 1 (non-empirical studies), Study Design 2 (empirical studies)
  - The CRPC and COVID-19 SLRs had four exclusion reasons each and so their pipelines had four models each: Population, Intervention, Outcomes, Study Design (empirical and non-empirical studies).

Figure 1: Few-Shot learning model for the NSCLC review



- The TIABS pipeline was developed as follow - **Figure 1**:
  - Five include and five exclude title and abstract human-labeled records were identified for each exclusion reason for each of the three sets of reviews and used to fine-tune a pre-trained transformer (SBERT-Sentence Bidirectional Encoder Representations from Transformers) model.
  - The fine-tuning process led to the generation of embeddings. These embeddings were then used for the training of few-shot binary classifiers - NSCLC: 30; CRPC: 25; and COVID-19: 30.
  - Each classifier was trained on one review exclusion criteria and the included records.
  - The trained few-shot classifiers were then assessed for robustness by validating their performance with human-labeled titles and abstracts previously unseen by the models - NSCLC: 1596; CRPC: 752; and COVID-19: 935.
  - The validated few-shot models were then used to assign an include and exclude status to unlabeled titles and abstracts - NSCLC: 7 147 (575 included and 6 572 excluded human labels); CRPC: 2 594 (702 included and 1 892 excluded human labels); and COVID-19: 4 002 (319 included and 3 683 excluded human labels).
  - Various combinations of SBERT models and ML models, such as Gaussian Naive Bayes, Logistic Regression, and Support Vector Machine, were experimented with. Ultimately, the SBERT 'paraphrase-mpnet-base-v2' model with a hyperparameter-tuned Support Vector Machine was found to yield optimal model performance.

RESULTS

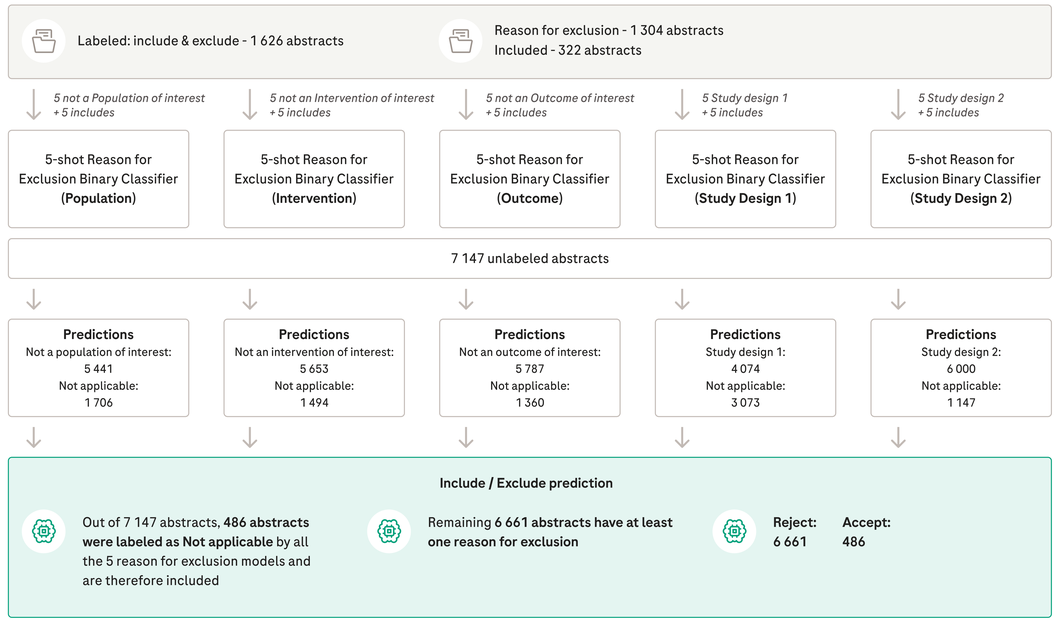
The evaluated SLR datasets were previously used to support research questions in non small cell lung cancer (NSCLC), castration-resistant prostate cancer (CRPC) and COVID-19.

The sizes of the human-labeled review title and abstracts datasets used to test the pipelines are:

NSCLC	CRPC	COVID-19
7 147	6 572	4 002

The datasets are used for both training and validation of the 5-shot classifiers. The accuracy of the 5-shot classifiers ranged from 0.64 to 0.95. The accuracy, recall, precision, and F-measure of the TIABS pipeline ranged between 0.70 and 0.90, 0.32 and 0.57, 0.33 and 0.59, 0.32 and 0.57, respectively.

Figure 2 - Schematic presentation of the title and abstract screening of NSCLC data using 'Explainable AI pipeline using 5 shot binary classifiers'



TIABS Pipeline Performance Results on the Test Data:

Table 1 - Results of SLR 1 (NSCLC) — Accuracy: 0.90

	Precision	Recall	F1-score	Data
Exclude	0.94	0.95	0.95	6 572
Include	0.38	0.32	0.35	575
macro avg	0.33	0.32	0.32	7 147
weighted avg	0.9	0.9	0.9	7 147

Table 2 - Results of SLR 2 (CRPC) — Accuracy: 0.70

	Precision	Recall	F1-score	Data
Exclude	0.76	0.86	0.81	1 892
Include	0.42	0.28	0.34	702
macro avg	0.59	0.57	0.57	2 594
weighted avg	0.67	0.7	0.68	2 594

Table 3 - Results of SLR 3 (COVID) — Accuracy: 0.84

	Precision	Recall	F1-score	Data
Exclude	0.93	0.89	0.91	3 683
Include	0.16	0.23	0.18	319
macro avg	0.36	0.37	0.37	4 002
weighted avg	0.87	0.84	0.85	4 002

DISCUSSION

We have developed an automated SLR workflow that uses an adapted transfer learning approach that requires very little data.

We leveraged a pre-trained BERT model, which has knowledge on a large and diverse dataset, and used it to train a new model on a smaller dataset with specific exclusion goals. By doing so, we were able to use transfer learning to significantly reduce the amount of training data needed to achieve high accuracy levels.

The conventional method of developing systematic literature reviews (SLRs) usually requires a considerable manual effort, and it can be challenging to generalize it for various studies. While automated text classification using machine learning, such as title and abstract screening (TIABS), can alleviate this issue, the pipeline used must be explainable by providing the reasons for exclusion.

CONCLUSION

The few-shot classifier metrics for each research question demonstrates automated TIABS screening can be conducted with a small training data set (i.e. 5 records for each class).

The use of retrospective SLRs eliminated the effort to identify the training data.

Further experimentation with a prospective SLR is required to compare and evaluate the speed and accuracy of manual versus AI-assisted screening using this approach, considering training data identification effort.