Natural language processing to support abstract selection in systematic literature reviews

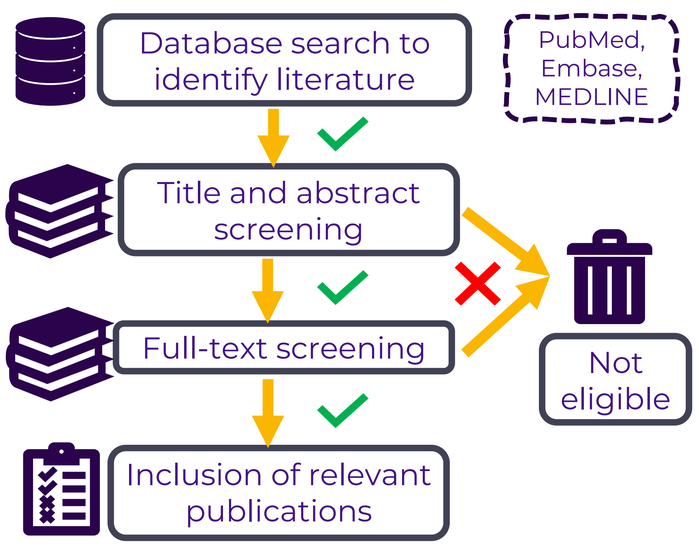
Putnam Inizio Advisory

M Damentko[1], B Hasni[2], B Abdellaoui[2], S Barbier[3], N Mateusz[1]
[1]Putnam, Krakow, Poland; [2]Putnam, Tunis, Tunisia; [3]Putnam, Lyon, France

Acceptance code MSR66

Introduction

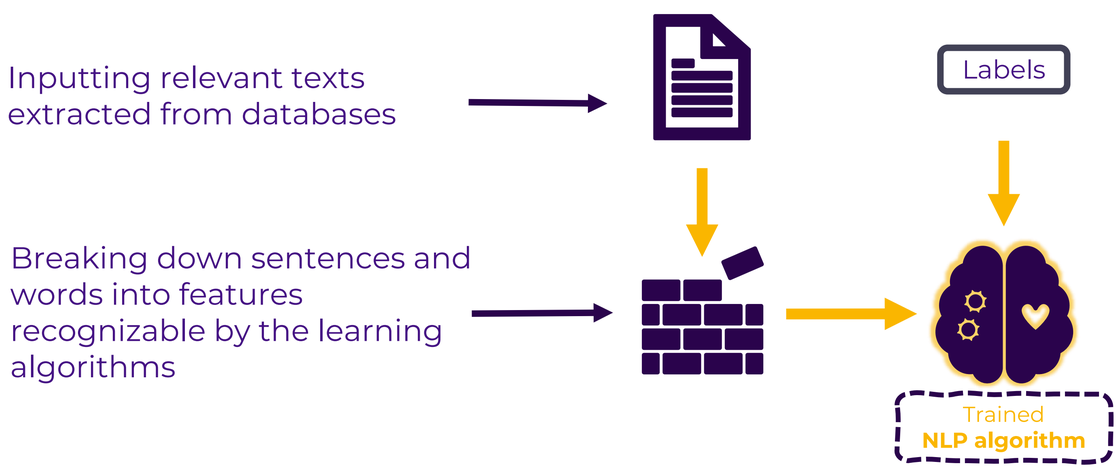
Figure 1. Simplified PRISMA SLR methodology



Literature review is a **time-consuming process** that requires many human resources (Figure 1). However, it is necessary for valid meta-analyses. The number of scientific publications continues to steadily increase, which may prove **difficult to inspect solely by human eyes**. Simultaneously, it is **crucial to maintain high-quality** reviews to meet the strict requirements of HTA institutions.

NLP and text-mining methods have been proposed to improve SLR processes (Figure 2) .

Figure 2. Training phases of the NLP algorithm



Features fed to machine-learning algorithms are **associated with pre-loaded labels.** For example, NLP learns which feature is a **positive sign** and which is **negative**. Thus, it can predict the probability of obtaining a specific classification.

Methods

The NLP algorithm we used was an **ensemble of 4 models**: SciBERT, BioBERT, BioMed-Robert, and ClinicalBERT. The BERT models were pre-trained on different, large databases.
The BERT architecture is based on a **multi-layer bidirectional transformer**.

This approach was selected based on parameters such as accuracy, precision, and sensitivity.

The algorithm was then **fine tuned.**

SLR data sets were divided into training and testing groups using **50/50** and **30/70** ratios. The training data were further divided into basic model training and validation data at a ratio of 90/10.

Labels were provided as **binary decisions** of the human reviewers (exclusion or inclusion).

To balance it, an **under-sampling method** was employed to mitigate the bias toward the majority class. The algorithm forced a ratio of at least 1/5.

The algorithm was **tested** on the remaining fraction of the SLR data set **to assess its performance against humans** for accuracy, sensitivity, and specificity.

Abbreviations: HTA, health technology assessment; NLP, natural language processing; SLR, systematic literature review

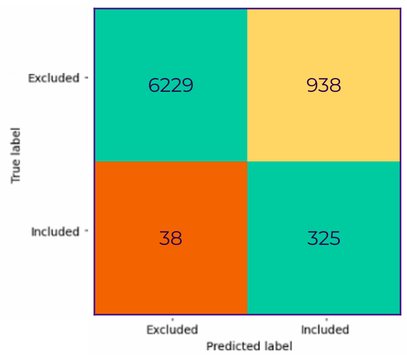
Results

We evaluated our NLP algorithm on 6 historical SLR data sets to check its effectiveness. The largest contained 15,059 records and the smallest 1,904. They intentionally covered a wide range of topics, from economic models to clinical endpoints.

When the algorithm replaced both reviewers, the sensitivity (true included rate) and specificity (true excluded rate) were 0.90 and 0.87 (Figure 3). If only 1 of the reviewers was replaced by the algorithm, the rates were 0.97 and 0.99, respectively. With this approach, without compromising the quality, the human reviewer could avoid reading >6800 abstracts, requiring additional review only to assess discrepancies between the human reviewer and the model. Assuming a human screening rate of 400 items per day, **this could save >17 person-days** and up to **38 spared days** when the NLP evaluated half of the abstracts instead of both reviewers.

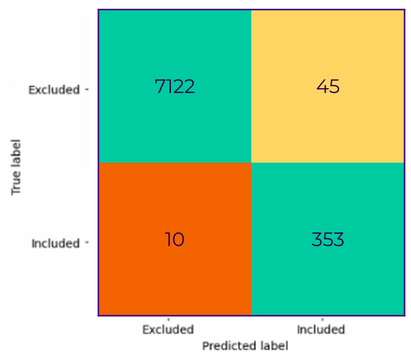
Figure 3. The largest data set: SLR on economic models (15,059 records)

Replacement for both reviewers
(data used for training vs used for testing):
7529/7530 records



	Accuracy	Sensitivity	Specificity
Results	0.87	0.90	0.87

Replacement for 1 reviewer



	Accuracy	Sensitivity	Specificity
Results	0.99	0.97	0.99

Accuracy: the overall probability that abstracts will be correctly classified
Sensitivity: the probability that included abstracts will be correctly classified as included
Specificity: the probability that excluded abstracts will be correctly classified as excluded

The effectiveness of the algorithm varied, depending not only on the number of abstracts used to train the model but also on the topic of the review, the disease state, and the population (Table 1). Effectiveness was also influenced by the imbalance in the binary decisions of the reviewers—usually, there were more excluded studies than included ones.

Table 1. Comparison of results for all historical SLRs

Data set	Ratio, training: testing	Number of abstracts	Accuracy	Sensitivity	Specificity	Spared days
Economic models	50/50	7529/7530	0.87	0.90	0.87	38
	30/70	4517/10,542	0.92	0.83	0.92	53
Chronic kidney disease	50/50	1695/1696	0.78	0.81	0.77	8
	30/70	1017/2374	0.83	0.76	0.84	12
Immunology-mediated diseases	50/50	2715/2715	0.76	0.75	0.76	14
	30/70	1629/3801	0.83	0.75	0.84	19
Heart failure	50/50	952/952	0.82	0.81	0.82	5
	30/70	571/1333	0.78	0.70	0.79	7
Metastatic cancers	50/50	1475/1475	0.76	0.84	0.73	7
	30/70	885/2065	0.73	0.82	0.69	10
Immune thrombocytopenia	50/50	1482/1482	0.77	0.75	0.70	7
	30/70	889/2075	0.69	0.54	0.70	10

Conclusions

Including NLP algorithms in the SLR process allowed for **considerable acceleration** while mantaining high-quality standards. Such algorithms may prove particularly useful when dealing with **large SLRs,** for which reviewing requires substantial time.

The NLP algorithm could act as a second reviewer, assessing in parallel to the human reviewer or doing part of the work for both analysts on the remaining unassessed abstracts.

More refinement is necessary to cater to more types of SLRs and to confirm that the quality is acceptable to HTA bodies.

Contact
Sylvaine Barbier
Sylvaine.Barbier@putassoc.com

Find out more at putassoc.com