MSR131

Prospects for Automation of Systematic Literature Reviews (SLRs) with Artificial Intelligence and Natural Language Processing

Royer J[1]; Wu EQ[2]; Ayyagari R[2]; Parravano S[2]; Pathare U[2]; Kisielinska M[2]

[1]Analysis Group, Inc., Montreal, QC, Canada; [2]Analysis Group, Inc., Boston, MA, USA

Introduction

- Systematic literature reviews (SLRs) are an essential component of evidence synthesis for health care economics and outcomes research (HEOR) and can inform clinical decision-making
- Currently, the process of preparing SLRs is greatly labor and resource-intensive, as it typically requires at least two expert human reviewers for the initial review, and then an additional reviewer to reconcile any discourse over studies to be included or excluded
- SLRs of high-quality (i.e., reviews that use explicit, reproducible, and applies systematic methods to minimize bias and maximize the recall rate) can take about 6 months to complete, or longer if the topic area is well-studied and has a large body of existing literature
- Given the value of these reviews in HEOR and clinical practice, and the exponentially growing volume of medical literature across study areas, it has become increasing important to assess potential applications of artificial intelligence (AI) in streamlining this process

Objective

- This research explores the performance of the latest AI techniques to assist with SLRs, with the goal of improving review time while maintaining high accuracy

Conclusions

- AI methods can be successfully applied to the process of developing SLRs to improve review time while maintaining high accuracy
- By implementing AI, human reviewers have the potential to focus their time more on uncertain/inconclusive model recommendations, while the AI screening methods excludes high-confidence irrelevant abstracts
- Further research should be undertaken to assess the capabilities of AI in completing abstract screening for other forms of literature reviews (i.e., targeted literature reviews), and across disease areas

Methods

Data source

- This study leverages data on abstracts from an SLR on attention-deficit/hyperactivity disorder (ADHD)-related studies, to assess the performance of AI methods in preparing a high quality SLR
- Of the 773 abstracts reviewed for the SLR on ADHD-related studies by human reviewers, 29.1% were included, and 70.9% were excluded
- In the data set, the eligibility classification of abstracts were based on a set of 2 inclusion criteria and 3 exclusion criteria

Table 1. Summary of ADHD data set

Total [N]	Criteria		Abstract decisions	
	Inclusion [N]	Exclusion [N]	Included [N (%)]	Excluded [N (%)]
768	2	3	217 (28.3 %)	551 (71.7%)

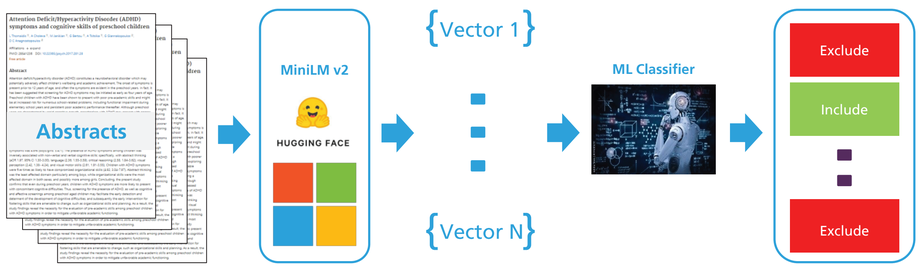
Study design

- A pre-trained sentence embedder (i.e., Microsoft's MiniLM v22) and OpenAI's GPT-3.5 (GPT-3.5) method were used

Embedding method

- The Embedding method required abstracts to be vectorized and then a series of classifiers were trained to predict the inclusion or exclusion of abstracts on the test dataset (**Figure 1**)
- To determine the best performing classifier for the Embedding method, the yield of the highest number of excluded abstracts at a given level of recall was used
- To obtain empirical confidence intervals (CI) for statistical estimates at a 90% confidence level, the process of training and assessing classifier performance was conducted over 1000 times

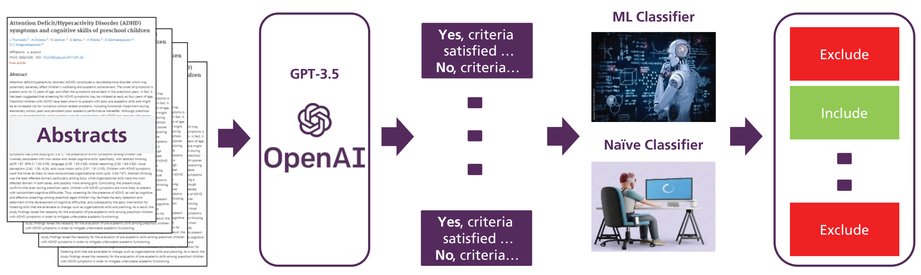
Figure 1. Embedding method



GPT-3.5 method

- The GPT-3.5 method required abstracts to be assessed using a set of questions for each inclusion and exclusion criteria, where GPT-3.5 responses were recorded and then converted into binary outputs and assigned to each abstract per criteria-related question asked
- To convert the GPT outputs into inclusion/exclusion predictions, we first computed an overall sum of relevance by averaging the binary values (i.e, "Yes"/"No") and then applied a naïve method (i.e., a method that requires no additional training), where we excluded the abstracts with the lowest 25th percentile of probability values [0.00 – 0.20] (**Figure 2**)
- Alternatively, we used a Logistic classifier trained on top of the binary inclusion/exclusion criteria values, by treating the GPT-3.5 responses to the inclusion/exclusion criteria as the independent variable, and the human labels as the dependent variable
- For both the Embedding method and GPT-3.5 methods, the training dataset consisted of 20% [154/773] of total abstract, and the testing dataset consisted of 80% [614/773] of total abstract

Figure 2. OpenAI's GPT-3.5 method



Study outcomes

- Model performance outcomes included recall rate (i.e., capture rate of relevant abstracts), proportion of abstracts excluded, type 1 errors (i.e., false-positive) and type 2 (i.e., false-negative) errors, and were assessed against abstract classifications in the SLR dataset based on human reviewers

Results

Overall

- The Embedding and GPT-3.5 methods were able to exclude irrelevant abstracts at a specified confidence level, and key advantages for each method were uncovered
- Moreover, applying a 20/80 split for the training and testing datasets allowed for the best predictions

Embedding method

- The Embedding method was able to exclude larger proportions of irrelevant studies
- More particularly, the proportion of irrelevant abstracts excluded was 40.1% (+1.6/-1.8), with a model recall rate of 94.2% (+1.2/-0.6) (CI: 90%) (**Table 2; Figure 3**)

GPT-3.5 method

- The GPT-3.5 method was able to exclude abstracts without being trained on a naïve dataset and provided a relevance score
- More particularly, the proportion of irrelevant abstracts excluded was 25.0%, with a model recall rate of 95.8% (**Table 2; Figure 3**)
- By using the GPT-3.5 responses in a logistic regression, the proportion of irrelevant abstracts excluded was 40.4% (+1.3/-1.6), with a model recall rate of 96.5% (+1.0/-0.9) (CI: 90%) (**Table 2; Figure 3**)

Figure 3. Excluded abstracts vs abstracts wrongly excluded (a) Embedding method (b) GPT 3.5 method

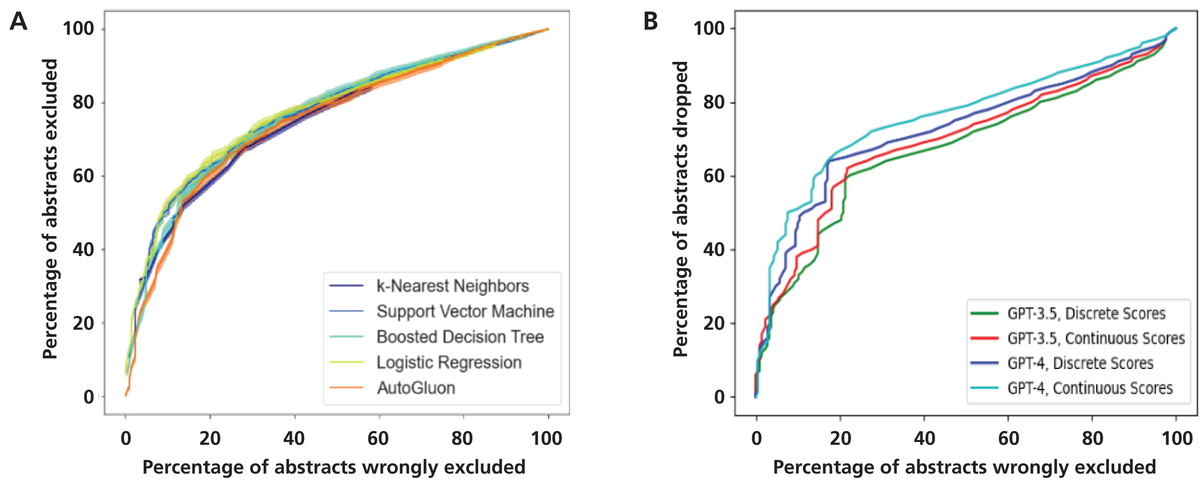


Table 2.Summary of study outcomes by method

	Excluded (%) +/- CI	Recall (%) +/- CI	Best performing classifier
Method applied			
Embedding method	$40.1^{+1.6/-1.8}$	$94.2^{+1.2/-0.6}$	Support vector machine
GPT 3.5[1]	25.0	95.8	Naïve Rankings
Plus logistic regression	$40.4^{+1.3/-1.6}$	$96.5^{+1.0/-0.9}$	

Note:
[1] Please note that GPT 3.5 required no training whatsoever, so it represents a significant cost saving over the other approaches despite the apparently lower performance.

Limitations

- The Embedding method needs 100 to 200 in the training dataset to accurately perform classifications for abstract inclusion and exclusion
- The GPT-3.5 method can be an iterative process, as questions about the inclusion and exclusion criteria prompted for each abstract can take time to generate

References

1. Cohen, A. M., Ambert, K., & McDonagh, M. (2009). Cross-Topic Learning for Work Prioritization in Systematic Review Creation and Update. *Journal of the American Medical Informatics Association*, 690–704.
2. MiniLM citation: https://arxiv.org/abs/2002.10957
3. GPT-3.5 citation: https://platform.openai.com/docs/model-index-for-researchers

Presented at ISPOR Europe 2023, Copenhagen, Denmark, 12–15 November 2023

ANALYSIS GROUP
ECONOMIC, FINANCIAL and STRATEGY CONSULTANTS