LASERAI

Fast, Accurate, Reliable: Prospective Follow-up Assessment of the SuperDeduper Module in Laser AI

Ewelina Sadowska, Joanna Konieczna, Ewa Borowiack, Monika Opalek, Artur Nowak

Evidence Prime, Krakow, Poland

Background

When conducting a systematic review, searching multiple bibliographic databases is essential to ensure the comprehensiveness of the retrieved evidence and to reduce the risk of missing eligible studies [1]. The PRISMA-Search extension to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement emphasizes that no single database can provide a complete set of studies meeting systematic review eligibility criteria, thus is it advised to search multiple databases and supplementary sources (e.g. clinical trial registries, preprint servers) to maximize recall of relevant studies [2]. Consequently, removing duplicate records ("deduplication") is an integral step in evidence synthesis workflow, preventing reviewers from screening the same reference multiple times.

Importantly, poorly executed deduplication can pose a significant risk, as relevant, unique studies may be erroneously treated as duplicates and removed from the reference list. The Cochrane Handbook [3] provides detailed guidance on identifying duplicate records, however, large and heterogeneous reference sets with high number of duplicate and near-duplicate records makes fully manual verification time-consuming and prone to errors. Therefore, it is common practice to use automated tools for deduplication.

Aim

To prospectively evaluate the performance of the rule-based algorithm used in the SuperDeduper deduplication module within Laser AI [Fig1]. This work extends our earlier retrospective assessment of SuperDeduper by providing a comprehensive evaluation of its performance [5].

Methods

Validation Dataset

The Gold-Standard Annotated Benchmark Dataset (in press) was developed to validate SuperDeduper tool performance. The dataset consists of three subsets that correspond to three seed systematic reviews which were used as reference sources. This reflects the outputs of large-scale, multi-database searches often encountered in real world systematic review projects. In our dataset, the number of references ranges from 3 645 to 31 740 per seed systematic review, covering from 7 to 10 databases. The dataset was manually deduplicated by two independent researchers, with additional true duplicates found during tool validation.

SuperDeduper

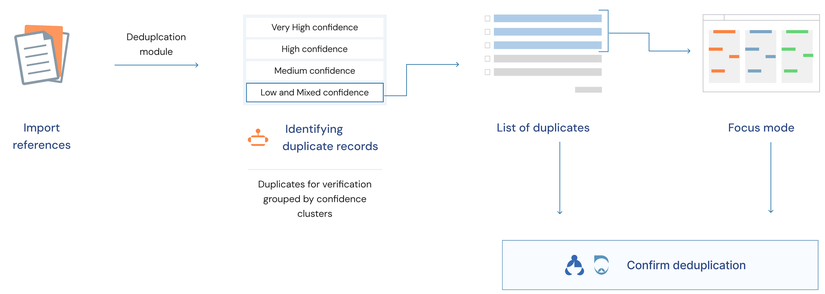
Development and functionality

We adopted and adapted the algorithm proposed by Forbes et al 2024 [4], which provides a robust framework for deduplication using predefined rule sets and string-matching techniques. It operates on metadata fields from RIS files, in particular: title, DOI, volume, authors, journal, pages and abstract. Our implementation of the Forbes algorithm involves several key operations:

- Field Preprocessing: Textual data in specified fields are normalized using "mutators", includes stages such as standardization of author names (e.g., resolving variations in initials, name order, and affiliation strings using complex regular expressions) and normalization of publication details such as page numbers and DOIs.
- Blocking/Grouping: Records are sorted and grouped by specific fields to limit pairwise comparisons to likely candidates.
- Pairwise Similarity Calculation: Within each group, similarity scores are computed for record pairs.

Records are first sorted and grouped by specific metadata fields and then pairwise similarity is calculated, i.e. within each group, similarity scores are computed for record pairs. This procedure is iterated across different field combinations and averaged to produce a final similarity score for each potential duplicate pair. A pair is flagged as a duplicate when its aggregated score exceeds a predefined threshold.

SuperDeduper Framework Fig 1



Confidence Band Stratification

Groups of records, i.e. potential duplicates, are assigned to one of four confidence buckets, representing the similarity of the weakest link within that cluster (Very High, High, Medium and Low/Mix Confidence) [Fig 2]. This overall score corresponds to the probability of a given group being a duplicate.

Fig 2

The thresholds used for this categorization

Very High Confidence:	Score ≥ 0.9
High Confidence	Score ≥ 0.7 and < 0.9
Medium Confidence	Score ≥ 0.4 and < 0.7
Low & Mix Confidence	Score ≥ 0.01 and < 0.4

Results

The total size of the benchmark dataset contained 40574 records. SuperDeduper identified the vast majority of duplicates pointed out by a human, and even pointed out a few [4] missed duplicates, simultaneously only 12 records were wrongly marked as duplicates (false positives). Average accuracy on all benchmark sets was 99.4%, average sensitivity 98.2%, average specificity over 99.9%. None of the false positive records were included in source systematic reviews and they were not automatically removed by the tool, but rather marked as duplicates of low confidence indicating necessity of human oversight.

MSR26

SuperDeduper performance metrics Table 1

Systematic Review ID	Retrieved References	False Negative	False Positive	True Negative	True Positive	Specificity	Sensitivity	Accuracy
Padilha 2018	31,748	184	11	18924	12621	0.999	0.986	0.994
Penington 2018	5,180	18	o	4190	972	1.000	0.982	0.997
Rosa 2018	3,646	25	1	2501	1118	1.000	0.978	0.993
					average	1.000	0.982	0.994

SuperDeduper performance metrics across different confidence bands Table 2

	Low Confidence				Medium Confidence				High Confidence				Very High Confidence				Unique references (One-element cluster)	
Padilha 2018	FN	FP	TN	TP	FN	FP	TN	TP	FN	FP	TN	TP	FN	FP	TN	TP	FN	TN
	15	11	784	1638	16	0	1529	3456	9	0	1624	2960	6	0	2675	4567	138	12248
	7.71%				15.76%				14.47%				22.84%					
	specificity			0.986	specificity			1.000	specificity			1.000	specificity			1.000		
	sensitivity			0.991	sensitivity			0.995	sensitivity			0.997	sensitivity			0.999		
	accuracy			0.989	accuracy			0.987	accuracy			0.998	accuracy			0.999		
Penington 2018	FN	FP	TN	TP	FN	FP	TN	TP	FN	FP	TN	TP	FN	FP	TN	TP	FN	TN
	0	0	83	153	1	0	135	283	1	0	197	306	0	0	160	230	16	3562
	4.56%				8.09%				9.73%				7.53%					
	specificity			1.000	specificity			1.000	specificity			1.000	specificity			1.000		
	sensitivity			1.000	sensitivity			0.996	sensitivity			0.997	sensitivity			1.000		
	accuracy			1.000	accuracy			0.998	accuracy			0.989	accuracy			1.000		
Rosa 2018	FN	FP	TN	TP	FN	FP	TN	TP	FN	FP	TN	TP	FN	FP	TN	TP	FN	TN
	1	1	87	188	2	0	171	382	3	0	172	343	0	0	115	205	19	1956
	7.60%				15.23%				14.21%				8.78%					
	specificity			0.989	specificity			1.000	specificity			1.000	specificity			1.000		
	sensitivity			0.995	sensitivity			0.995	sensitivity			0.991	sensitivity			1.000		
	accuracy			0.983	accuracy			0.996	accuracy			0.994	accuracy			1.000		

Conclusions

The SuperDeduper module aligns with the human-in-the-loop approach, classifying potential duplicate clusters into several confidence bands, providing users with actionable insights for their deduplication process. The results confirm that it is an effective and safe method to remove duplicates. According to our validation study (in press), only the "Low and Mixed Confidence" bucket requires manual verification, whereas duplicate clusters assigned to the remaining confidence categories can be accepted directly. The tool may speed up the process of review by reducing human burden and the time spent on verifying duplicates.

References

[1] Ewald H, Klerings I, Wagner G, Heise TL, Stratil JM, Lhachimi SK, Hemkens LG, Gartlehner G, Armijo-Olivo S, Nussbaumer-Streit B. Searching two or more databases decreased the risk of missing relevant studies: a metaresearch study. J Clin Epidemiol. 2022 Sep;149:154-164. doi: 10.1016/j.jclinepi.2022.05.022. Epub 2022 May 30. PMID: 35654269.

[2] Rethlefsen, M.L., Kirtley, S., Waffenschmidt, S. et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 10, 39 (2021). https://doi.org/10.1186/s13643-020-01542-z

[3] https://training.cochrane.org/handbook/current/chapter-04

[4] Forbes, Connor & Greenwood, Hannah & Carter, Matt & Clark, Justin. (2024). Automation of duplicate record detection for systematic reviews: Deduplicator. Systematic Reviews. 13. 10.1186/s13643-024-02619-9.

[5] Nowak A, Sadowska E, Borowiack E. Superdeduper: Testing New AI-Powered System for Deduplicating References in Literature Reviewst. Value in Health, Volume 27, Issue 12, S2 (December 2024)