SuperDeduper: Testing the New AI-Powered System for Deduplicating References in Literature Reviews

Artur Nowak, Ewelina Sadowska, Ewa Borowiack
Evidence Prime, Krakow, Poland

Key points

SuperDeduper is a Laser AI module that performs AI-assisted deduplication of references.

SuperDeduper achieved 100% specificity (0 lost references) as the only one among the general-purpose deduplication tools.

It also achieved 98.05% accuracy, which is second only to SRA Deduplicator (98.53%).

Our results indicate that the SuperDeduper module is a reliable and efficient tool for removing duplicates without requiring human supervision.

The publicly available dataset has limitations. Therefore, as part of the planned prospective validation, we will create a more comprehensive dataset, which we intend to publish.

Introduction

Removing duplicate records is a crucial step while conducting systematic reviews, including cost and cost-effectiveness outcomes. It's often a long and laborious process, but properly performed deduplication prevents researchers from reviewing the same references from different databases, reducing the time spent on screening. At the same time, errors in deduplication pose a risk of losing relevant studies before screening even starts.

This problem is addressed by some of the search engines, which allow deduplication of the results at the search stage (e.g., Ovid). However, this approach has limitations since it cannot be applied to all databases.

Deduplication tools can be categorized according to the level of automation and artificial intelligence (AI) support that guides them. They range from manual comparison of key fields in reference management softwares, like Endnote, to tools that manage reviews where deduplication is supported by AI models (e.g., Deduklick). One of the tools that use AI models to remove duplicate references is SuperDeduper which combines classic approaches of comparing fields as well as delegating checks of close matches to AI (semantic similarity). SuperDeduper was built to provide 100% specificity (no references erroneously classified as duplicates) while maximizing accuracy and sensitivity (to minimize unnecessary manual work of reviewing duplicates).

Objective

Our aim was to retrospectively evaluate the accuracy of SuperDeduper, using publicly available benchmark datasets that compared multiple existing tools.

Method

We looked for gold-standard datasets that were used to evaluate automatic deduplication processes in at least one reference management or literature review software. To identify existing datasets that could potentially be used to validate SuperDeduper, we conducted a rapid review of validation studies for other deduplication tools. Our search yielded seven relevant articles [1–7]. Of these, three provided access to the datasets used in their validation processes, which could serve as potential resources for validation [1, 5, 7]. We used dataset [1,7] which has been used to validate nearly all available deduplication tools. This approach allows us to directly compare the performance of SuperDeduper with other existing tools, providing a robust benchmark for evaluating its effectiveness in the deduplication process. The benchmark set of deduplicated references was created manually in an Excel sheet based on 3,130 records identified through searches in MEDLINE, Embase, PsycINFO, and the Cochrane Central Register of Controlled Trials (Cochrane CENTRAL).

We ran SuperDeduper on the same dataset to compare the number of false positives (number of unique references erroneously marked as duplicates) and false negatives (duplicate references that were missed by the method).

SuperDeduper is designed to resemble the traditional approach of applying sets of rules for matching bibliographic metadata [8,9], but uses AI to review the potential duplicate pairs (the work that is typically done by humans). This hybrid approach (Fig. 2) has multiple advantages. First, it provides more transparency because it automates the existing best practices instead of replacing them with a black box. Second, it allows customization of the deduplication algorithm, especially if users have a tried-and-tested approach for the set of databases in which they search. Third, it splits the potential duplicate pairs into multiple confidence bands, allowing users to decide which tasks to delegate to AI and which to check manually.

Figure 1. shows the user interface of the Laser AI tool that allows for the review and correcting of SuperDeduper's decisions.
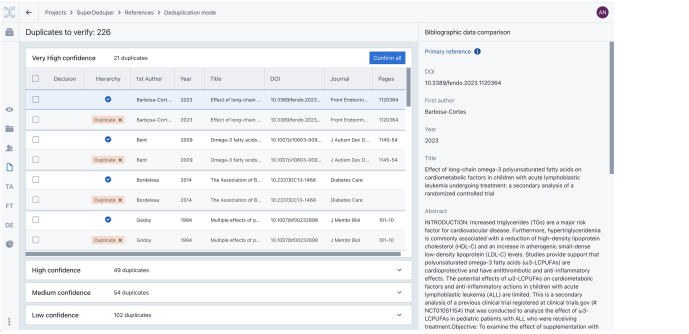

Figure 1: User interface of the deduplication module in Laser AI

Results

We compared the results of SuperDeduper with the following 12 tools evaluated on the McKeown dataset [1, 7]: Ovid Multifile search, review softwares: Covidence, Rayyan, Deduklick, Systematic Review Accelerator's Deduplicator and reference management softwares: EndNote desktop X9, Manual EndNote online classic, EndNote 20, Mendeley Desktop, Zotero, ProQuest RefWorks (similar and exact match). On the same sample of references, SuperDeduper identified 1177 duplicates. There were no false positives (compared to a range of 0–208 false positives in other tools) and 61 false negatives (with other tools showing a range of 35–718 false negatives) [Table 1]. The calculated sensitivity was 95%, ranking third out of 13 tools, while the specificity was 100%.

Among the solutions that achieved 100% specificity, SuperDeduper ranked first in terms of sensitivity. The accuracy was 98%, placing SuperDeduper second among all deduplication tools.
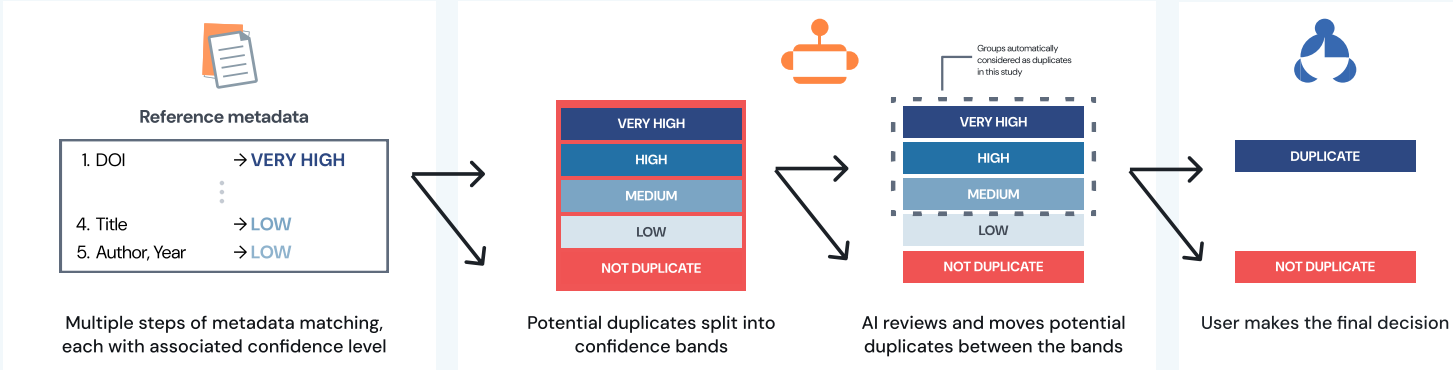

Figure 2: Diagram explaining the methodology

	Systematic review tools					Reference management and other tools								
	SuperDeduper	**Deduklick**	**Systematic Review Accelerator**	**Rayyan**	**Covidence**	**EndNote desktop X9**	**Manual EndNote online classic**	**EndNote 20**	**Mendeley**	**Zotero**	**RefWorks (Similar)**	**RefWorks (Exact)**	**Ovid Multifile search**
Accuracy	98.0%	96.5%	98.5%	96.8%	96.1%	76.4%	77.0%	91.9%	93.2%	80.2%	94.3%	80.2%	97.1%
Sensitivity	95.0%	91.2%	97.2%	96.0%	90.3%	57.1%	42.0%	79.9%	84.3%	51.6%	86.4%	50.2%	92.7%
Specificity	100.0%	99.9%	99.4%	97.3%	99.9%	89.0%	99.8%	99.8%	99.1%	98.9%	99.5%	99.8%	100.0%
False positives	0	2	11	52	2	208	3	4	17	20	9	4	0
False negatives	61	109	35	49	120	531	718	249	195	599	168	617	90

Tab 1 SuperDeduper performance results versus other deduplication tools/methods [7]

Conclusion

Our results suggest that the SuperDeduper module is a time-saving and effective method to remove duplicates while minimizing human supervision.

The limitation of this experiment is that we used only one dataset to validate the efficacy. However, most identified studies did not make their datasets publicly available or used proprietary data. These findings underscore the need for accessible, standardized datasets to support robust validation efforts in the deduplication space, and they highlight the potential value of sharing datasets for future research and development.

Using the chosen dataset allows for direct comparison with a large number of tools that support the systematic review process. However, this dataset has some limitations, as mentioned in articles [1,7]. It was created from searches conducted on a single platform, Ovid, and dates back to 2018, before databases like Embase and Cochrane Central began including unpublished sources such as clinical trial registries.

In the next step, we will prospectively validate SuperDeduper using more comprehensive datasets. These datasets will cover a larger number of databases and platforms—at least seven per dataset. They will span/cover various clinical areas and research topics, ensuring that the validation process reflects diverse scientific domains. This diversity in sources and topics will provide a robust foundation for assessing SuperDeduper's effectiveness across different contexts and data sources. We are planning to make the deduplication datasets publicly available.

References

[1] McKeown, S., Mir, Z.M. Considerations for conducting systematic reviews: evaluating the performance of different methods for de-duplicating references. Syst Rev 10, 38 (2021). https://doi.org/10.1186/s13643-021-01583-y

[2] Borissov N, Haas Q, Minder B, Kopp-Heim D, von Gernler M, Janka H, Teodoro D, Amini P. Reducing systematic review burden using Deduklick: a novel, automated, reliable, and explainable deduplication algorithm to foster medical research. Syst Rev. 2022 Aug 17;11(1):172. doi: 10.1186/s13643-022-02045-9. PMID: 35978441; PMCID: PMC9382798.

[3] Guimarães NS, Ferreira AJF, Ribeiro Silva RC, de Paula AA, Lisboa CS, Magno L, Ichiara MY, Barreto ML. Deduplicating records in systematic reviews: there are free, accurate automated ways to do so. J Clin Epidemiol. 2022 Dec;152:110–115. doi: 10.1016/j.jclinepi.2022.10.009. Epub 2022 Oct 12. PMID: 36241035.

[4] Rathbone J, Carter M, Hoffmann T, Glasziou P. Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module. Syst Rev. 2015 Jan 14;4(1):6. doi: 10.1186/2046-4053-4-6. PMID: 25588387; PMCID: PMC4320616.

[5] Hair K, Bahor Z, Macleod M, Liao J, Sena ES. The Automated Systematic Search Deduplicator (ASySD): a rapid, open-source, interoperable tool to remove duplicate citations in biomedical systematic reviews. BMC Biol. 2023 Sep 7;21(1):189. doi: 10.1186/s12915-023-01686-z. PMID: 37674179; PMCID: PMC10483700.

[6] Janka H, Metzendorf M-I. High precision but variable recall – comparing the performance of five deduplication tools. J Eur Assoc Health Info Libr [Internet]. 2024 Mar. 17 [cited 2024 Oct. 11];20(1):12–7. Available from: https://ojs.eahil.eu/JEAHIL/article/view/607

[7] McKeown S, Mir ZM. Considerations for conducting systematic reviews: A follow-up study to evaluate the performance of various automated methods for reference de-duplication. Res Synth Methods. 2024 Jul 25. doi: 10.1002/jrsm.1736. Epub ahead of print. PMID: 39051574.

[8] Bramer WM, Giustini D, de Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. J Med Libr Assoc. 2016 Jul;104(3):240–3. doi: 10.3163/1536-5050.104.3.014. Erratum in: J Med Libr Assoc. 2017 Jan;105(1):111. doi: 10.5195/jmla.2017.128. PMID: 27366130; PMCID: PMC4915647.

[9] University of Leeds. Duplicate Checking Guidance. Information Specialists, University of Leeds, Mar. 2019, https://information-specialists.leeds.ac.uk/wp-content/uploads/sites/71/2019/03/Duplicate_checking_guidance.pdf. (online access 8/11/2024)