Assessing the Performance of ASReview for Abstract Screening in Systematic Reviews in Education: A Benchmarking and Evaluation Project

Diego G. Campos[1], Tim Fütterer[2], Thomas Gfrörer[2], Rosa Lavelle-Hill[2], Kou Murayama[2], Lars König[3], Martin Hecht[3], Steffen Zitzmann[2], and Ronny Scherer[1]

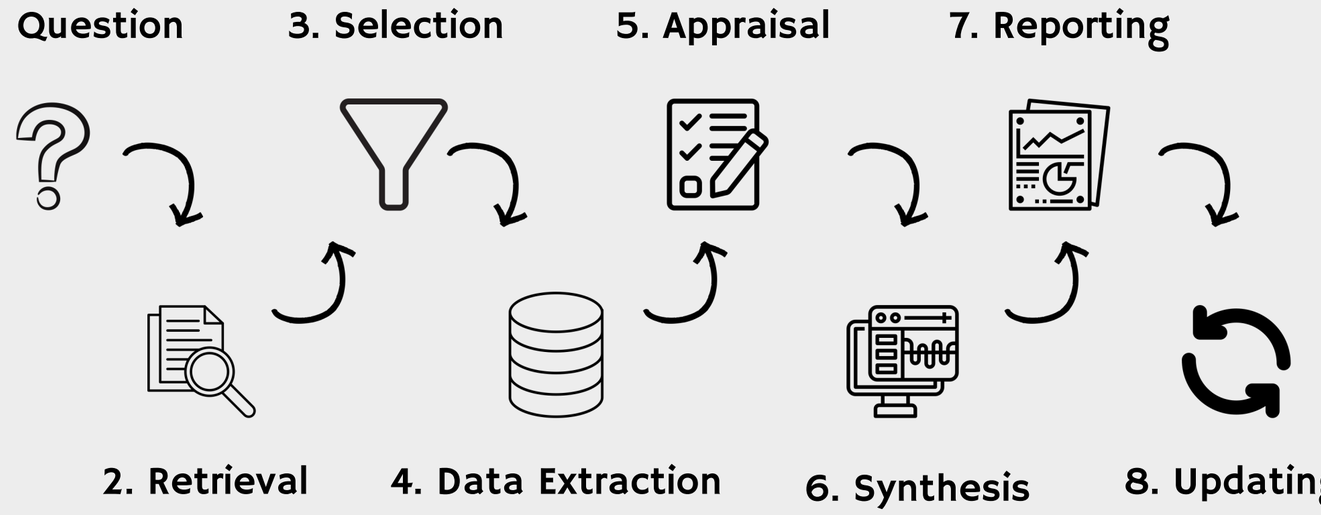
[1] University of Oslo, Norway; [2] University of Tübingen, Germany; [3]Helmut-Schmidt-Universität, Germany

LEAD Intramural Research Grant Application, University of Tübingen

Introduction

- Systematic reviews and meta-analyses are critical for advancing practice, research, and policy (e.g., Schneider & Preckel, 2017; van de Schoot et al., 2021; van Huizen & Plantenga, 2018).
- Understand teaching and learning processes and to derive measures for an effective educational system (Taylor & Hedges, 2023).
- What works clearinghouses (Hedges, 2018).

Systematic Review

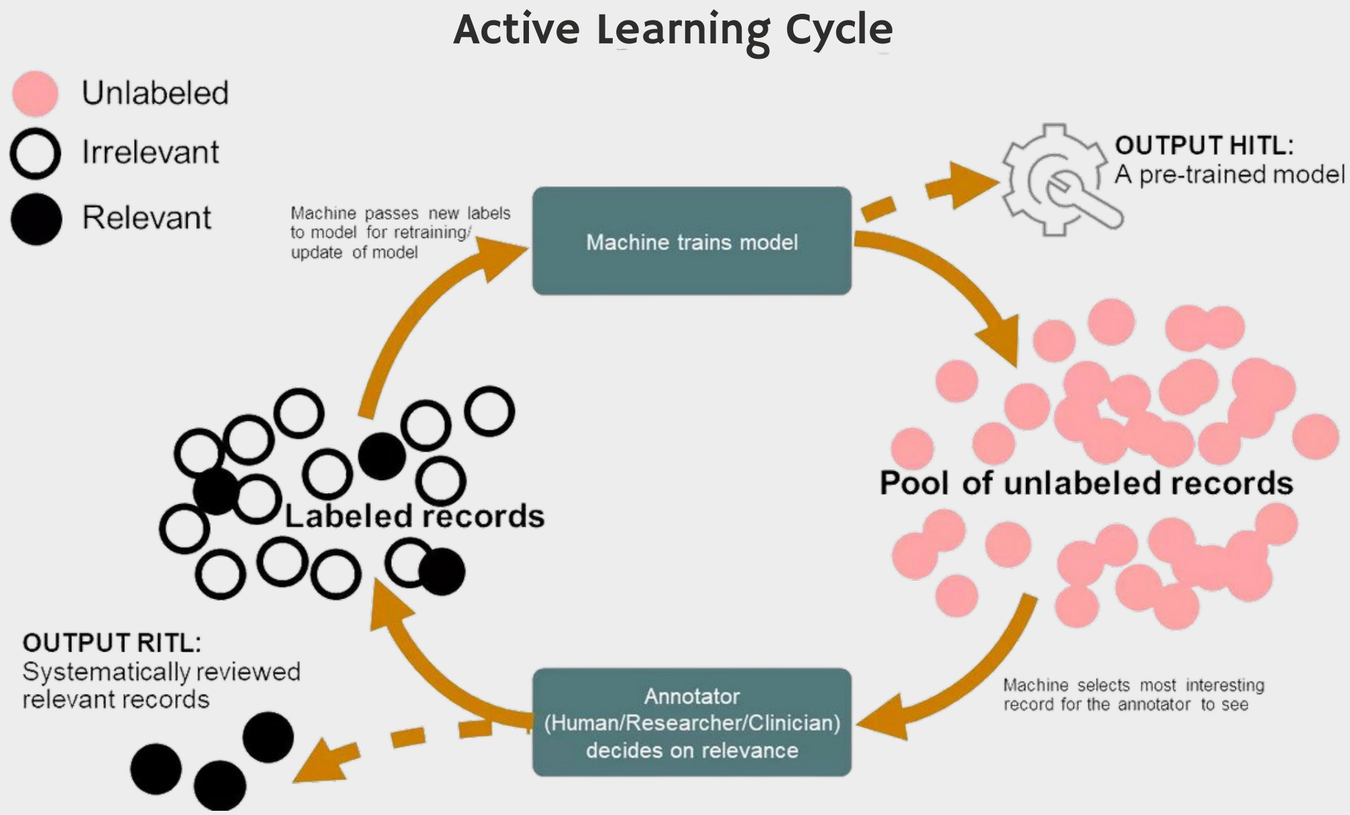


- **John Hattie and his team needed 15 years** to complete their meta-analysis “visible learning” (Hattie, 2009).
- An experienced reviewer can take between **30 seconds and 1 minute per reference** (Przybyła, et al. 2017).

Machine Learning and Literature Screening

ASReview can help researchers to find relevant records by using **screening prioritization** with **active learning.**

Screening prioritization re-arranges the records to be screened from **random** to a more **intelligent order**



From: https://asreview.nl/blog/active-learning-explained/

Integrating AI tools into the Systematic Review

- Simulation studies in health sciences have shown that the **AI screening process within ASReview is reliable and can save more than 90% of the time** compared to purely human-based screening (van de Schoot et al., 2021).
- The **performance of AI tools varies** across research domains (van de Schoot et al., 2021).
- Clear **stopping rules need to be defined** (Yu & Menzies, 2019).

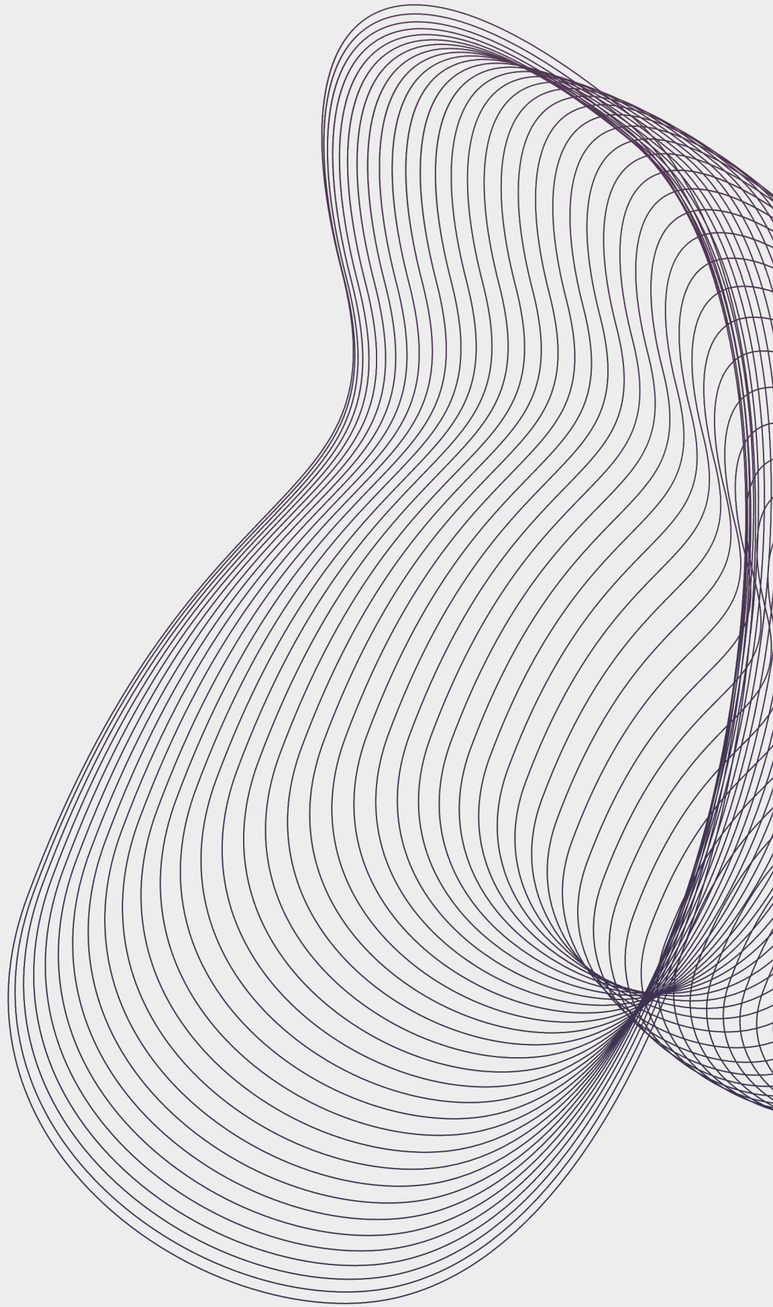
Stopping Rules

Statistic Approaches

The decision to stop screening is based on deriving an estimate of the total number of relevant papers within the initial training set (e.g.; van Haastrecht et al., 2021).

Heuristic Approaches

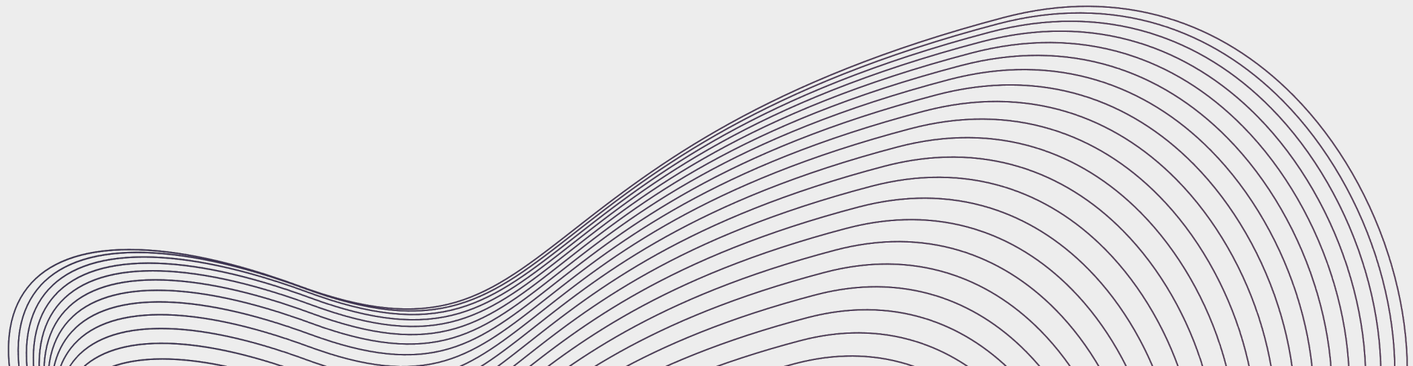
- Data-driven (Ros et al., 2017).
- Time-based (Wallace et al., 2010).
- Mixed-data (Hamel et al., 2021).

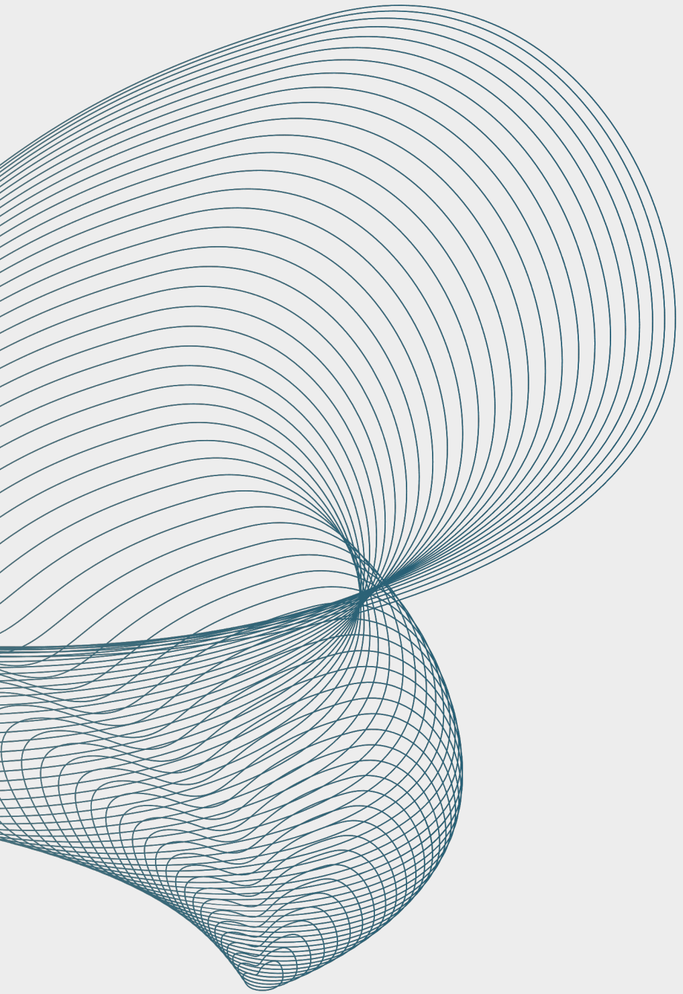


Research questions

1. How do AI-based NLP algorithms perform with respect to WSS and ETS on educational data sets?

2. How do heuristic stopping criteria perform with respect to recall, WSS, and ETS on educational data sets?





Methodology

- The study used **a systematic approach to compare the accuracy, speed, and cost-effectiveness of AI tools for abstract screening in education.** Data were collected from multiple academic databases and analyzed using the ASReview software and R.

Sample

- Journals in education and educational psychology.
  - Campbell Reviews (n = 17).
  - Review of Educational Research (n = 95).
  - Educational Research Review (n = 83).
  - Educational Psychology Review (n = 37).
- Experts in the field (n = 84).
- Total records sought for retrieval (n = 316).
- Reports excluded (n = 12).
- Total reports included (n = 28).

Sample

Descriptive statistics of databases included in the simulation study.

Database	Total References	References Selected for Full Text Screening	Percentage of References Selected for Full Text Screening
Anmarkrud, et al., (2022)	2550	230	8.5 %
Backfisch, et al., (2020)	2353	719	31.2 %
Capparozza, et al., (2021)	2739	171	6.5 %
Endedijk, et al., (2022)	4452	365	8.5 %
Filges, et al., (2018)	8065	374	4.1 %
Filges, et al., (2022)	16221	6429	9.8 %
Fong, et al., (2021)	2471	199	8.7 %
Fütterer, et al., (2023)	19497	44	0.2 %
Jäger-Dengler-Harles, et al., (2020)	6798	1212	20.2 %
Kupers, et al., (2019)	3094	962	31.1%
Lesperance, et al., (2022)	8465	149	1.8 %
Neri & Retelsdorf, (2022)	686	115	16.8 %
Noetel, et al., (2022)	1368	86	5.8 %
Pico & Woods, (2023)	30	14	46.7 %
Roberts, et al., (2022)	5065	168	3.3 %
Rowan, et al., (2021)	416	209	50.1 %
Saqr, et al., (2022)	4581	1818	39.5 %

Method

- ASReview (van de Schoot et al., 2021)
- R (R Core Team, 2022)

Feature Extraction Technique	Classifier	Model
Doc2Vec	Naïve Bayes (NB)	LR + Doc2Vec
Sentence BERT (SBERT)	Logistic Regression (LR)	LR + SBERT
Term Frequency-Inverse Abstract Frequency (TF-IDF)	Random Forest (RF)	LR + TFIDF
	Support Vector Machine (SVM)	NB + TF-IDF
		RF + Doc2Vec
		RF + SBERT
		RF + TF-IDF
		SVM + Doc2Vec
		SVM + SBERT
		SVM + TF-IDF

Model Evaluation

Work Saved over Sampling (WSS@95%)

$$WSS@95\% = 0.95 - \frac{TP(i_{R95}) + FP(i_{R95})}{N}$$

Przybyła et al., (2018)

Estimated time savings (ETS)

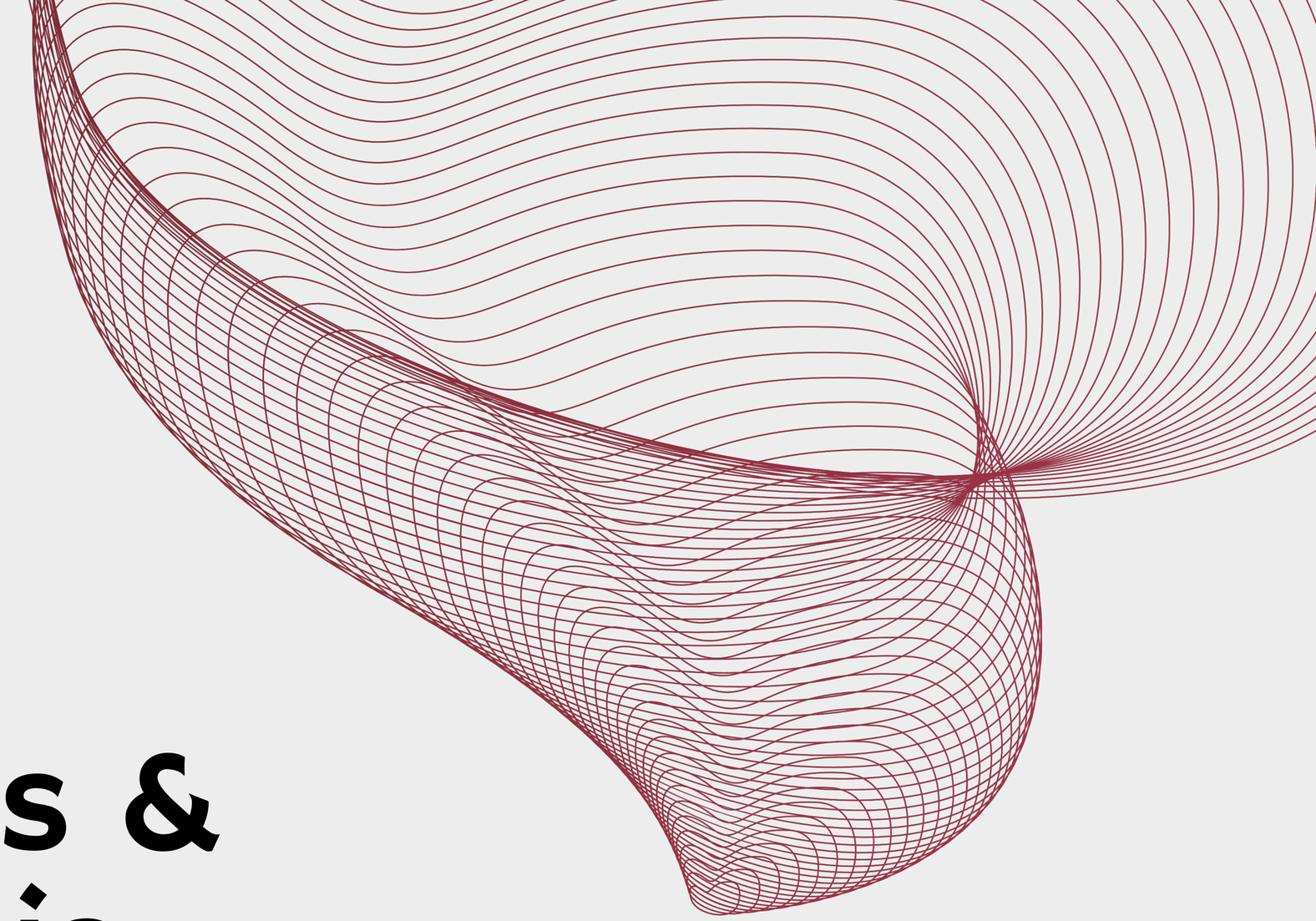
$$ETS = \frac{RN\ (i) * 0.5}{8}$$

Gates et al., (2019)

Recall and missed records

$$recall\ (i) = \frac{TP(i)}{TP(i) + FN(i)} \qquad \text{MR (i)} = 1 - \left(\frac{TP(i)}{TP(i) + FN(i)}\right)$$

Przybyła et al., (2018)



Results & Analysis

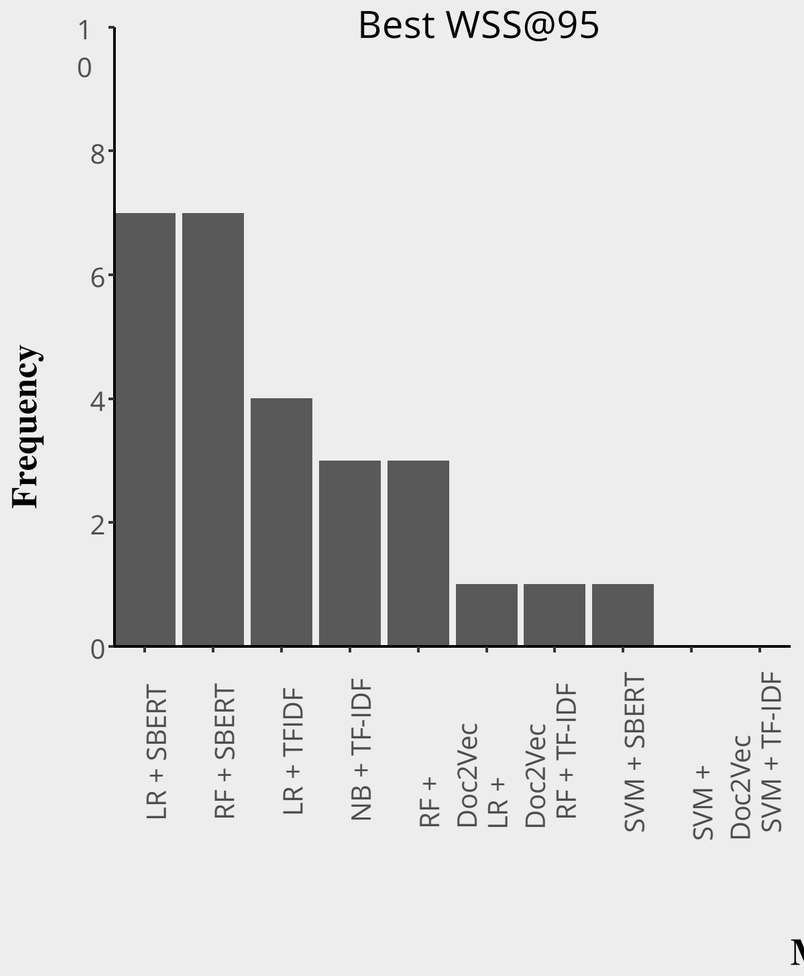
Model Evaluation

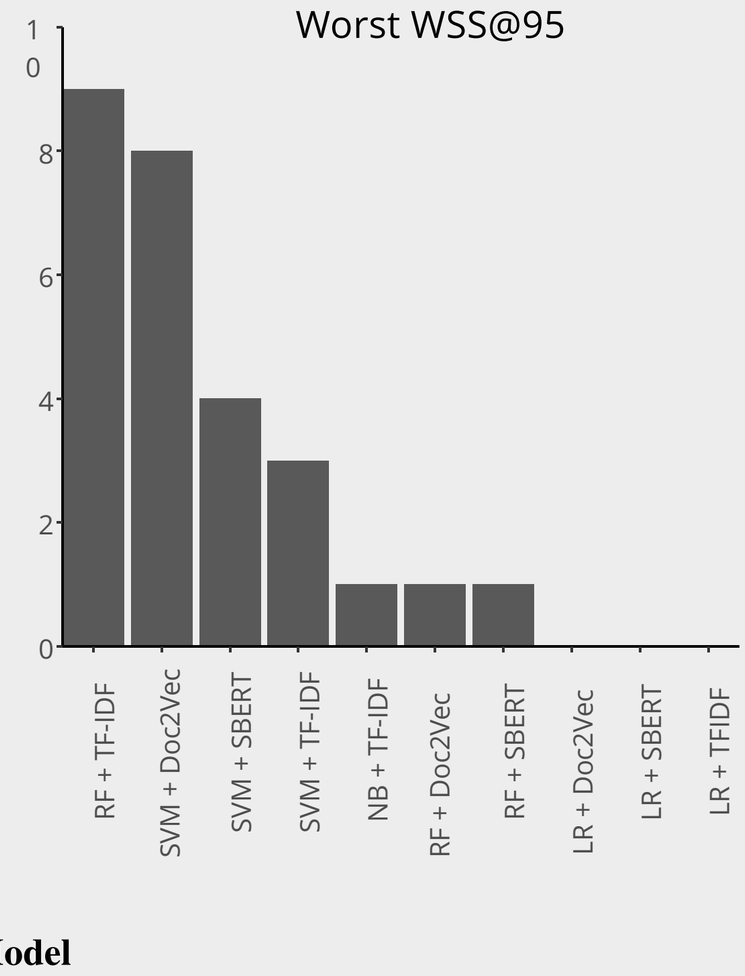
Model	Mean (WSS@95%)	Median (WSS@95%)	SD (WSS@95%)	Mean (ETS@95%)	Median (ETS@95%)	SD (ETS@95%)
LR + Doc2Vec	49,9%	49,7%	20,8%	130	74	214
LR + SBERT	55,9%	57,1%	21,4%	140	77	221
LR + TFIDF	52,7%	55,9%	22,0%	137	77	224
NB + TF-IDF	49,7%	51,2%	22,6%	129	65	223
RF + Doc2Vec	49,4%	52,4%	19,9%	117	72	152
RF + SBERT	54,2%	59,4%	21,5%	137	83	220
RF + TF-IDF	47,9%	44,2%	20,7%	125	68	205
SVM + Doc2Vec	46,7%	43,0%	21,2%	126	67	211
SVM + SBERT	51,1%	50,5%	21,0%	131	58	203
SVM + TF-IDF	49,6%	53,0%	21,6%	130	73	221

Model Evaluation

- The performance of classifiers and feature extraction techniques are similar for educational datasets.

- The LR + SBERT and RF + SBERT models were more frequently identified as the best performing models and produced the highest WSS@95%.

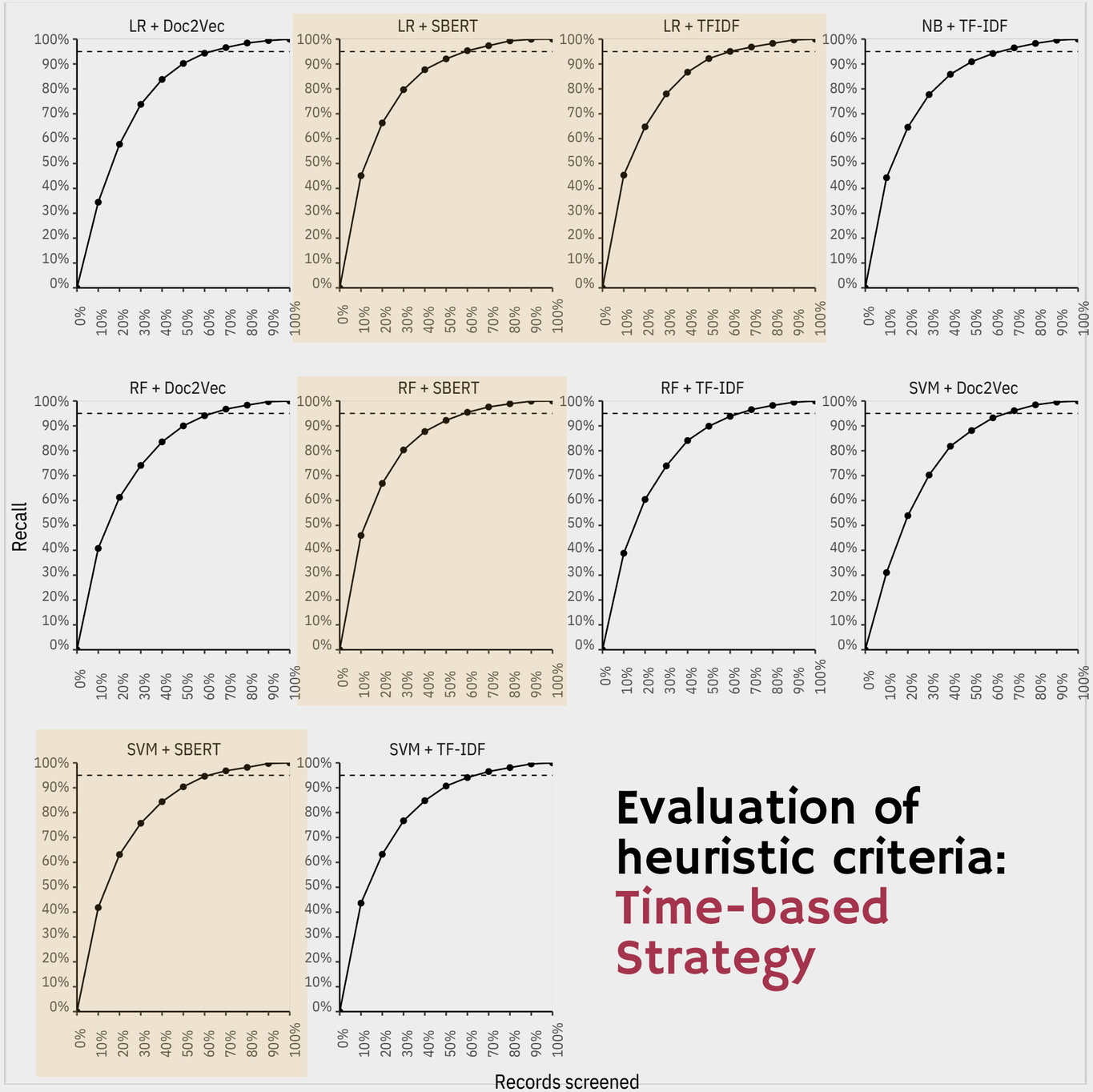




Evaluation of heuristic criteria:
Time-based Strategy

Percentage reviewed	Recall (Mean)	Recall (SD)	Missed Records (Mean)	Missed Records (SD)	ETS (Mean)	ETS (SD)
10%	41%	5%	59%	5%	185	0
20%	62%	4%	38%	4%	164	0
30%	76%	3%	24%	3%	144	0
40%	85%	2%	15%	2%	123	0
50%	91%	1%	9%	1%	103	0
60%	94%	1%	6%	1%	82	0
70%	97%	0%	3%	0%	62	0
80%	98%	0%	2%	0%	41	0
90%	100%	0%	0%	0%	21	0
100%	100%	0%	0%	0%	0	0

- The time required to find 95% of all relevant records showed minimal differences between the models.
- The performance of the stopping criteria varied greatly between the databases.
- The number of records you have to screen to find 95% of all relevant records is positively correlated with the number of relevant records in your database (r = 0.81).



Evaluation of heuristic criteria: Time-based Strategy

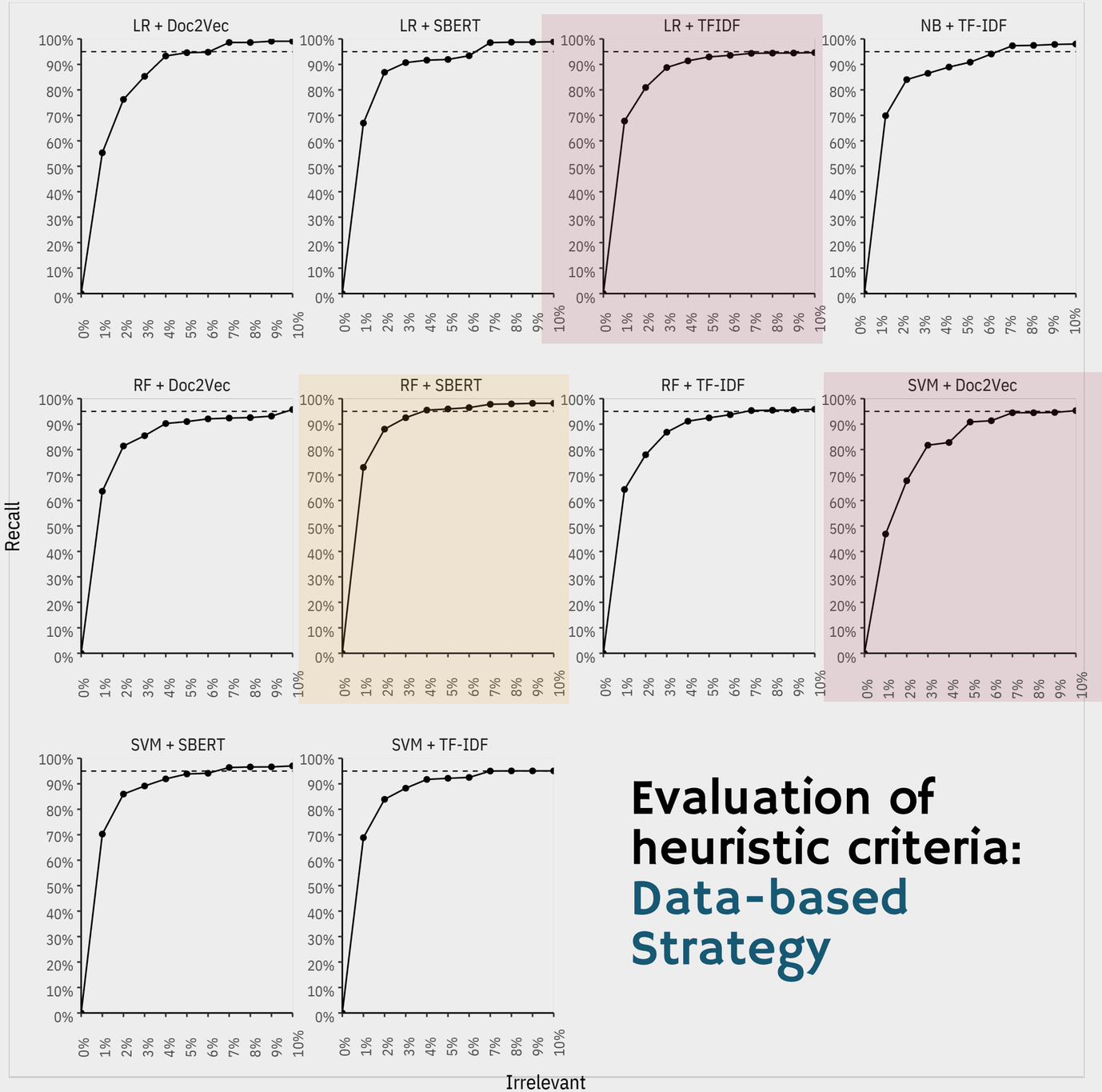
Evaluation of heuristic criteria:
Data-based Strategy

Percentage Irrelevant	Recall (Mean)	Recall (SD)	ETS (Mean)	ETS (SD)	Missed Records (Mean)	Missed Records (SD)	WSS (Mean)	WSS (SD)
1 %	65%	8%	154	2	35%	8%	78%	2%
2 %	81%	6%	133	3	19%	6%	63%	2%
3 %	88%	3%	123	3	12%	3%	55%	2%
4 %	91%	3%	115	4	9%	3%	51%	2%
5 %	93%	2%	110	4	7%	2%	49%	2%
6 %	94%	1%	107	2	6%	1%	46%	2%
7 %	96%	2%	105	3	4%	2%	44%	2%
8 %	96%	2%	103	3	4%	2%	43%	2%
9 %	96%	2%	102	3	4%	2%	43%	2%
10 %	97%	2%	101	3	3%	2%	42%	2%

- There were large variations between different models.

- The performance of the stopping criteria varied greatly between the databases.

  - 1% to 9%.
  - In three of the databases, a cut-down percentage of 10% was insufficient to identify 95% of the records.

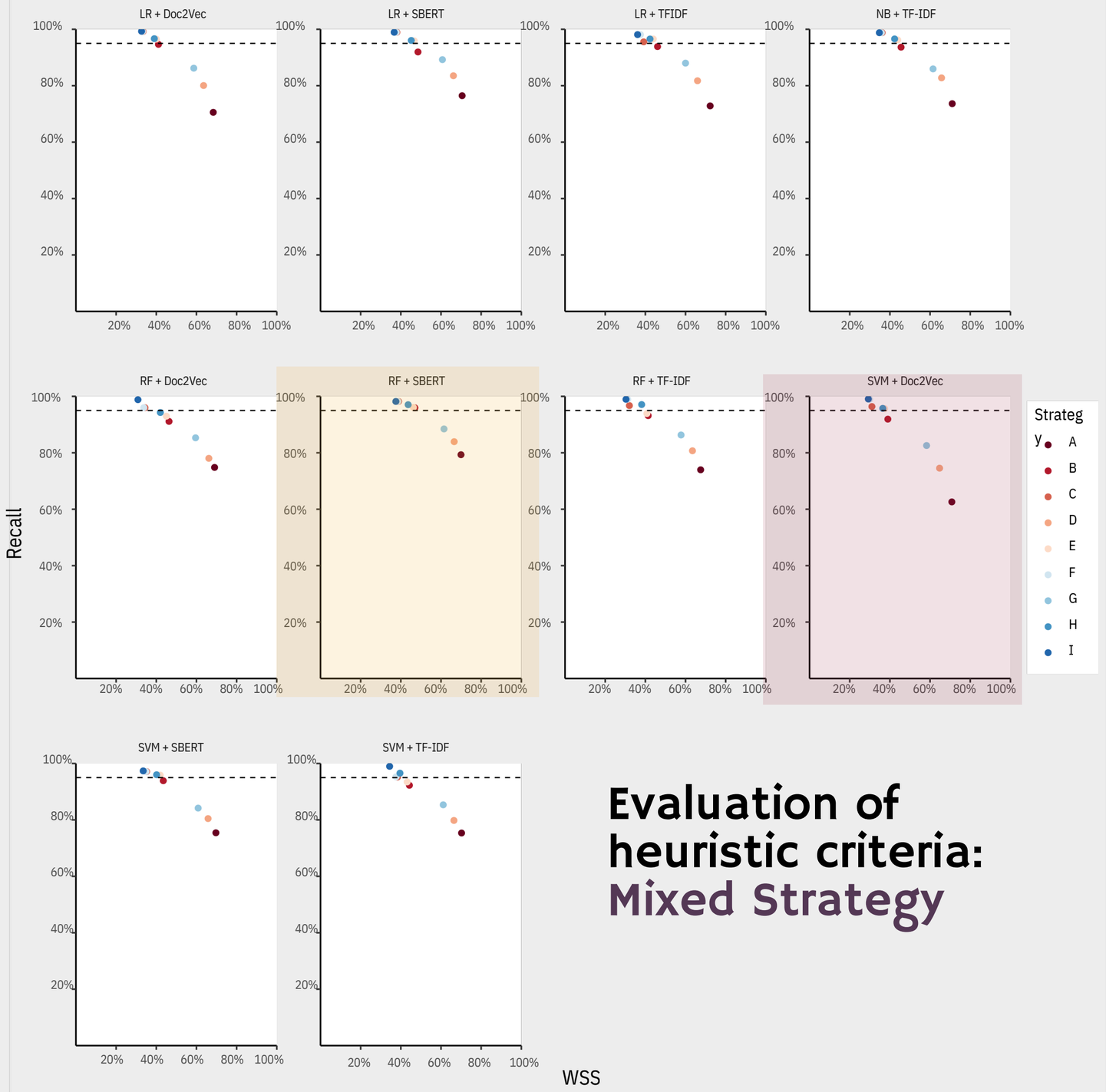


Evaluation of heuristic criteria: Data-based Strategy

Evaluation of heuristic criteria: Mixed Strategy

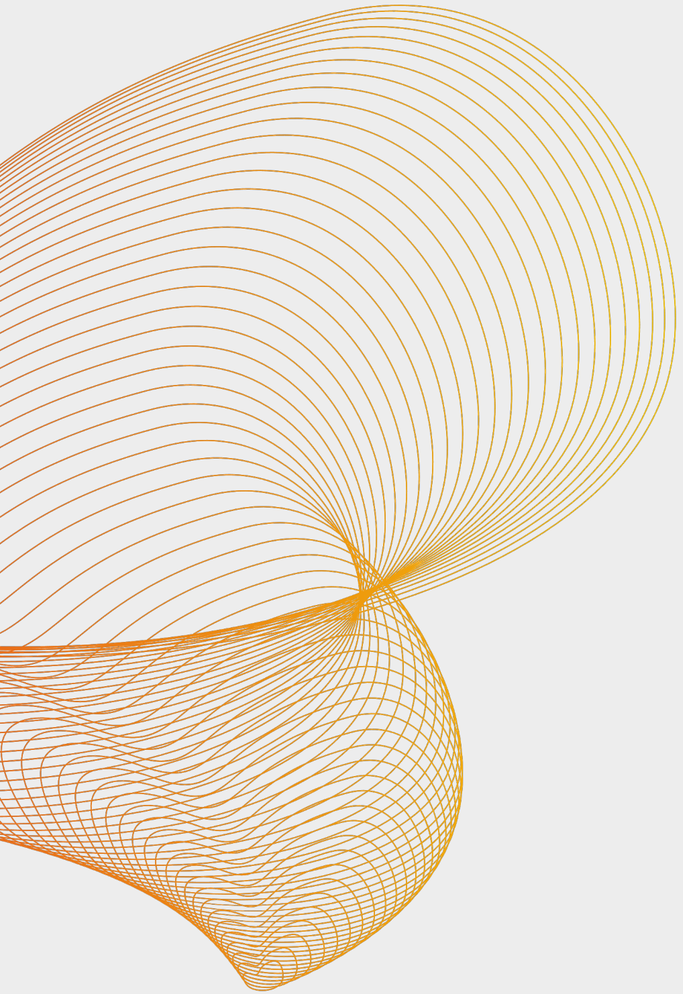
Mixed Strategy	Recall (Mean)	Recall (Median)	Recall (SD)	Missed Records (Mean)	Missed Records (Median)	Missed Records (SD)	WSS (Mean)	WSS (Median)	WSS (SD)
A	73%	83%	25%	27%	17%	25%	70%	81%	21%
B	93%	99%	18%	7%	1%	18%	44%	48%	29%
C	97%	100%	12%	3%	0%	12%	36%	38%	29%
D	81%	89%	19%	19%	11%	19%	65%	75%	18%
E	95%	99%	12%	5%	1%	12%	43%	48%	27%
F	98%	100%	9%	2%	0%	9%	35%	38%	28%
G	86%	91%	15%	14%	9%	15%	60%	68%	14%
H	96%	99%	10%	4%	1%	10%	41%	48%	25%
I	99%	100%	5%	1%	0%	5%	34%	37%	27%

- Performance of heuristic strategies varies across models.
- Performance of the models varied depending on the characteristics of the databases and the screening strategies employed.
    - Positive correlations between recall and the number of records.
    - Negative correlation between recall and the number of relevant records.



Evaluation of heuristic criteria: Mixed Strategy

Conclusions



1. Using **active learning for abstract screening allows educational researchers to find relevant records faster** than random screening.

2. The **performance of the heuristic strategy varied depending on the characteristics of the databases** and the models used.

3. Future studies should **evaluate statistical stopping criterion** to determine its performance.

4. It is **important to examine how the performance of the models and the stopping criteria changes** once more prior information is considered.

References

Chai, K. E. K., Lines, R. L. J., Gucciardi, D. F., & Ng, L. (2021). Research Screener: A machine learning tool to semi-automate abstract screening for systematic reviews. Systematic Reviews, 10(1), 93. https://doi.org/10.1186/s13643-021-01635-3

Ferdinands, G., Schram, R. D., Bruin, J. de, Bagheri, A., Oberski, D. L., Tummers, L., & Schoot, R. van de. (2020). Active learning for screening prioritization in systematic reviews—A simulation study. OSF Preprints. https://doi.org/10.31219/osf.io/w6qbg

Gates, A., Guitard, S., Pillay, J., Elliott, S. A., Dyson, M. P., Newton, A. S., & Hartling, L. (2019). Performance and usability of machine learning for screening in systematic reviews: A comparative evaluation of three tools. Systematic Reviews, 8(1), 278. https://doi.org/10.1186/s13643-019-1222-2

Gates, A., Johnson, C., & Hartling, L. (2018). Technology-assisted title and abstract screening for systematic reviews: A retrospective evaluation of the Abstrackr machine learning tool. Systematic Reviews, 7(1), 45. https://doi.org/10.1186/s13643-018-0707-8

Olofsson, H., Brolund, A., Hellberg, C., Silverstein, R., Stenström, K., Österberg, M., & Dagerhamn, J. (2017). Can abstract screening workload be reduced using text mining? User experiences of the tool Rayyan. Research Synthesis Methods, 8(3), 275–280. https://doi.org/10.1002/jrsm.1237

Przybyła, P., Brockmeier, A. J., Kontonatsios, G., Le Pogam, M.-A., McNaught, J., von Elm, E., Nolan, K., & Ananiadou, S. (2018). Prioritising references for systematic reviews with RobotAnalyst: A user study. Research Synthesis Methods, 9(3), 470–488. https://doi.org/10.1002/jrsm.1311

Teijema, J., Hofstee, L., Brouwer, M., Bruin, J. de, Ferdinands, G., Boer, J. de, Siso, P. V., Brand, S. van den, Bockting, C., Schoot, R. van de, & Bagheri, A. (2022). Active learning-based Systematic reviewing using switching classification models: The case of the onset, maintenance, and relapse of depressive disorders. PsyArXiv. https://doi.org/10.31234/osf.io/t7bpd
van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A.,

Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. Nature Machine Intelligence, 3(2), Article 2. https://doi.org/10.1038/s42256-020-00287-7

Thank You

Diego G. Campos

@diegc_

diegogc@uio.no

www.diegocampos.co