Towards Performance Comparability:

An Implementation of New Metrics into the ASReview Active Learning Screening Prioritization Software for Systematic Literature Reviews

Author: Lesley Spedener[1]

Supervisors: Rens van de Schoot[2], Laura Hofstee[2]

Second examiner: Ayoub Bagheri[2]

[1] Applied Data Science Master's Student Department of Information and Computing Science, Faculty of Science, Utrecht University
[2] Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Utrecht University

Acknowledgements

I would like to thank my supervisors Rens van de Schoot and Laura Hofstee for their enthusiasm, guidance, feedback, and support throughout this project. I would also like to express a special thanks to the ASReview team, specifically Jelle Teijema, Jonathan de Bruin, and Peter Lombaers for their valuable insights.

Abstract

Simulation-based Active Learning (AL) studies have demonstrated the potential of machine learning methods in reducing manual screening workload in systematic literature reviews. The second most used performance metric in this field is Work Saved Over Sampling (WSS), which aims to measure the reduction in screening effort. A drawback of the WSS metric, however, is its sensitivity to dataset class imbalance, which leads to biased performance comparisons across datasets. In this light, two main features were added to the state-of-the-art and open-source simulation software ASReview, which offers a unique infrastructure for testing different AL model and feature extractor combinations across datasets. First, the confusion matrix was implemented into the ASReview software, which was subsequently used to implement the True Negative Rate (TNR), shown to be equal to the normalized WSS (Kusa et al., 2023). These advancements, previously absent in the software, represent a step towards achieving a more comprehensive understanding of AL performance in SLR tasks. Specifically, the adjustment for class imbalance facilitates further study of data characteristics related to model performance beyond class imbalance. This enhanced understanding enables researchers and practitioners to make more informed decisions in selecting and fine tuning AL models, ultimately leading to more efficient screening in practice.

.

Keywords: systematic literature review (SLR), active learning (AL), evaluation, comparability, Work Saved Over Sampling (WSS), True Negative Rate (TNR), specificity

Table of Contents

1. Introduction

To lessen the substantial workload associated with manual systematic literature reviews (SLRs), researchers are working towards the automation of the citation screening task (van Dinter et al., 2021). Specifically, screening prioritization through active learning (AL) can help reviewers save a significant amount of time by stopping the reviewing process once enough relevant articles are found (Yu & Menzies, 2019). In this light, various simulation studies have been conducted to gain a comprehensive understanding of the performance of AL for the SLR task (Teijema et al., 2023). Metrics play a crucial role in driving advancements in machine learning fields as they provide a means to quantitatively compare the effectiveness of different models. By selecting appropriate metrics, researchers can focus on specific aspects of model performance relevant to the task at hand. The SLR task, which typically involves data with a very small proportion of relevant records, is characterized by the aim of maintaining high recall while attempting to reduce workload of manual screening (O'Mara-Eves et al., 2015). In the literature, a wide diversity of metrics exists to measure the extent to which this aim is achieved (Teijema et al., 2023). Work Saved over Sampling (WSS), measured at recall, stands as the second most employed metric to evaluate AL-assisted systematic literature reviews (Teijema et al., 2023). The WSS was introduced as a custom metric for the SLR task to balance high recall and optimal precision. It is defined as "*the percentage of papers that meet the original search criteria that the reviewers do not have to read (because they have been screened out by the classifier)*" (Cohen et al., 2006).

However, studies found a drawback of the WSS measure: its maximum and minimum values depend on the class imbalance of a dataset. For instance, for a dataset with 5% relevant records the maximum value of WSS@95% is 90%, whereas for a perfectly balanced dataset (i.e., equal amount of relevant and irrelevant records), the maximum value of WSS@95% is only 45%

(van Dinter et al., 2021; Kusa et al., 2022). In simpler terms, given the high recall requirement, if a dataset has a higher proportion of relevant records there is less work to be saved. Therefore, if the datasets differ in class imbalance, comparing model performance across datasets using the WSS metric leads to a biased comparison.

The issue described above is relevant for the implementation of AL models for the SLR task, in terms of generalizability of performance across datasets. Comparing models on different datasets helps discover strengths and limitations of the models in relation to (shared or individual) data characteristics. For instance, if consistent differences in performance occur in social science datasets versus medicine datasets, one can investigate the causes of the differences and adapt the model and its parameters to the specific context. However, not adjusting for dataset class imbalance might obscure performance conclusions. To take class imbalance out of the equation, Kusa et al. (2023) propose a min-max normalization of the WSS metric resulting in the normalized WSS (nWSS). In addition, after factorization they find that the nWSS is equal to the standard True Negative Rate (TNR) metric, i.e., specificity.

ASReview is a free and open-source software, which includes unique infrastructure for running large AL simulation studies in a transparent manner (van de Schoot et al., 2021). It has many model configuration options and supports new customized functionalities. Therefore, it offers many opportunities for comparisons across datasets and models, which are necessary to gain a broad understanding of AL performance. However, the TNR is missing in the ASReview Insights Extension v1.1.2 (ASReview LAB Developers, n.d.), which outputs performance metrics.

In light of the context described above, the goal of the present study is to advance performance comparability across datasets and models by first adding a feature to the ASReview Insights Extension (ASReview LAB Developers, n.d.) to output all confusion matrix components, which underlie most metrics and plots. Second, the True Negatives (TN) of the confusion matrix components are then used to add the normalized WSS (Kusa et al., 2023), i.e., True Negative Rate (TNR) to the output, thereby facilitating comparison of AL model performance across datasets adjusting for dataset class imbalance.

2. Method

2.1 ASReview Insights Extension

The ASReview simulation mode is used to evaluate the performance of AL models on fully labelled data (van de Schoot et al., 2021). Since the correct labels are known, a simulation study can mimic the AL screening process by taking on the labelling role of the reviewer. The ASReview Insights Extension (ASReview LAB Developers, n.d.) is responsible for reporting performance metrics and plots after a simulation. It outputs the following metrics: Relevant Record Found (RRF) at % screened, Work Saved Over Sampling (WSS@r%), Extra Relevant Records Found (ERF), Time to Discovery (TD), and Average Time to Discovery (ATD) along with plots such as recall at % of records screened and WSS over recall.

However, some features are missing in the ASReview Insights Extension. Up to version v1.1.2. ASReview Insights does not calculate or output confusion matrix components. The confusion matrix underlies most of the metrics currently calculated in ASReview Insights as well as other metrics not (yet) present (e.g., Recall, Precision, Specificity, F1-score, WSS, Utility etc.). See O’Mara-Eves et al. (2015) for an overview of metrics. The True Negatives of the confusion matrix can be used to calculate the normalised WSS (nWSS), respectively the TNR@r%, which is not yet implemented in ASReview Insights v1.1.2 either.

2.2 Definitions

2.2.1 Confusion matrix

The confusion matrix is composed of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). The confusion matrix components were defined taking inspiration of the description of another screening simulation software (DistillerSR, 2022), see Table 1. Note that these components are designed in line with the certainty-based query strategy, which has been shown to be effective for AL and the SLR task (Miwa et al., 2014).

2.2.2 Work Saved Over Sampling (WSS)

The formula of the WSS metric (Cohen et al., 2006) aims to measure how the percentage of work was saved through the classifier

$$\text{WSS} = \frac{(TN+FN)}{N} - (1-r), \text{ where } r = \frac{TP}{TP+FN} \quad (1)$$

It is typically measured at a 95% recall. Therefore, work saved over sampling at 95% recall is

$$\text{WSS@95\%} = \frac{(TN+FN)}{N} - 0.05 \quad (2)$$

Note that the second term is subtracted to adjust for work saved from random manual screening.

2.2.3 True Negative Rate (TNR)

Here, the True Negative Rate (TNR) is the proportion of irrelevant records that were correctly not reviewed. TNR at recall (TNR@r%) is calculated by dividing the number of True Negatives at a given recall by the total number of irrelevant records

$$\frac{TN@r\%}{E} \text{ where E = number of irrelevant records}$$

A calculation example of the confusion matrix components, the WSS@r% and TNR@r% is given in Table 2.

Table 1

Definition and Calculation of Confusion Matrix Components for Active Learning (AL)

	Definition	Calculation
True Positives (TP)	The number of relevant records found.	Relevant Records * r%
False Positives (FP)	The number of irrelevant records falsely reviewed.	Records Reviewed – TP = Records Reviewed – (Relevant Records * r%)
True Negatives (TN)	The number of irrelevant records correctly not reviewed.	Irrelevant Records – FP = Irrelevant Records – (Records Reviewed – (Relevant Records * r%))
False Negatives (FN)	The number of relevant records not found.	Relevant Records – TP = Relevant Records – Relevant Records * r%

Note. The components can be retrieved at recall (r%) and at the number of records screened.

Table 2

Metric Calculation Example at 95% recall

Class imbalance	5% (Relevant Records)	20% (Relevant Records)
Total records	2000	2000
Records Reviewed	1100	1100
Relevant Records	100	400
Irrelevant Records	1900	1600
TP	100 * 95% = 95	400 * 95% = 380
FP	Records Reviewed – TP = 1100 – 95 = 1005	Records Reviewed – TP = 1100 – 380 = 720
TN	Irrelevant Records – FP = 1900 – 1005 = 895	Irrelevant Records – FP = 1600 – 720 = 880
FN	Relevant records – TP = 100 – 95 = 5	Relevant records – TP = 400 – 380 = 20
WSS95%	(TN + FN) / N – (1 – TP / (TP + FN)) = (895 + 5) / 2000 – (1 – 0.95) = 0.40	(TN + FN) / N – (1 – TP / (TP + FN) = (880 + 20) / 2000 – (1 – 0.95) = 0.40
TNR95%	TN / Irrelevant Records = 895 / 1900 = 0.47	TN / Irrelevant Records = 880 / 1600 = 0.55

Note. TP = True Positives, FP = False Positives, TN = True Negatives, FN = False Negatives. The WSS formula was introduced in Cohen et al. (2006).

Moreover, the example in Table 2 demonstrates how two datasets of equal size but differing class imbalance can have the same WSS@r% score, whereas the TNR@r% can differ substantially. Without considering the difference in class imbalance, one could wrongly conclude that a given model performed equally well on both datasets.

2.3. Implementation

The ASReview Insights extension is comprised of the following python modules: algorithms.py, metrics.py, plots.py, entrypoint.py, utils.py, and __init__.py. Note that the implementation steps are presented in chronological order.

In `algorithms.py`, to implement the confusion matrix components, functions were added to retrieve the TP, FP, TN, and FN values respectively. The functions were designed to return the values at recall (for metrics output).

In `metrics.py` the TP, FP, TN, and FN values are imported, and functions were added to slice the values at a given recall intercept. Then the sliced values were added to the metrics JSON file through the `get_metrics()` function, which also specifies default recall intercepts. The TNR was added to `metrics.py` by first dividing each element of the TN values by the number of irrelevant records. Similarly, the sliced value was then added to the metrics JSON file through the `get_metrics()` function. After review, this functionality has been made available in ASReview Insights v.1.1.2.

In addition to the scripts available in Insights v.1.1.2, two additional features were created. First, the TP, FP, TN, FN functions were adapted to also return the values at number of records screened. If the `x_screened` argument is specified as true, the values are returned at number of records screened. If argument is specified as false, the values are returned at recall. In `plot.py` the adaptation was used as input to plot the counts of TP, FN, TN, and FN as the number of records screened increases.

All python code and documentation are stored in the following GitHub repository: https://github.com/LSped/ASReview-metrics-comparability.

2.4 Simulation Design

2.4.1 Set up

To evaluate the difference in WSS95% and TNR95% scores, a simulation study was run using ASReview (v1.2). ASReview's Makita workflow generator (v0.6.3) (Teijema et al., n.d.) with the Makita basic template was used to create a folder structure and generate a jobs file. The jobs file includes commands line commands to run the naïve Bayes (NB) and TF-IDF simulations. The simulations are run with a default seed, with which two random prior knowledge records are selected (one relevant and one irrelevant). Moreover, the certainty-based query strategy and dynamic resampling are default settings. The NB + TF-IDF model and feature extractor combination was chosen based on Ferdinands et al. (2023).

2.4.2 Data

The 24 datasets used in this simulation study are part of the synergy dataset, which includes manually labeled data of 26 systematic reviews (de Bruin et al., 2023). Out of the 26 datasets, Brouwer_2019 and Walker_2018 were excluded from the analysis due to their large number of records (38114 and 48375 records respectively), which was expected to result in high computation time.

2.4.3 Statistical Analysis

First, differences in rank order based on the WSS95% and TNR@95% metric were examined, and a Wilcoxon rank-sum test was computed. The assumption of normality was tested with a Shapiro-Wilk test. Spearman's correlations were computed to examine the associations between the following variables: class imbalance and WSS@95%, class imbalance and TNR@95%, class imbalance and TNR@95% - WSS@95%.

2.5 Analytic strategy

The results section is structured as follows. First, the confusion matrix output, including metrics and plots, is presented. Next, the TNR metrics output is presented. Finally, the results of the evaluation of differences between the WSS95% metric and TNR95% metric across datasets are presented.

3. Results

3.1 Confusion Matrix Output

3.1.1 Metrics File

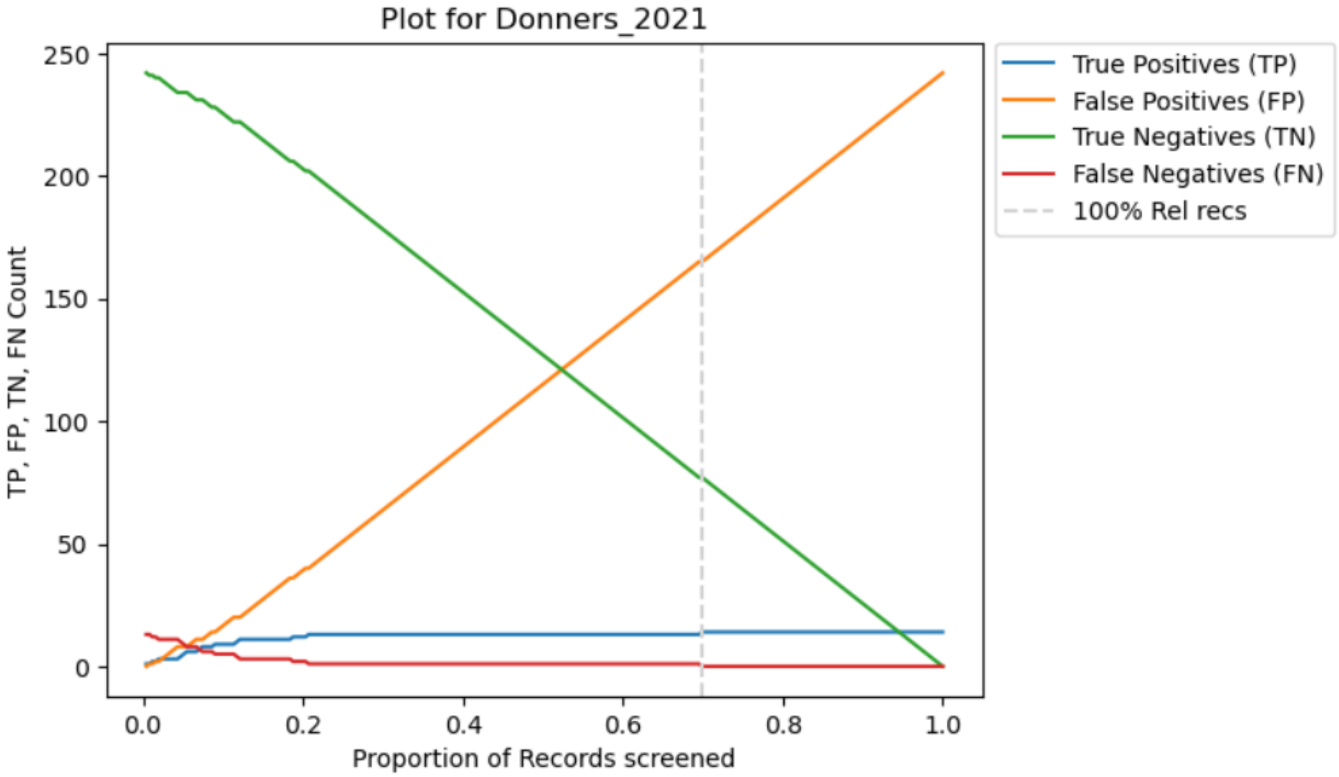
An example of the generated output JSON file with the added confusion matrix components at a range of recall intercepts can be found in Appendix A - D. The output format is in line with previously reported metrics and is chosen to prevent any downstream issues in the software. The values can be retrieved at any specified recall intercept via the command line. The example in Appendix A and Appendix B stems from the NB + TF-IDF simulation on the Donners_2021 dataset. For illustrative purposes, the values are shown at the recall intercepts: 0.1, 0.25, 0.5, 0.75, 0.8, 0.85, 0.9, 0.95, and 1. Moreover, the metrics output for all datasets included in the simulation is stored in an Excel file, which can be used for further analysis. The resulting excel file (with a large range of intercepts) is available in the following GitHub repository: https://github.com/LSped/asreview-insights-metrics-comparability-main/tree/main/output.

3.1.2 Plots

Figures 1 and 2 show the frequencies of the confusion matrix values over the number (or proportion) of records screened on two example datasets: Donners_2021 & Nelson 2002. The TP curve is mirrored by the FN curve and the same applies to the FP and TN curves. The mirrored pattern is present because at a given number or proportion of records screened, the sums of TP + FN = Relevant Records, and FP + TN = Irrelevant Records respectively, are constant. The main additional insight from the plots below is that they show the number of False Positives at a given number (or proportion) of relevant records, which shows how many irrelevant records had to be reviewed to reach the TP count. In addition, the number of True Negatives and False Negatives can be easily read at any point. Since the confusion matrix components are interrelated, this information could be deduced from the recall plot present in ASReview Insights v1.1.2 (ASReview LAB Developers, n.d.) when knowing the total number of relevant and irrelevant records in the dataset. However, the added curves provide additional insights into the how the rate of True Positives and False Positives progress over records screened. For instance, the Nelson 2002 dataset stands out compared to all other datasets since from 0% to close to 30% records screened the FP rate is lower than the TP rate. In other words, precision increases as the number of records screened increases until the curves intersect. This could perhaps indicate a cluster of very similar relevant records, which could not be read from the TP curve alone. The plot provides a picture over performance throughout the screening process. It indicates at which point half of the relevant records are found, namely when the TP curve crosses the FN curve. Most datasets have a small percentage of relevant records, which is now more clearly reflected in the plots, however, it renders the TP curve less readable (see Appendix F-H). The plots of all 24 datasets are available in better resolution at the following GitHub repository https://github.com/LSped/asreview-insights-metrics-comparability-main/blob/main/output/New%20Output%20.ipynb

Figure 1

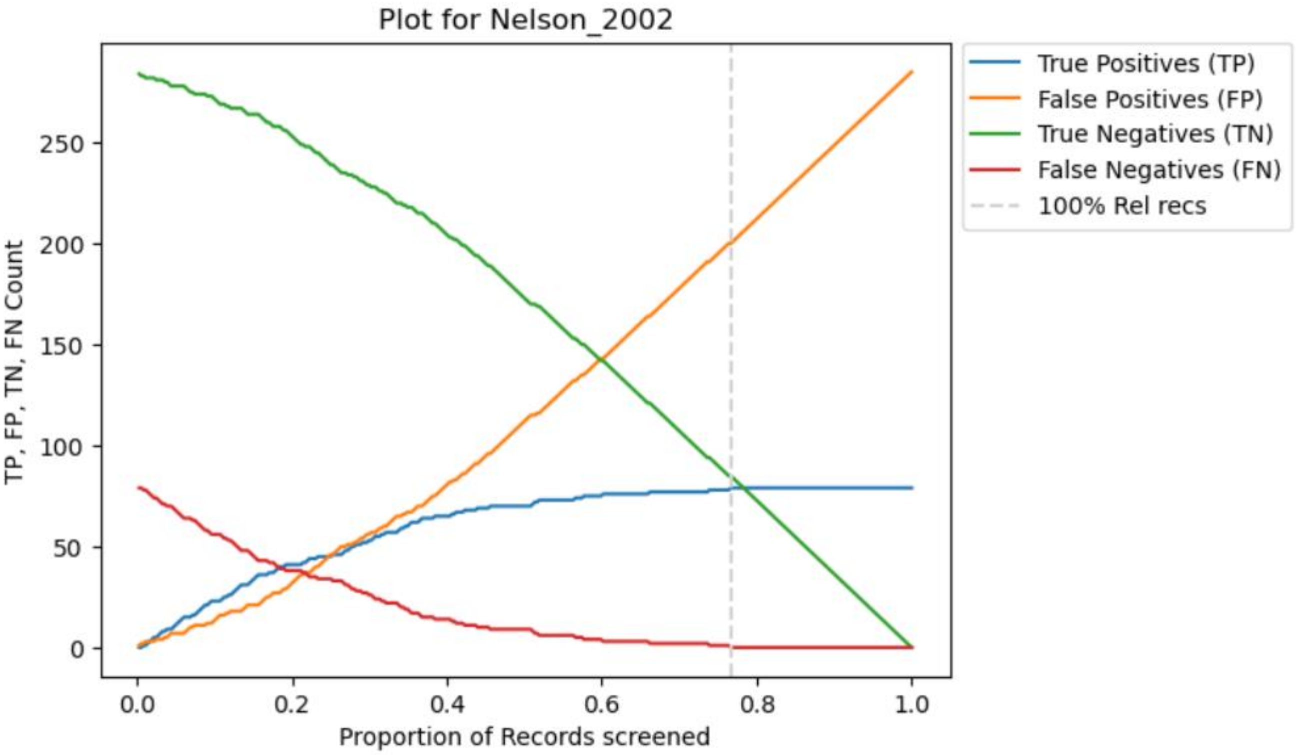
Confusion matrix components at % of records screened (NB+TF-IDF) simulation Donners_2021 dataset



Note. The dataset is part of the synergy dataset which consists of 26 manually labeled systematic literature reviews.

Figure 2

Confusion matrix components at % of records screened (NB+TF-IDF) simulation Nelson_2002 dataset



Note. The dataset is part of the synergy dataset which consists of 26 manually labeled systematic literature reviews.

3.2 True Negative Rate (TNR) Output

The TNR@r% is outputted in the same JSON file as the confusion matrix. The TNR@r% output can be seen in Appendix E. Like the confusion matrix values, the TNR@r% can be returned at any specified intercept via the command line. Moreover, for all datasets part of a simulation the TNR@r% metrics output is stored in an Excel file, which can be used for further analysis.

The resulting excel file (with a large range of intercepts) is available in the following GitHub repository https://github.com/LSped/asreview-insights-metrics-comparability-main/tree/main/output.

3.3 Simulation study

The results of the comparison of the WSS@r% and the TNR@r% metrics across datasets are presented in Table 3. The rank order of datasets in terms of highest performance score changes whether the WSS@95% or TNR@95% metric is used. For instance, using the WSS95% score, Hall_2012 is the best performing dataset, while when using the TNR95% score it switches positions with Leenaars_2019. However, the Wilcoxon rank-sum test, which tests for a difference in ranks, was not significant. Since the class imbalance, WSS95%, and TNR95% variables do not follow normal distributions based on the Shapiro-Wilk test, a Spearman correlation test was performed. The Spearman correlation between class imbalance and WSS@95% is negative with a correlation coefficient ρ= -0.62 and P = 0.001. The correlation between class imbalance and TNR@95% is negative with a correlation coefficient ρ= -0.61 and P = 0.001. Both results are significant with similar and large correlation coefficients, whereas the last is slightly smaller in absolute terms. The Spearman correlation between the difference in TNR95% and WSS95% scores and class imbalance is positive and not significant, with ρ= 0.15 and P = 0.46 respectively.

Table 3

Comparison of WSS95% and TNR95% scores of (NB + TF-IDF) simulations on 24 synergy datasets

Dataset	wss95%	tnr95%	rank (wss95%)	rank (tnr95%)	rank difference	difference (tnr95%-wss95%)	class imbalance
Appenzeller-Herzog_2019	0.795	0.902	7	6	-1	0.107	0.9
Bos_2018	0.815	0.983	5	3	-2	0.168	0.2
Chou_2003	0.449	0.56	18	18	0	0.111	0.8
Chou_2004	0.209	0.398	22	22	0	0.189	0.6
Donners_2021	0.688	0.835	12	11	-1	0.147	5.8
Hall_2012	0.911	0.985	1	2	1	0.074	1.2
Jeyaraman_2020	0.543	0.648	17	17	0	0.105	8.2
Leenaars_2019	0.895	0.991	2	1	-1	0.096	0.3
Leenaars_2020	0.577	0.68	16	16	0	0.103	8.1
Meijboom_2021	0.689	0.788	11	12	1	0.099	4.2
Menon_2022	0.646	0.76	15	14	-1	0.114	7.6
Moran_2021	0.145	0.207	24	24	0	0.062	2.1
Muthu_2021	0.334	0.433	21	21	0	0.099	12.4
Nelson_2002	0.368	0.526	20	19	-1	0.158	21.9
Oud_2018	0.675	0.767	13	13	0	0.092	2.1
Radjenovic_2013	0.867	0.948	4	5	1	0.081	0.8
Sep_2021	0.167	0.261	23	23	0	0.094	14.8
Smid_2020	0.792	0.896	8	7	-1	0.104	1
van_de_Schoot_2018	0.894	0.969	3	4	1	0.075	0.8
Valk_2021	0.398	0.517	19	20	1	0.118	12.3
van_der_Waal_2022	0.756	0.846	10	9	-1	0.09	1.7
van_Dis_2020	0.662	0.73	14	15	1	0.069	0.8
Wassenaar_2017	0.775	0.846	9	10	1	0.07	1.4
Wolters_2018	0.809	0.896	6	8	2	0.087	0.4

Note. The datasets are part of the synergy dataset which consists of 26 manually labeled systematic literature reviews.

.

4. Discussion

The main aim of the present study was to contribute towards model performance comparability across datasets within the field of simulation-based AL for systematic reviews. The goal was achieved by enhancing the metrics reported by the ASReview screening software (van de Schoot et al., 2021).

The first step constituted of adding the confusion matrix components at recall to the ASReview Insights Extension v1.1.2 (ASReview LAB Developers, n.d.), which were successfully implemented into the software. Moreover, two additional functionalities were created. First, functions were adapted such that the confusion matrix could be retrieved at number of records screened. Second, this allowed the creation of plots representing the confusion matrix component counts at number or proportion of records screened. These two additional features have been made available in the project's GitHub repository. Furthermore, the True Negative Rate (TNR@r%) at recall, equal to the normalized WSS (Kusa et al., 2023), was successfully integrated into the output of ASReview Insights v.1.1.2.

Moreover, the NB + TF-IDF simulation performed on 24 datasets shows that the order of the datasets based on the performance scores changes depending on whether the WSS@95% metric or the TNR@95% metric is used. The correlation between WSS95% and class imbalance is negative and significant. Similarly, the correlation between TNR95% and class imbalance is negative and significant. The negative relationship between WSS95% and class imbalance is expected. However, the negative relationship between TNR95% and class imbalance is less intuitive. After adjusting for class imbalance, the correlation coefficient is only slightly smaller. Therefore, class imbalance seems to be related to the TNR@95% score, influencing model performance in other ways. Moreover, while the correlation between class imbalance and the

TNR95% - WSS95% difference is positive as expected, this result was not significant. This result suggests that the adjustment for class imbalance does not lead to significant differences in performance scores on the present datasets in comparison to the WSS95% metric. Note that the task of systematic literature reviews is characterized by high class imbalance. Therefore, the range of imbalance does not span to a fully balanced dataset, perhaps making the impact less visible. In the 24 synergy datasets, the percentage of relevant records ranges from 0.2% to 21.9%, with a median of 1.9% relevant records.

While the implementation of the new features represents a step towards comparability, certain limitations remain. First, concerning the new confusion matrix output, the components are only retrievable in absolute numbers and not in proportions. A future implementation step would be to adapt the TP, FP, TN, FN components to be displayed as proportions as well, resulting in the True Positive Rate (TPR), the False Positive Rate (FPR), the True Negative Rate (TNR), and the False Negative Rate (FNR). Moreover, future work could consider adapting the present confusion matrix functions such that they can be used to calculate the utility metric comprised of yield and burden, which is an AL specific metric (Miwa et al., 2014). It makes a distinction between labeled and unlabeled TP, FP, TN, and FN, which requires increased coding effort and considerations regarding class imbalance. A broader limitation of this study is the premise of the certainty-based query strategy for defining the confusion matrix components. While the certainty-based query strategy is an effective choice for the high class imbalance task, to address uncertainty-based, mixed queries or clustering, the definitions of the confusion matrix components would need to be reconsidered.

Another limitation of the current study is that the comparison of model performance on 24 datasets was examined only with the basic simulation template, which does not run a simulation with different prior knowledge. Similarly, the performance of only one model and feature extractor combination was examined. Therefore, the results do not provide the full picture of using the TNR@r% metric instead of the WSS@r% metric to evaluate model performance across all present datasets. Further simulation studies should compare differences in terms of WSS95% and TNR95% across different models and feature extractor combinations and across different prior knowledge to make definite conclusions on the datasets used. Nonetheless, the TNR metric encourages future research efforts to isolate the influence of data characteristics other than class imbalance on performance. Such researchable data characteristics may include (un)controlled vocabulary, clustering, heterogeneity of relevant records, inclusion criteria etc.

Note that performance conclusions across datasets without class imbalance adjustment are made in the field (Ferdinands et al., 2023; Cohen et al., 2006). Therefore, the inclusion of the TNR metric is a valuable addition for data scientists who want to study and improve model performance of AL-assisted systematic reviews. The TNR@r% metric allows researchers to gain deeper insights into the reasons behind performance discrepancies across different datasets while adjusting for dataset class imbalance. In turn, a deeper understanding of performance discrepancies not only aids model refinements but also enables the identification of datasets that consistently exhibit poor performance.

In addition, a better understanding of AL performance benefits users of AL screening software by having more effective model configurations at their disposal. Having a comprehensive understanding of which model and feature extractor combinations work best for certain datasets can greatly assist users in selecting the most suitable model combination for a new dataset. Finally, if users are in possession of a manually labeled dataset and want to perform AL-assisted screening on a new but similar dataset, they can make use of the simulations on the labeled dataset themselves to choose the best performing model for the new dataset.

In conclusion, there is a need for more extensive and comparative performance analysis across various models and datasets in the field of AL for SLR tasks. The implementation of the TNR@r% into ASReview Insights v1.1.2 contributes towards addressing this need. Future studies should prioritize further investigation into the use of the TNR@r% metric, as a normalized WSS metric, in the AL context. In sum, the current study calls for appropriate metrics to address the issue of class imbalance when evaluating model performance across datasets.

References

ASReview LAB developers. ASReview Insights - Insights and plotting tool for the ASReview project [Computer software]. https://github.com/asreview/asreview-insights

Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P. Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association : JAMIA*, *13*(2), 206–219. https://doi.org/10.1197/jamia.M1929

De Bruin, J., Ma, Y., Ferdinands, G., Teijema, J., Van de Schoot, R. (2023), SYNERGY - Open machine learning dataset on study selection in systematic reviews. *DataverseNL, V1*. https://doi.org/10.34894/HE6NAQ,

DistillerSR. (2022, June 28). *Re-Rank Simulation - DistillerSR User Guide - 1*. DistillerSR User Guide. http://v2dis-help.evidencepartners.com/1/en/topic/ai-simulation

Ferdinands, G., Schram, R., de Bruin, J., Bagheri, A., Oberski, D. L., Tummers, L., Teijema J. J., & van de Schoot, R. (2023). Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the Average Time to Discover relevant records. *Systematic Reviews*, *12*(1), 100.

Kusa, W., Lipani, A., Knoth, P., & Hanbury, A. (2023). An analysis of work saved over sampling in the evaluation of automated citation screening in systematic literature reviews. *Intelligent Systems with Applications*, *18*, 200193.

Kusa, W., Hanbury, A., & Knoth, P. (2022). Automation of citation screening for systematic literature reviews using neural networks: A replicability study. In Advances in information retrieval, *44th European conference on IR research, ECIR 2022*. https://doi.org/10.1007/978-3-030-99736-6_39

Miwa, M., Thomas, J., O'Mara-Eves, A., & Ananiadou, S. (2014). Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics*, *51*, 242-253.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, *4*(1), 1-22.

Teijema, J. J., Seuren, S., Anadria, D., Bagheri, A., & van de Schoot, R. (2023, June 29). Simulation-based Active Learning for Systematic Reviews: A Systematic Review of the Literature. https://doi.org/10.31234/osf.io/67zmt

Teijema, J., Van de Schoot, R., Ferdinands, G., Lombaers, P., & De Bruin, J. ASReview Makita: a workflow generator for simulation studies using the command line interface of ASReview LAB [Computer software]. https://github.com/asreview/asreview-makita

Van Dinter, R., Catal, C., & Tekinerdogan, B. (2021). A multi-channel convolutional neural network approach to automate the citation screening process. Applied Soft Computing, 112, Article 107765. https://doi.org/10.1016/J.ASOC.2021.107765.

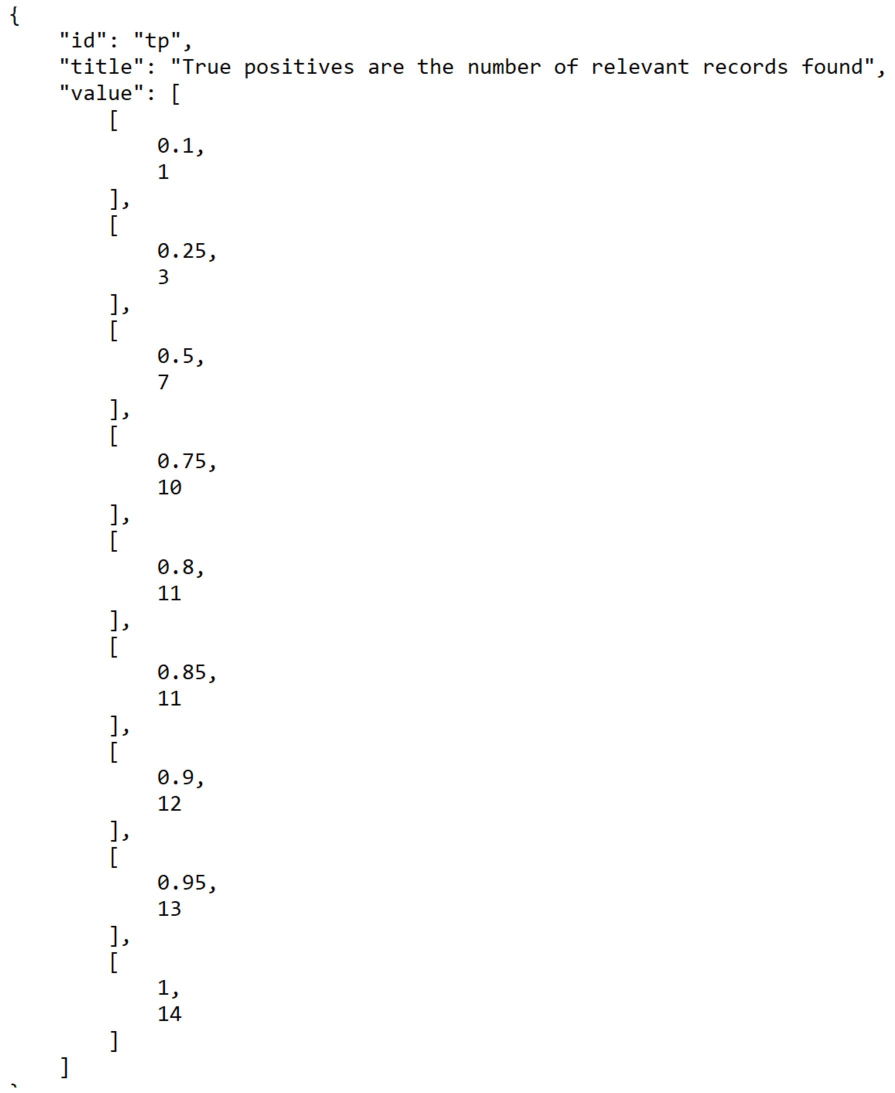
van de Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdema, F., ... & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature machine intelligence*, *3*(2), 125-133.

Yu, Z., & Menzies, T. (2019). FAST2: An intelligent assistant for finding relevant papers. *Expert Systems with Applications*, *120*, 57-71.

Appendix A

Figure 1

True Positives (TP) Example Metrics Output (NB + TF-IDF) simulation on Donners 2021

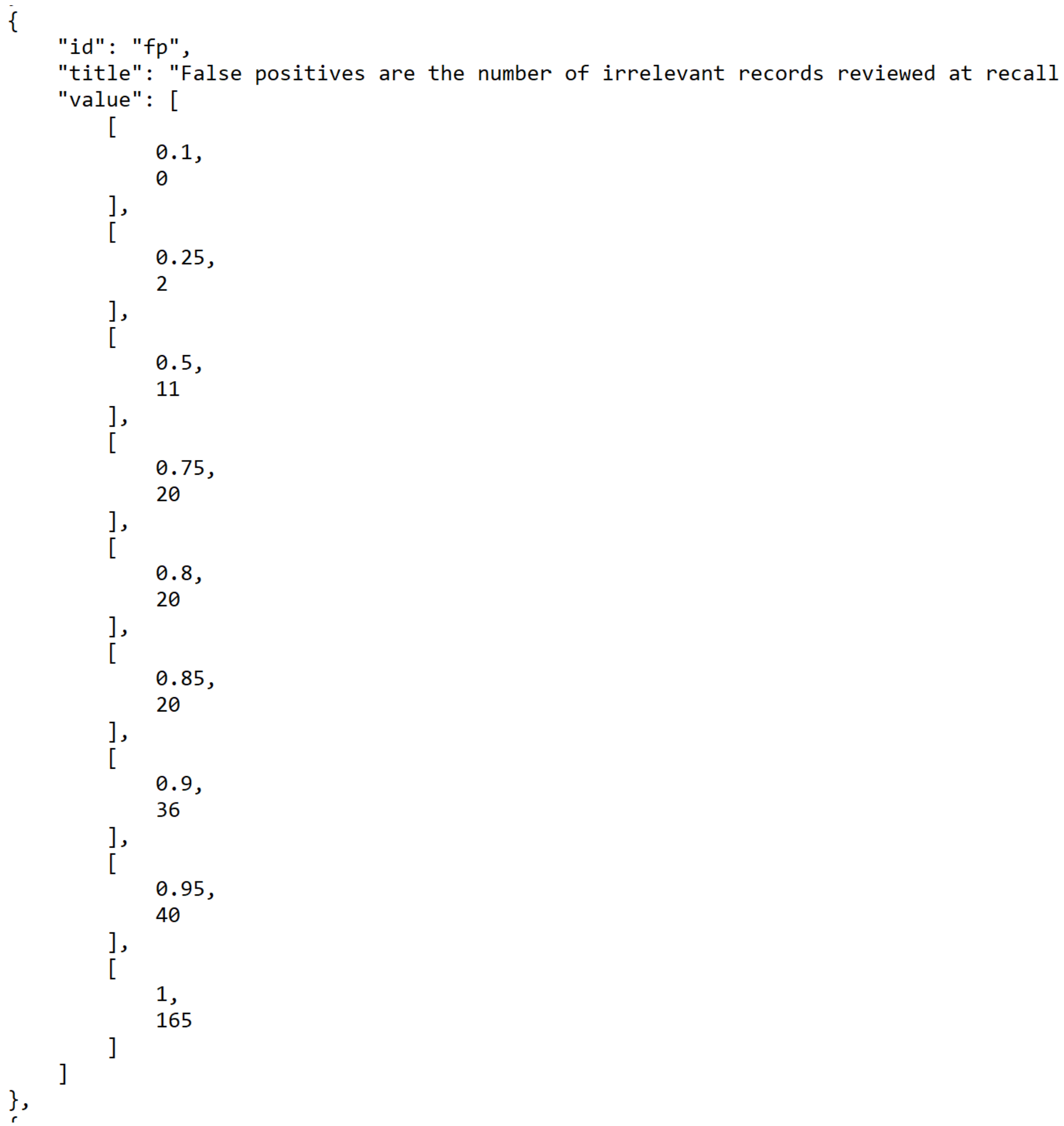


```
{
    "id": "tp",
    "title": "True positives are the number of relevant records found",
    "value": [
        [
            0.1,
            1
        ],
        [
            0.25,
            3
        ],
        [
            0.5,
            7
        ],
        [
            0.75,
            10
        ],
        [
            0.8,
            11
        ],
        [
            0.85,
            11
        ],
        [
            0.9,
            12
        ],
        [
            0.95,
            13
        ],
        [
            1,
            14
        ]
    ]
```

Appendix B

Figure 2

False Positives (FP) Example Metrics Output (NB + TF-IDF) simulation on Donners 2021

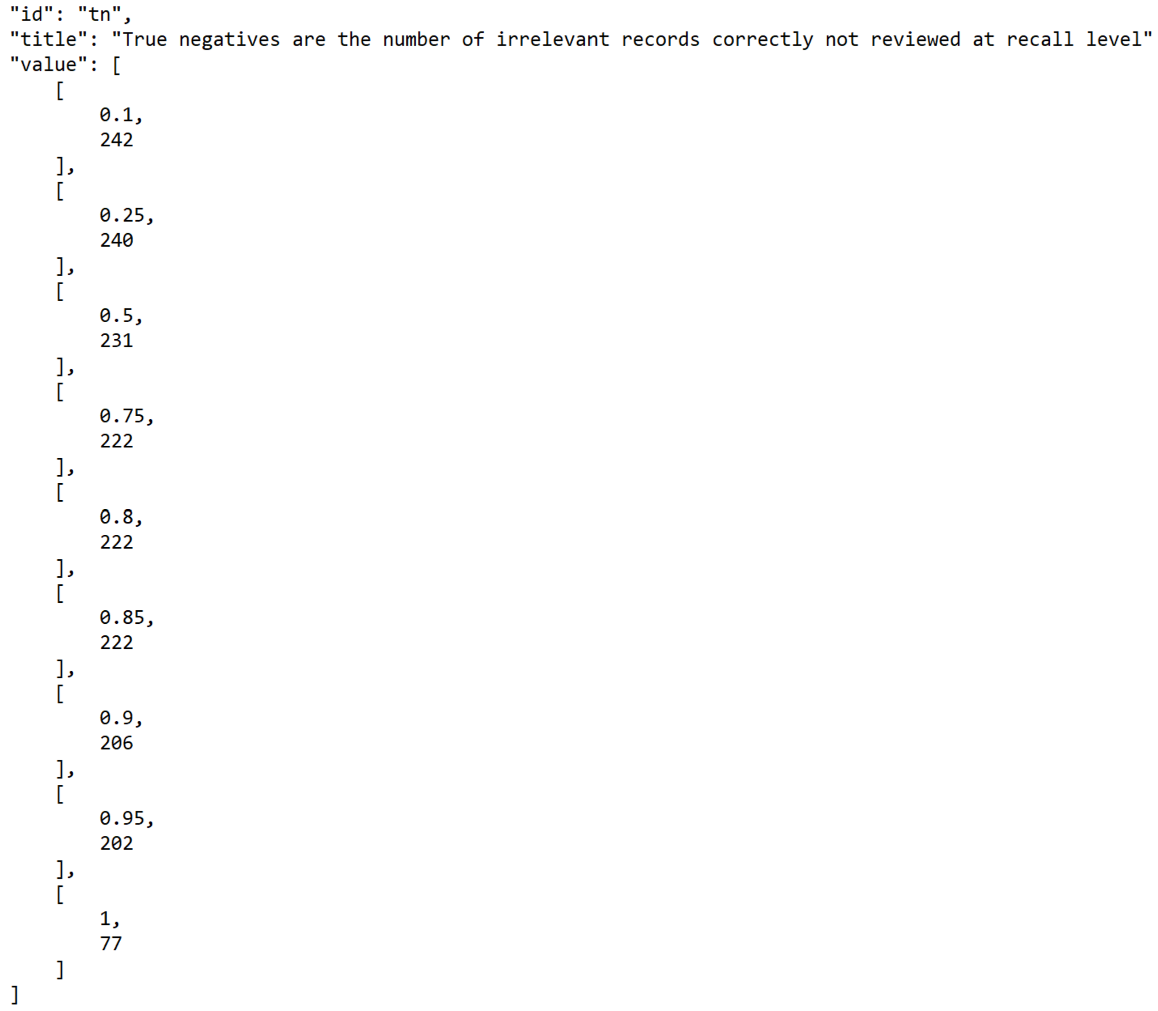


```
{
    "id": "fp",
    "title": "False positives are the number of irrelevant records reviewed at recall
    "value": [
        [
            0.1,
            0
        ],
        [
            0.25,
            2
        ],
        [
            0.5,
            11
        ],
        [
            0.75,
            20
        ],
        [
            0.8,
            20
        ],
        [
            0.85,
            20
        ],
        [
            0.9,
            36
        ],
        [
            0.95,
            40
        ],
        [
            1,
            165
        ]
    ]
},
```

Appendix C

Figure 3

True Negatives (TN) Example Metrics Output (NB + TF-IDF) simulation on Donners 2021

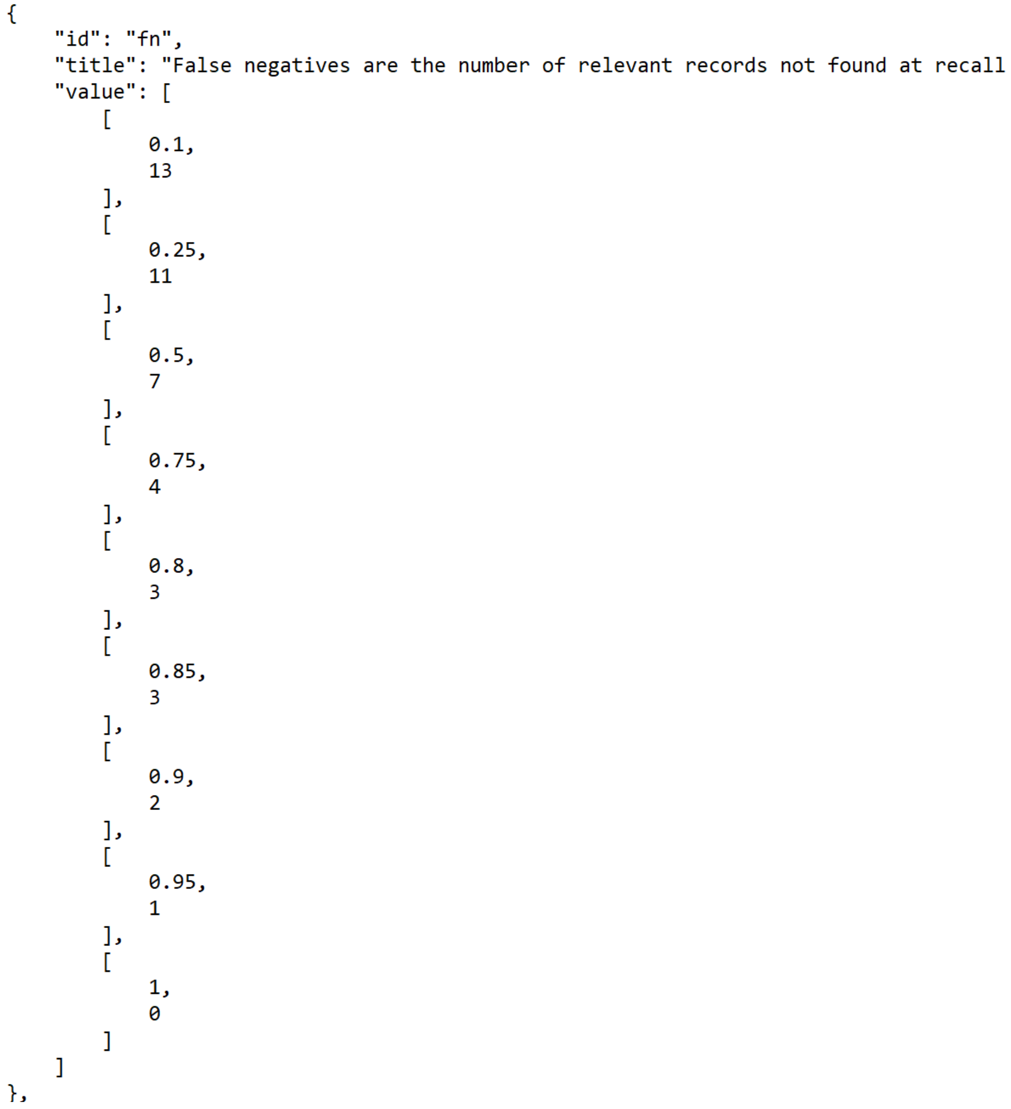


```
"id": "tn",
"title": "True negatives are the number of irrelevant records correctly not reviewed at recall level"
"value": [
    [
        0.1,
        242
    ],
    [
        0.25,
        240
    ],
    [
        0.5,
        231
    ],
    [
        0.75,
        222
    ],
    [
        0.8,
        222
    ],
    [
        0.85,
        222
    ],
    [
        0.9,
        206
    ],
    [
        0.95,
        202
    ],
    [
        1,
        77
    ]
]
```

Appendix D

Figure 4

False Negatives (FN) Example Metrics Output (NB + TF-IDF) simulation on Donners 2021

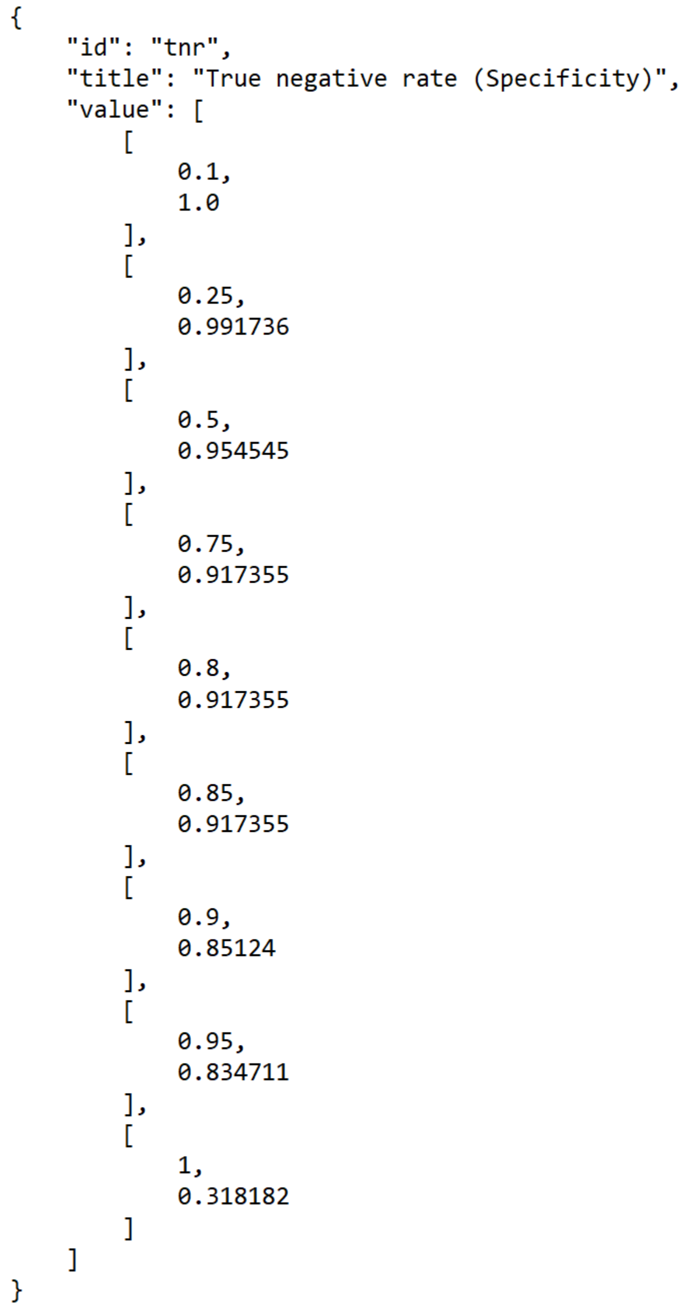


```
{
    "id": "fn",
    "title": "False negatives are the number of relevant records not found at recall
    "value": [
        [
            0.1,
            13
        ],
        [
            0.25,
            11
        ],
        [
            0.5,
            7
        ],
        [
            0.75,
            4
        ],
        [
            0.8,
            3
        ],
        [
            0.85,
            3
        ],
        [
            0.9,
            2
        ],
        [
            0.95,
            1
        ],
        [
            1,
            0
        ]
    ]
},
```

Appendix E

Figure 5

True Negative Rate (TNR@r%) Example Output (NB + TF-IDF simulation on Donners 2021)

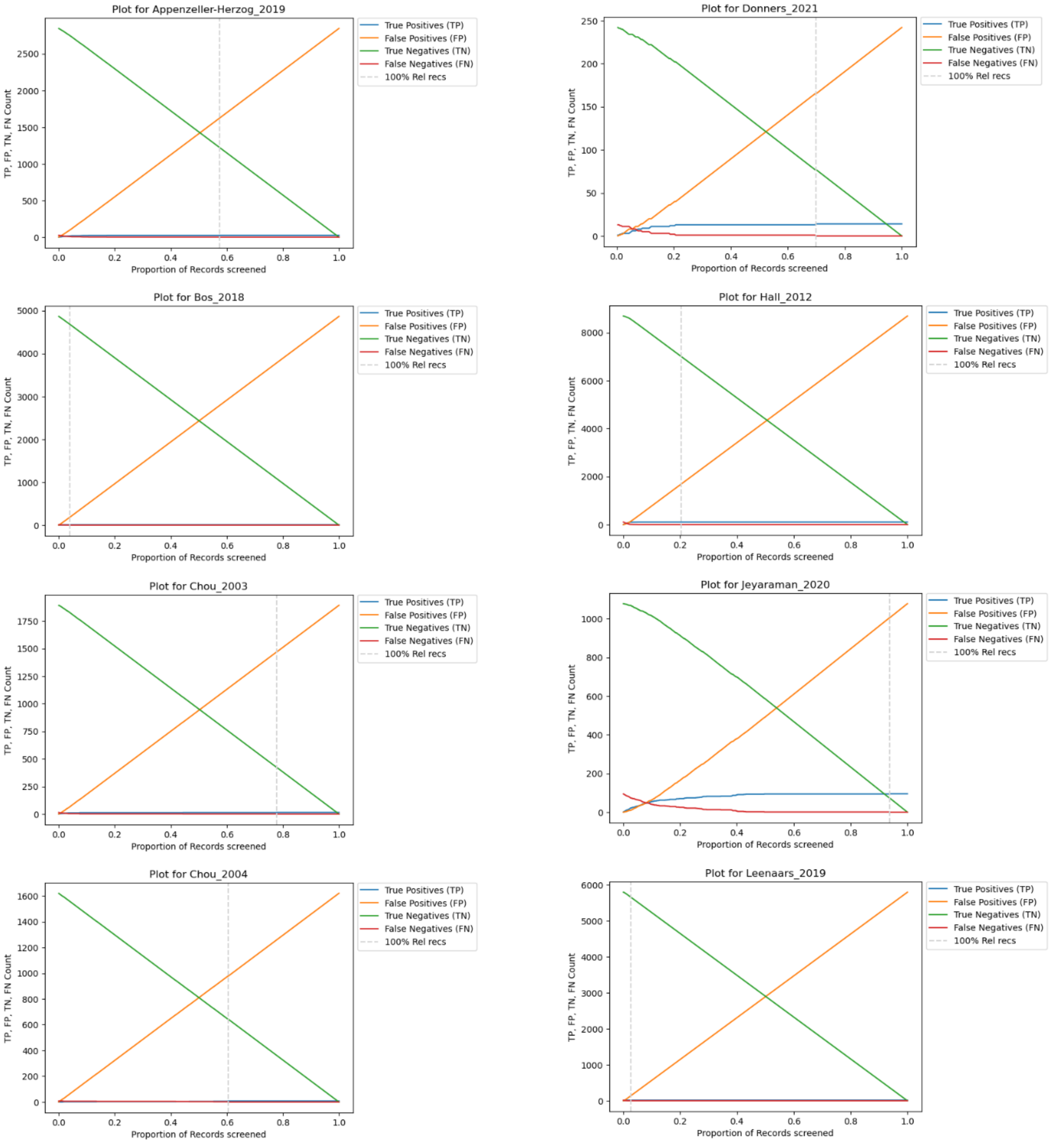


```
{
    "id": "tnr",
    "title": "True negative rate (Specificity)",
    "value": [
        [
            0.1,
            1.0
        ],
        [
            0.25,
            0.991736
        ],
        [
            0.5,
            0.954545
        ],
        [
            0.75,
            0.917355
        ],
        [
            0.8,
            0.917355
        ],
        [
            0.85,
            0.917355
        ],
        [
            0.9,
            0.85124
        ],
        [
            0.95,
            0.834711
        ],
        [
            1,
            0.318182
        ]
    ]
}
```

Appendix F

Figure 6a

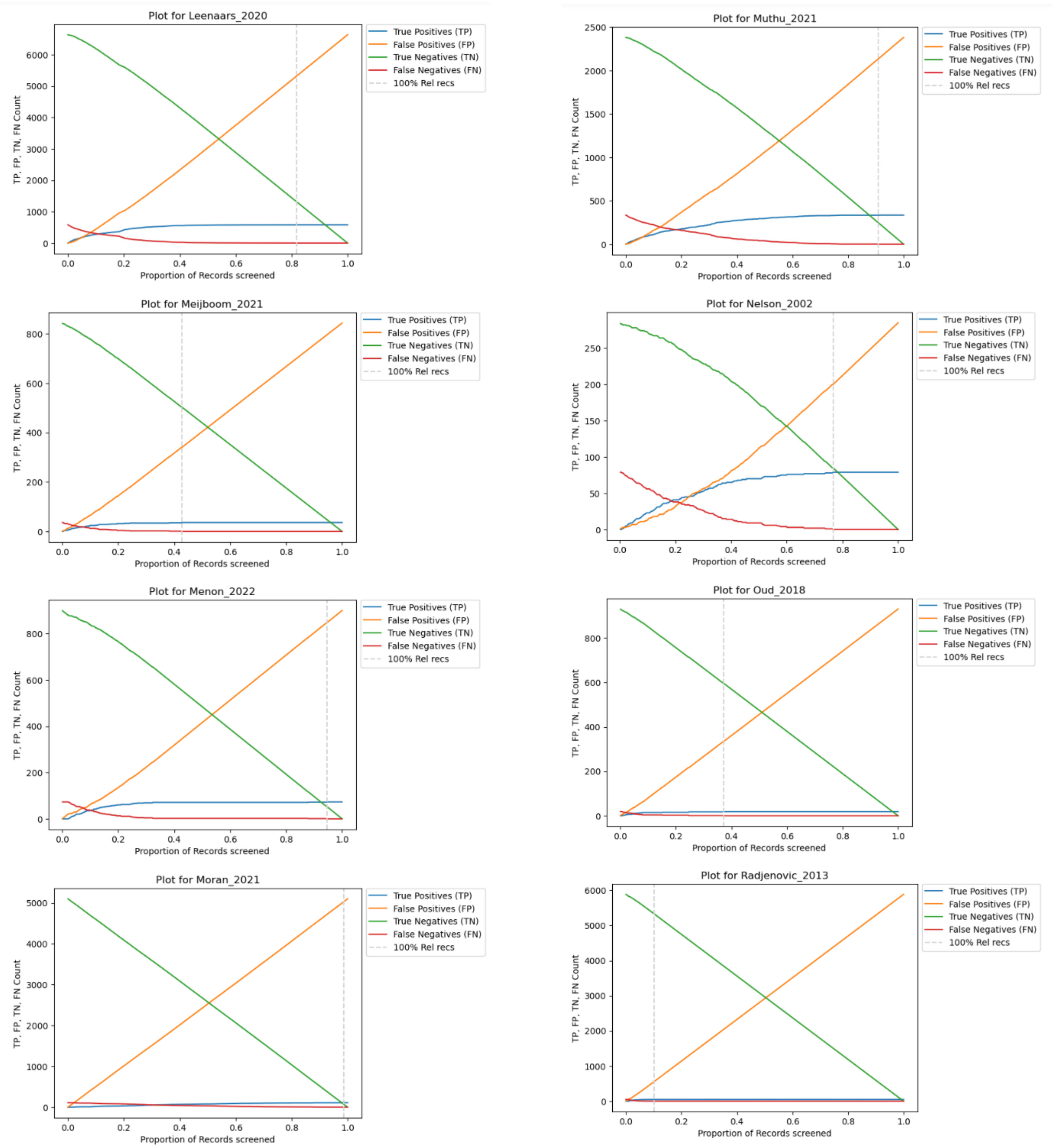
Confusion matrix components at % of records screened (NB+TF-IDF) simulation on synergy dataset



Appendix G

Figure 6b

Confusion matrix components at % of records screened (NB+TF-IDF) simulation on synergy dataset



Plot for Leenaars_2020

Plot for Muthu_2021

Plot for Meijboom_2021

Plot for Nelson_2002

Plot for Menon_2022

Plot for Oud_2018

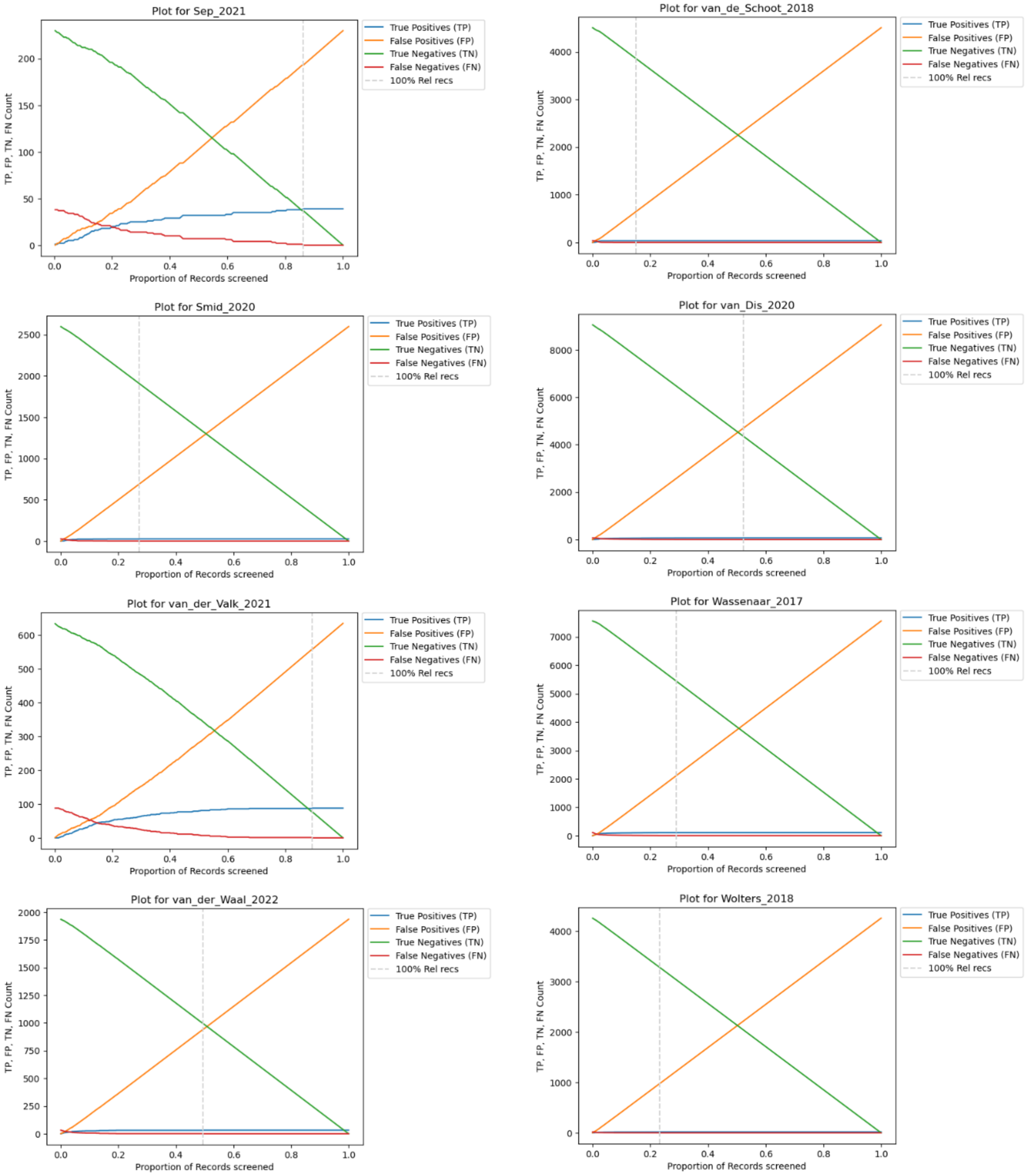
Plot for Moran_2021

Plot for Radjenovic_2013

Appendix H

Figure 6c

Confusion matrix components at % of records screened (NB+TF-IDF) simulation on synergy dataset



Data and Code availability

The data used in the present study did not contain any sensitive personal information. The data consists of digital object identifies (DOI), titles, abstracts, and inclusion labels of manually labeled systematic literature reviews. It is a free and open dataset accessible on GitHub https://github.com/asreview/synergy-dataset. All code to reproduce the results described in this paper is available at the following GitHub repository (https://github.com/LSped/asreview-insights-metrics-comparability-main/tree/main)

Acknowledgement of AI tools

ChatGPT, an advanced language model developed by OpenAI, served as a supplementary aid for content review with all final interpretations and content decisions falling under sole responsibility of the author.