Intelligent Systems with Applications 18 (2023) 200193

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

An analysis of work saved over sampling in the evaluation of automated citation screening in systematic literature reviews

Wojciech Kusa[a,*], Aldo Lipani[b], Petr Knoth[c], Allan Hanbury[a,d]

[a] TU Wien, Vienna, Austria
[b] University College London, London, United Kingdom
[c] The Open University, Milton Keynes, United Kingdom
[d] Complexity Science Hub, Vienna, Austria

ARTICLE INFO

Keywords:
Citation screening
Study selection
Evaluation
Work Saved over Sampling (WSS)
Systematic Literature Review (SLR)

ABSTRACT

Citation screening is an essential and time-consuming step of the systematic literature review process in medicine. Multiple previous studies have proposed various automation techniques to assist manual annotators in this tedious task. The most widely used measure for the evaluation of automated citation screening techniques is Work Saved over Sampling (WSS). In this work, we analyse this measure and examine its drawbacks. We subsequently propose to normalise WSS which enables citation screening performance comparisons across different systematic reviews. We analytically show that normalised WSS is equivalent to the True Negative Rate (TNR). Finally, we provide benchmark scores for fifteen systematic review datasets with TNR@95% recall measure and compare the measure with Precision and AUC.

1. Introduction

Systematic literature reviews are recall-focused, secondary studies that synthesise all relevant data providing an answer to a specific clinical question. Since their conclusions are considered as the gold standard in evidence-based medicine, systematic reviews follow strict criteria. Conducting systematic reviews is a slow, repetitive, and time-consuming process that relies primarily on human labour.

Out of all stages of a systematic literature review process, citation screening (also known as the selection of primary studies) is estimated to be one of the most time-consuming steps (Bannach-Brown et al., 2019). It often requires screening (tens of) thousands of studies for eligibility with respect to the study criteria. Traditionally, the process is divided into two stages. In the first stage, only titles and abstracts are appraised to save time and resources. This is followed by an appraisal of the full texts of articles, a more detailed and more time-consuming assessment of all papers included from the first stage (Tsafnat et al., 2018).

To this date, many machine learning algorithms have been proposed to automate citation screening. According to the recent systematic review on this topic (van Dinter, Tekinerdogan et al., 2021), there were 25 papers published on automation of the selection of primary studies. An older systematic review from 2014 found in total 44 studies dealing implicitly or explicitly with the problem of citation screening (O'Mara-Eves et al., 2015). Already several commercial systems offer, to some degree, automation of the screening process.

All automated citation screening models can be coarsely classified into either classification or ranking approaches. Both follow a similar approach and use natural language processing to train a supervised model on an annotated sample of a dataset to determine whether a paper should be included or excluded from the review. A successful automated citation screening algorithm should miss as few relevant papers as possible and also save time for the reviewers by removing irrelevant papers.

In the field of citation screening, the most commonly used custom evaluation measure is Work Saved over Sampling at $r\%$ recall (WSS@r%). It was introduced by Cohen et al. (2006) as a measure be-

* Corresponding author.
E-mail addresses: wojciech.kusa@tuwien.ac.at (W. Kusa), aldo.lipani@acm.org (A. Lipani), petr.knoth@open.ac.uk (P. Knoth), allan.hanbury@tuwien.ac.at (A. Hanbury).
URLs: https://informatics.tuwien.ac.at/people/wojciech-kusa (W. Kusa), https://www.ucl.ac.uk/civil-environmental-geomatic-engineering/people/dr-aldo-lipani (A. Lipani), https://www.open.ac.uk/people/pk3295 (P. Knoth), https://informatics.tuwien.ac.at/people/allan-hanbury (A. Hanbury).

https://doi.org/10.1016/j.iswa.2023.200193
Received 30 May 2022; Received in revised form 30 November 2022; Accepted 29 January 2023
Available online 10 February 2023

ing able to balance between very high recall and optimal precision. They describe WSS@r% 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).*" It estimates the human screening workload reduction by using automation tools, assuming a fixed recall level of $r\%$.

Work Saved over Sampling given a recall set at $r\%$, is defined as follows:

$$WSS@r\% = \frac{TN + FN}{N} - (1 - r), \tag{1}$$

where TN is the number of true negatives (excludes that were correctly removed), FN is the number of false negatives (includes that were incorrectly marked as irrelevant documents), and N is the total number of documents.

The choice of recall level is influenced by the domain and characteristics of the review. Past studies on the automation of citation screening in medicine typically used 95% recall as the threshold to preserve a satisfactory quality of the systematic literature review in medicine (Cohen et al., 2006). In other technology-assisted review systems, e.g. e-discovery, recall levels might be lower, and sometimes this choice is influenced by time or money limitations.

This paper examines WSS and investigates its properties and terms with their influence on the final score. Similarly to the Discounted Cumulative Gain (DCG) metric (Järvelin & Kekäläinen, 2002), we propose to normalise the WSS in order to be able to compare the scores between multiple models and datasets. This representation preserves all the features of the WSS and simultaneously removes some constants from the equation. Furthermore, we show that the normalised WSS is equivalent to the True Negative Rate (TNR) also known as specificity. Using the derived equation, we calculate and provide benchmark scores for fifteen systematic review datasets with the TNR@95% recall measure. Finally, we recommend using TNR at $r\%$ recall as the evaluation measure for technology-assisted reviews. However, before starting, we introduce the notation used.

1.1. Notation

The basic symbols and sets used in this paper are given in the following table.

$\mathcal{I}$	set of relevant documents that should be included in the review, *includes*
$\mathcal{E}$	set of irrelevant documents that should be excluded in the review, *excludes*
$\lvert\mathcal{I}\rvert$	number of *includes*
$\lvert\mathcal{E}\rvert$	number of *excludes*
N	total number of documents $\lvert\mathcal{I}\rvert + \lvert\mathcal{E}\rvert$
TP	number of *true positives*, i.e., includes classified correctly
TN	number of *true negatives*, i.e., excludes classified correctly
FP	number of *false positives*, i.e., excludes classified incorrectly
FN	number of *false negatives*, i.e., includes classified incorrectly
$r\%$	a recall value of $r\%$
$n_{r\%}$	rank of a document for which the recall level of $r\%$ is achieved
$X@r\%$	evaluation measure X calculated at a fixed recall value of $r\%$

2. Related work

WSS was previously used in multiple studies to evaluate the effectiveness of a supervised machine learning system for citation screening (Matwin et al., 2010, Howard et al., 2016, Scells et al., 2019, Kontonatsios et al., 2020, Kusa et al., 2022). It was also used as one of evaluation measures for the Technology Assisted Review shared task at CLEF by Kanoulas et al. (2017, 2018).

O'Mara-Eves et al. (2015) mention that there is a subjective component for metrics like Fβ-score and WSS. Evaluators determine thresholds and parameters, making it difficult to compare across studies. It is also not always transparent or justified how the thresholds/weights are chosen. Cohen (2008) in their later study abandoned WSS in favour of Area Under the ROC Curve (AUC) as they argue that the former metric fails to capture different recall-precision trade-offs in different reviews. On the other hand, Cormack and Grossman (2017) mention that cumulative measures like area under the cumulative recall curve and average precision yield very little insight into the actual or hypothetical effectiveness of the models.

Norman (2020) notices that despite WSS being relatively easy to interpret in the context of automation of systematic reviews, it is also strongly influenced by random effects and tends to have a large variance. Recall versus effort plots using the *knee method* (Cormack & Grossman, 2016) can be used as a more generalised extension of the WSS metric, plotting the scores over the full range of values of recall.

Previous studies suggested that the drawback of WSS is that its maximum value depends on the ratio of included to excluded samples (van Dinter, Catal et al., 2021, Kusa et al., 2022). They showed that for a perfectly balanced dataset the maximum value of WSS@95% is 0.45. At the same time, when the number of relevant documents is lower (as it is very common in the case of systematic reviews), the maximum value of WSS is higher. For instance, when the total number of relevant documents is 5% of all the documents, then the maximum value of WSS@95% is 0.90.

Evaluation of models using active learning was conducted with different measures that account for labelled and unlabelled samples. *Burden*, *utility* and *coverage* were introduced for evaluating active learning models in the context of citation screening (Wallace et al., 2010, Miwa et al., 2014). Burden represents the fraction of positive instances annotated manually by reviewers, whereas utility is a weighted sum of recall and (1-burden) using a β parameter, similar to Fβ-measure. Coverage indicates the ratio of positive instances in the data pool annotated during active learning.

3. Analysis of the work saved over sampling measure

In this section, we first present an example of the evaluation of automated citation screening with WSS. Later, we examine WSS properties and its terms and their influence on the final score.

3.1. Citation screening example

Let us assume an example systematic review with a citation list containing the total number of documents $N = 2000$. Out of them, only 200 are relevant to the systematic review study and should be included in the final review (also known as *includes*, $\mathcal{I}$). The remaining 1800 documents are irrelevant to the review topic and should be excluded (also known as *excludes*, $\mathcal{E}$). In a manual screening scenario, annotators need to screen all 2000 documents to select only the 200 relevant ones.

Fixing the level of recall also assumes that the number of true positives and false negatives is static. A recall of 95% is achieved when the model correctly predicts 190 relevant documents (TP). The remaining 10 includes are treated as false negatives (FN). In practice, different models vary from each other by how many excludes they can screen out automatically (i.e., good models maximise the number of $\mathcal{E}$ classified as true negatives (TN) while minimising the number of false positives (FP)). The WSS measure can be applied both to ranking (where the rank of $r\%$ relevant documents is used) and classification (where we a posteriori assume that the model used a specific prediction threshold to achieve the recall level of $r\%$.)

3.2. The $(1 - r)$ term

The $(1 - r)$ term was introduced to measure the advantage of a model when compared to the work saved with respect to a simple random sampling. A recall level of 95% is on average achieved when 95% of a dataset is randomly sampled, and this provides a 5% saving for reviewers. With the $(1 - r)$ term, the WSS@r% score above 0 means that a model performs better than the random sampling. If the WSS score is below 0, the model performs worse than random.

We argue that the $(1-r)$ term does not impact the WSS score as it was originally assumed, as it is just a constant value that is being subtracted from all scores from the same level $r\%$ of recall. In particular, for $r=0.95$, this term will always subtract 0.05 from the final WSS score, which can be seen as redundant if we want to compare multiple results.

3.3. The FN term

WSS at a specific $r\%$ recall assumes that exactly $(1-r)\%$ of documents that should be included will be misclassified. For a specific $r\%$ recall, the number of False Negatives (FN) is always equal to $\lfloor |\mathcal{I}| \cdot (1-r) \rfloor$, where with $\lfloor \cdot \rfloor$ we indicate the floor operator. This means that the FN term will also be a constant for every model for the same dataset. Consequently, for a fixed level of recall, true positives (TP) are equal to $r \cdot |\mathcal{I}|$.

Furthermore, the usage of the FN term in the WSS formula complicates its understanding. In the numerator (which should be maximised since the formula measures work saved), there is a sum of true negatives (a factor that should be maximised) and false negatives (a factor which should instead be minimised). A single evaluation measure should not maximise the sum of correct and wrong decisions simultaneously.

3.4. The maximum and minimum WSS value

For every dataset, we can calculate the maximum and minimum values of the WSS score as follows:

$$max(WSS@r\%) = \frac{|\mathcal{E}| + \lfloor |\mathcal{I}| \cdot (1-r) \rfloor}{N} - (1-r), \tag{2}$$

$$min(WSS@r\%) = \frac{0 + \lfloor |\mathcal{I}| \cdot (1-r) \rfloor}{N} - (1-r). \tag{3}$$

The maximum value of WSS is achieved when at least $r\%$ of included documents are presented first, before any irrelevant document (or in the classification nomenclature $TN = |\mathcal{E}|$). On the other hand, the minimum WSS value is obtained when all excluded documents are ranked before at least one relevant document ($TN = 0$).

The absolute maximum and minimum values of WSS depend on the dataset, and its excludes/includes ratio. $max(WSS)$ approaches 0 in datasets significantly imbalanced towards the positive class (includes):

$$\lim_{|\mathcal{E}|\to 0} max(WSS@r\%) = \lim_{|\mathcal{E}|\to 0} \frac{|\mathcal{E}| + |\mathcal{I}| \cdot (1-r)}{|\mathcal{E}| + |\mathcal{I}|} - (1-r) = 0. \tag{4}$$

On the other hand, as the ratio of irrelevant to relevant documents ($|\mathcal{E}|/|\mathcal{I}|$) gets higher, the maximum achievable score by WSS also gets higher (impact of includes in both nominator and denominator gets smaller, and the final score depends more on the excludes). Therefore, $max(WSS)$ approaches r in datasets heavily imbalanced towards the negative class (excludes):

$$\lim_{|\mathcal{I}|\to 0} max(WSS@r\%) = \lim_{|\mathcal{I}|\to 0} \frac{|\mathcal{E}| + |\mathcal{I}| \cdot (1-r)}{|\mathcal{E}| + |\mathcal{I}|} - (1-r) = r. \tag{5}$$

Similar considerations can be applied to $min(WSS)$, and its upper and lower bound also depends on the excludes/includes ratio. Moreover, $min(WSS)$ will not be negative only in the case when the dataset contains only documents that should be included ($|\mathcal{E}| = 0$). These properties of maximum and minimum values of WSS mean that this measure does not fulfil the zero and maximum Axiom #3 proposed by Busin and Mizzaro (2013).

3.5. Evaluation with cross-validation

Most of the automated citation screening models require some seed of manually labelled documents to train the machine learning model, which can rank or predict the category of remaining documents. This assumes preparation of the training set, i.e., manually annotating documents for their eligibility. In previous work, evaluation was usually done using stratified 5 × 2-fold cross-validation that splits the dataset into two equally sized subsets with an even distribution of label classes which are subsequently used to train and test the model (Matwin et al., 2010, Cohen, 2011, Howard et al., 2016, Kontonatsios et al., 2020, van Dinter, Catal et al., 2021, Kusa et al., 2022). The actual work saved would be measured on the second half of the initial dataset. Effectively, in the example dataset and when using 5 × 2-fold cross-validation, there would be total of $|\mathcal{N}| = 1000$ documents for the evaluation with WSS, out of which 100 includes $\mathcal{I}$ and 900 excludes $\mathcal{E}$.

This approach implies another practical consideration with the $(1-r)$ term in the WSS measure. If in the dataset the total number of includes $\mathcal{I}$ is small, such that for a specific level of recall r, $(1-r)\%$ of relevant items would be fewer than one document (i.e., $|\mathcal{I}| \cdot (1-r) < 1$), the number of false negatives will be equal to 0 for all recalls $\geq r$. Thus, the following equation holds:

$$WSS@r\% = WSS@100\% - (1-r). \tag{6}$$

This means that even when comparing WSS scores for different levels of recall r, they will differ only by the constant $(1-r)$ term, and it does not depend on the total number of documents N. For WSS@95%, the equation above is true for all datasets where the total number of relevant documents used in the evaluation is fewer than 20 ($|\mathcal{I}| < 20$). Moreover, when a common practice of using stratified 5 × 2-fold cross-validation is applied to evaluating a model, and one only calculates the scores on half of the dataset, this, in practice, means that the total size of includes in the dataset for which this equation holds is twice as high (40 relevant examples in the case of $r = 95\%$).

From our analysis of 23 commonly used benchmark datasets (Kusa et al., 2022), five have less than 40 includes in total (three of these datasets have even less than 20 includes). This means that there is no difference if one evaluates the same model at 95% or 100% recall, as these two scores will always only differ by 0.05 for the dataset considered.

4. The normalised WSS

As it was done in the case of the DCG metric (Järvelin & Kekäläinen, 2002), we propose to normalise the WSS metric. As for the nDCG, the normalised WSS will allow for comparison across multiple models and benchmark systematic review datasets. The approach is presented below:

$$nWSS@r\% = \frac{WSS@r\% - min(WSS@r\%)}{max(WSS@r\%) - min(WSS@r\%)} \tag{7}$$

With the assumptions from the previous section, we further formulate the equation as:

$$\begin{aligned}
&nWSS@r\% \\
&\quad = \frac{(TN + \lfloor |\mathcal{I}| \cdot (1-r) \rfloor)/N - \cancel{(1-r)} - \lfloor |\mathcal{I}| \cdot (1-r) \rfloor / N + \cancel{(1-r)}}{(|\mathcal{E}| + \lfloor |\mathcal{I}| \cdot (1-r) \rfloor)/N - \cancel{(1-r)} - \lfloor |\mathcal{I}| \cdot (1-r) \rfloor / N + \cancel{(1-r)}} \\
&\quad = \frac{(TN + \lfloor |\mathcal{I}| \cdot (1-r) \rfloor)/\cancel{N} - \lfloor |\mathcal{I}| \cdot (1-r) \rfloor / \cancel{N}}{(|\mathcal{E}| + \lfloor |\mathcal{I}| \cdot (1-r) \rfloor)/\cancel{N} - \lfloor |\mathcal{I}| \cdot (1-r) \rfloor / \cancel{N}} \\
&\quad = \frac{TN + \cancel{\lfloor |\mathcal{I}| \cdot (1-r) \rceil} - \cancel{\lfloor |\mathcal{I}| \cdot (1-r) \rceil}}{|\mathcal{E}| + \cancel{\lfloor |\mathcal{I}| \cdot (1-r) \rceil} - \cancel{\lfloor |\mathcal{I}| \cdot (1-r) \rceil}} \\
&\quad = \frac{TN}{|\mathcal{E}|}
\end{aligned} \tag{8}$$

Applying this normalization makes all the constant terms of WSS (FN and $(1-r)$) cancel themselves. The nWSS score for every dataset is always in the range $[0,1]$. An ideal score is achieved when all the excluded documents are classified as true negatives, and then the nWSS is equal to 1. Conversely, when all the documents that should be excluded are classified incorrectly, $TN = 0$ and thus nWSS = 0.

In the case of a recall threshold at 95%, the nWSS equation is:

$$nWSS@95\% = \frac{TN@95\%}{|\mathcal{E}|}, \tag{9}$$

meaning that we only need to estimate the number of true negatives produced by a ranking/classification model when it achieves 95% recall.

Furthermore, as $|\mathcal{E}|$ is equal to all the negatives that should be excluded, i.e., $|\mathcal{E}| = TN + FP$, this allows us to produce another version of the nWSS:

$$nWSS = \frac{TN}{TN + FP}, \tag{10}$$

which is equal to the True Negative Rate (TNR), also known as specificity. This means that nWSS@r% is equal to specificity at a recall rate of $r\%$ (S@r%).

$$nWSS@r\% = TNR@r\% = \frac{TN@r\%}{|\mathcal{E}|}. \tag{11}$$

4.1. Alternative demonstration for rank-based evaluation

Here we propose an alternative demonstration that uses rank-based evaluation terms. We assume that $n_{r\%}$ is the rank of the document in the ordered dataset, which is the last manually screened document in order to achieve $r\%$ of recall. $TN + FN$ is thus equal to $N - n_{r\%}$, and we can then re-write the WSS equations as follows:

$$WSS@r\% = \frac{TN + FN}{N} - (1-r) = \frac{N - n_{r\%}}{N} - (1-r). \tag{12}$$

In this equation, both N and r are fixed, and the only model and dataset-dependent parameter is $n_{r\%}$. The minimum value of WSS is when the rank is the lowest possible (only $(1-r)$ of relevant documents were still not seen): $n_{r\%} = N - (1-r) \cdot |\mathcal{I}|$. The maximum value of WSS is when the rank is equal to $r\%$ of relevant documents: $n_{r\%} = r \cdot |\mathcal{I}|$. We can then write the minimum as:

$$\begin{aligned}
min(WSS@r\%) &= \frac{N - (N - (1-r)\cdot|\mathcal{I}|)}{N} - (1-r)\\
min(WSS@r\%) &= \frac{(1-r)\cdot|\mathcal{I}|}{N} - (1-r),
\end{aligned} \tag{13}$$

and the maximum as:

$$\begin{aligned}
max(WSS@r\%) &= \frac{N - r\cdot|\mathcal{I}|}{N} - (1-r)\\
max(WSS@r\%) &= \frac{(|\mathcal{E}| + |\mathcal{I}|) - r\cdot|\mathcal{I}|}{N} - (1-r)\\
max(WSS@r\%) &= \frac{|\mathcal{E}| + (1-r)\cdot|\mathcal{I}|}{N} - (1-r).
\end{aligned} \tag{14}$$

We can then write the formula for normalised WSS@r% using document ranking terms:

$$\begin{aligned}
nWSS@r\% &= \frac{\frac{N-n_{r\%}}{N} - \cancel{(1-r)} - \frac{(1-r)\cdot|\mathcal{I}|}{N} + \cancel{(1-r)}}{\frac{|\mathcal{E}|+(1-r)\cdot|\mathcal{I}|}{N} - \cancel{(1-r)} - \frac{(1-r)\cdot|\mathcal{I}|}{N} + \cancel{(1-r)}}\\
nWSS@r\% &= \frac{\frac{N-n_{r\%}}{\cancel{N}} - \frac{(1-r)\cdot|\mathcal{I}|}{\cancel{N}}}{\frac{|\mathcal{E}|+(1-r)\cdot|\mathcal{I}|}{\cancel{N}} - \frac{(1-r)\cdot|\mathcal{I}|}{\cancel{N}}}\\
nWSS@r\% &= \frac{N - n_{r\%} - (1-r)\cdot|\mathcal{I}|}{|\mathcal{E}| + \cancel{(1-r)\cdot|\mathcal{I}|} - \cancel{(1-r)\cdot|\mathcal{I}|}}\\
nWSS@r\% &= \frac{N - n_{r\%} - (1-r)\cdot|\mathcal{I}|}{|\mathcal{E}|}\\
nWSS@r\% &= \frac{|\mathcal{E}| + r\cdot|\mathcal{I}| - n_{r\%}}{|\mathcal{E}|}.
\end{aligned} \tag{15}$$

Equation (15) is the rank-based version of the nWSS equation. Furthermore, if we substitute the rank-based terms with confusion matrix terms ($n_{r\%} = TP + FP$), we can show that this formula is identical to Equation (10):

$$\begin{aligned}
nWSS@r\% &= \frac{|\mathcal{E}| + r\cdot|\mathcal{I}| - n_{r\%}}{|\mathcal{E}|}\\
nWSS@r\% &= \frac{(TN + \cancel{FP}) + r\cdot|\mathcal{I}| - (TP + \cancel{FP})}{|\mathcal{E}|}\\
nWSS@r\% &= \frac{TN + r\cdot|\mathcal{I}| - TP}{|\mathcal{E}|}\\
nWSS@r\% &= \frac{TN + \cancel{TP} - \cancel{TP}}{|\mathcal{E}|}\\
nWSS@r\% &= \frac{TN}{TN + FP}.
\end{aligned} \tag{16}$$

5. Benchmark results with $TNR@95\%$

In this section, we calculate the TNR scores on previous benchmark results of citation screening datasets from Cohen et al. (2006). We used Equation (7) to convert WSS@95% scores reported by previous studies to the TNR@95% recall scores. The performance of past models evaluated with TNR@95%, together with averaged WSS@95% is presented in Table 1. Other researchers can use these scores to compare their models on the most popular citation screening benchmark collection.

Furthermore, compared with the average WSS scores aggregated from these 15 datasets, we can notice that the model's ordering changes when evaluated with averaged TNR. When ordered by their average WSS score, models from best to lowest score are D, E, C, G, **F**, **B** and A. However, when evaluated with TNR, the order is the following: D, E, C, G, **B**, **F** and A. Hence, with only seven models, we have already noticed that the incorrect usage of WSS to compare averaged performance across several datasets proved to yield erroneous order of models.

6. Discussion

6.1. Comparison with precision

Fig. 1 presents the dynamic of evaluation measures' scores as a function of the number of true negatives detected by an algorithm for a fixed recall level of 95%. We consider two types of datasets having the same total number of documents $N = 2000$ but differing in the $|\mathcal{I}|/|\mathcal{E}|$ ratio: heavily imbalanced towards the negative class with only 5% of positive examples (Fig. 1a), and perfectly balanced dataset (Fig. 1b). On both datasets, WSS and TNR scores rise linearly with the rising number of true negatives detected by the algorithm, but a change in the Precision scores is not linear, and its derivative depends on the class imbalance. In addition, out of these three measures, only TNR is always bounded by 0 and 1. Again, minimum Precision value depends on the class imbalance, which for WSS is the case for both minimum and maximum values.

TNR score can also be directly translated to the number of documents reviewers do not need to screen manually. Furthermore, when used with appropriate multipliers, assuming all documents are equal, one can convert the TNR score into the time and money saved by using automation tools.

6.2. Comparison with AUC

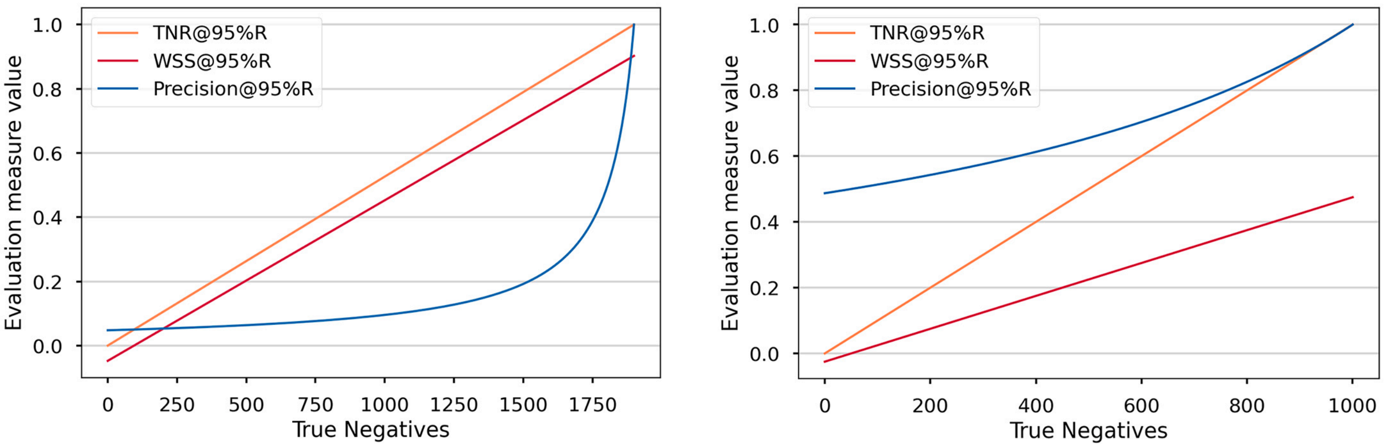
As already mentioned, measures like ROC or Precision-Recall curve are more suitable for comparing a model's effectiveness across multiple recall levels. However, they do not allow for automatic comparisons across multiple models and are not suitable for score aggregations across several datasets. Fawcett (2006) mentions that even though ROC curves may be used to evaluate classifiers, care should be taken when using them to conclude classifier superiority.

Fig. 2 presents ROC curves and corresponding AUC scores for two hypothetical models on the same dataset. Model A, which obtains a higher AUC score, quickly achieves >60% recall, but its score plateaus and only manages to exceed recall of 80% at the very end. On the other hand, model B, which "struggles" initially but reaches perfect recall at

Table 1
Evaluation results with WSS and TNR at 95% recall on systematic review datasets from Cohen et al. (2006). The following models were used: A: Cohen et al. (2006), B: Matwin et al. (2010), C: Cohen (2008), D: Howard et al. (2016), E: Kontonatsios et al. (2020), F: van Dinter, Catal et al. (2021), G: Kusa et al. (2022). **Bold** indicates highest score.

No	Dataset name	Dataset size	Percentage of includes	Models						
WSS@95%				A	B	C	D	E	F	G
1	ACEInhibitors	2544	1.6%	0.566	0.523	0.733	**0.801**	0.787	0.783	0.783
2	ADHD	851	2.4%	0.680	0.622	0.526	**0.793**	0.665	0.698	0.424
3	Antihistamines	310	5.2%	0.000	0.149	0.236	0.137	**0.310**	0.168	0.047
4	Atypical Antipsychotics	1120	13.0%	0.141	0.206	0.170	0.251	**0.329**	0.212	0.218
5	Beta Blockers	2072	2.0%	0.284	0.367	0.465	0.428	**0.587**	0.504	0.419
6	Calcium Channel Blockers	1218	8.2%	0.122	0.234	0.430	**0.448**	0.424	0.159	0.178
7	Estrogens	368	21.7%	0.183	0.375	0.414	**0.471**	0.397	0.119	0.306
8	NSAIDs	393	10.4%	0.497	0.528	0.672	**0.730**	0.723	0.571	0.620
9	Opioids	1915	0.8%	0.133	0.554	0.364	**0.826**	0.533	0.295	0.559
10	Oral Hypoglycemics	503	27.0%	0.090	0.085	**0.136**	0.117	0.095	0.065	0.098
11	Proton PumpInhibitors	1333	3.8%	0.277	0.229	0.328	0.378	**0.400**	0.243	0.283
12	Skeletal Muscle Relaxants	1643	0.5%	0.000	0.265	0.374	**0.556**	0.286	0.229	0.090
13	Statins	3465	2.5%	0.247	0.315	0.491	0.435	**0.566**	0.443	0.409
14	Triptans	671	3.6%	0.034	0.274	0.346	0.412	**0.434**	0.266	0.210
15	Urinary Incontinence	327	12.2%	0.261	0.296	0.432	0.531	**0.531**	0.272	0.439
	Average WSS@95% score			0.2343	0.3348	0.4078	**0.4876**	0.4711	0.3351	0.3388
	Rank based on average score			7	6	3	1	2	5	4

TNR@95%				A	B	C	D	E	F	G
1	ACEInhibitors			0.625	0.582	0.795	**0.864**	0.850	0.846	0.846
2	ADHD			0.746	0.687	0.589	**0.862**	0.731	0.765	0.484
3	Antihistamines			0.053	0.210	0.302	0.197	**0.380**	0.230	0.102
4	Atypical Antipsychotics			0.212	0.287	0.246	0.339	**0.429**	0.294	0.301
5	Beta Blockers			0.340	0.425	0.525	0.487	**0.649**	0.564	0.478
6	Calcium Channel Blockers			0.183	0.305	0.518	**0.538**	0.512	0.223	0.244
7	Estrogens			0.284	0.529	0.579	**0.652**	0.557	0.202	0.441
8	NSAIDs			0.605	0.640	0.800	**0.865**	0.857	0.688	0.742
9	Opioids			0.184	0.609	0.417	**0.883**	0.588	0.348	0.614
10	Oral Hypoglycemics			0.176	0.169	**0.239**	0.213	0.182	0.141	0.186
11	Proton PumpInhibitors			0.338	0.289	0.391	0.443	**0.466**	0.303	0.345
12	Skeletal Muscle Relaxants			0.050	0.317	0.426	**0.609**	0.338	0.281	0.141
13	Statins			0.303	0.373	0.553	0.496	**0.630**	0.504	0.469
14	Triptans			0.086	0.334	0.409	0.478	**0.500**	0.326	0.268
15	Urinary Incontinence			0.347	0.387	0.542	**0.655**	**0.655**	0.360	0.550
	Average TNR@95% score			0.3022	0.4094	0.4888	**0.5721**	0.5550	0.4050	0.4141
	Rank based on average score			7	5	3	1	2	6	4



(a) Evaluation measures' scores versus the number of True Negatives for an imbalanced dataset with 5% of positive examples ($|\mathcal{I}| = 100$, $|\mathcal{E}| = 1900$).

(b) Evaluation measures' scores versus the number of True Negatives for a perfectly balanced dataset ($|\mathcal{I}| = |\mathcal{E}| = 1000$).

Fig. 1. Dynamics of evaluation measures (WSS, TNR (nWSS) and Precision) scores as a function of the number of True Negatives (TN) at 95% recall for two sample datasets.

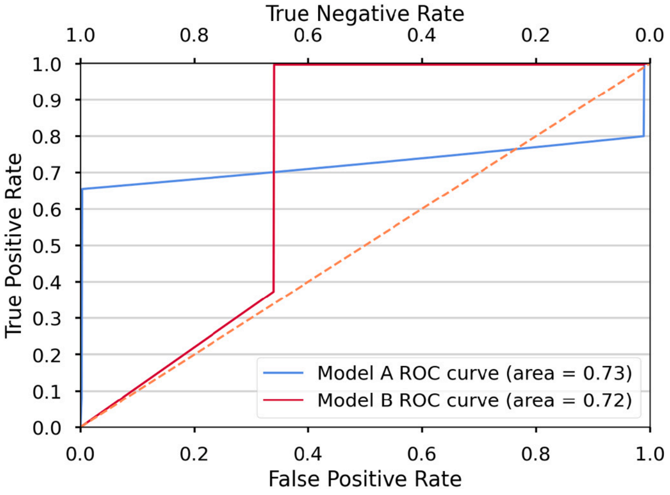


Fig. 2. Receiver Operating Characteristic (ROC) curves for two hypothetical models with their corresponding AUC scores. Model A achieves a higher value of AUC, despite the fact that its TPR performance reaches 80% only at the FPR level almost equal to 100%, and model B achieves maximum recall at FPR level of 35%.

an FPR level of 0.35, obtains a lower AUC score. For the general search task, model A might be more suitable. However, for technology-assisted reviews where we want to ensure that the model achieves very high recall (and even in the case of rapid reviews or e-discovery, this should very rarely be lower than 70%), model B is the only one which delivers some gain to the user.

Hence, we believe that compared to TNR, AUC scores can favour models that achieve good recall scores at low values of FPR, which are of no value for citation screening tasks. An alternative can be to calculate partial AUC score (pAUC), a practice for highly sensitive diagnostic tests (McClish, 1989, Jiang et al., 1996). Similarly to the TNR@r% and WSS@r% calculations, one could parameterise AUC by the desired minimum recall (TPR) level. Then, the pAUC is computed in the part of the ROC space where the recall is greater than a given threshold r.

6.3. Limitations

Work Saved over Sampling cannot account for the amount of manual work required to kick-start the automated screening. Current classification approaches use some type of cross-validation to train and evaluate their models. Usage of different train/test splits provides another challenge as TNR@r% (unlike burden or utility) does not measure the amount of data that needs to be labelled manually before training the classifier. To overcome this problem, plotting the learning curves for TNR@r% could be one way to compare the performance of these models.

Alternatively, a set of standard benchmarks with fixed train/test splits would need to be introduced. Future work will focus on this aspect and the applicability of the measure in the active learning scenario.

7. Conclusions

This paper analyses Work Saved over Sampling (WSS), a measure commonly used to evaluate automated citation screening models. We inspect the terms and properties of WSS and show drawbacks of the measure.

We propose min-max normalisation of Work Saved over Sampling at r% recall (nWSS@r%). It improves on the commonly used WSS measure as it normalises possible scores into the $[0, 1]$ range. This enables fair comparison between different models and score aggregations from multiple datasets. nWSS also simplifies over WSS as it does not contain two WSS terms that were shown to be constants by our analysis. Moreover, we show that nWSS is equal to True Negative Rate (TNR), further simplifying the understanding of the measure.

TNR has a linear correlation with the number of documents that a manual reviewer does not need to screen and can be directly translated to the time (and money) saved when using automation tools. We suggest the usage of TNR at r% of recall as an evaluation measure for the citation screening task if the score is to be compared between multiple models across several datasets.

CRediT authorship contribution statement

Wojciech Kusa: Conceptualization, Formal analysis, Methodology, Writing – original draft. **Aldo Lipani:** Formal analysis, Supervision, Validation, Writing – review & editing. **Petr Knoth:** Supervision, Validation, Writing – review & editing. **Allan Hanbury:** Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

This work was supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval – DoSSIER (H2020-EU.1.3.1., ID: 860721).

References

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S. C., Ananiadou, S., Liao, J., & Macleod, M. R. (2019). Machine learning algorithms for systematic review: Reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews*, *8*, Article 23. https://doi.org/10.1186/S13643-019-0942-7.

Busin, L., & Mizzaro, S. (2013). Axiometrics: An axiomatic approach to information retrieval effectiveness metrics. In *Proceedings of the 2013 conference on the theory of information retrieval* (pp. 22–29).

Cohen, A. M. (2008). Optimizing feature representation for automated systematic review work prioritization. In *AMIA annual symposium proceedings* (p. 121). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656096/.

Cohen, A. M. (2011). Letter: Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@95 measure. *Journal of the American Medical Informatics Association*, *18*, 104. https://doi.org/10.1136/JAMIA.2010.008177. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3005879/.

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*, *13*, 206–219. https://doi.org/10.1197/jamia.M1929. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1447545/.

Cormack, G. V., & Grossman, M. R. (2016). Engineering quality and reliability in technology-assisted review. In *SIGIR 2016 - proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval* (pp. 75–84).

Cormack, G. V., & Grossman, M. R. (2017). Technology-assisted review in empirical medicine: Waterloo participation in CLEF eHealth 2017. In *CLEF (working notes)*.

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 Dinter, R., Tekinerdogan, B., & Catal, C. (2021). Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology*, *136*, Article 106589. https://doi.org/10.1016/j.infsof.2021.106589.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.

Howard, B. E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M. R., Holmgren, S., Pelch, K. E., Walker, V., Rooney, A. A., Macleod, M., Shah, R. R., & Thayer, K. (2016). SWIFT-review: A text-mining workbench for systematic review. *Systematic Reviews*, *5*, 1–16. https://doi.org/10.1186/s13643-016-0263-z.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*, 422–446. https://doi.org/10.1145/582415.582418.

Jiang, Y., Metz, C. E., & Nishikawa, R. M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*, *201*, 745–750.

Kanoulas, E., Li, D., Azzopardi, L., & Spijker, R. (2017). CLEF 2017 technologically assisted reviews in empirical medicine overview. *CEUR Workshop Proceedings*, *1866*, 1–29. https://pureportal.strath.ac.uk/en/publications/clef-2017-technologically-assisted-reviews-in-empirical-medicine-.

Kanoulas, E., Li, D., Azzopardi, L., & Spijker, R. (2018). CLEF 2018 technologically assisted reviews in empirical medicine overview. *CEUR Workshop Proceedings*, *2125*. https://pureportal.strath.ac.uk/en/publications/clef-2018-technologically-assisted-reviews-in-empirical-medicine-.

Kontonatsios, G., Spencer, S., Matthew, P., & Korkontzelos, I. (2020). Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews. *Expert Systems with Applications: X*, *6*, Article 100030. https://doi.org/10.1016/j.eswax.2020.100030.

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.

Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O., & O'Blenis, P. (2010). A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association*, *17*, 446–453. https://doi.org/10.1136/JAMIA.2010.004325.

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making*, *9*, 190–195.

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. https://doi.org/10.1016/J.JBI.2014.06.005.

Norman, C. (2020). Systematic review automation methods. Ph.D. thesis, Université Paris-Saclay; Universiteit van Amsterdam. https://tel.archives-ouvertes.fr/tel-03060620.

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*, 5. https://doi.org/10.1186/2046-4053-4-5.

Scells, H., Zuccon, G., & Koopman, B. (2019). Automatic Boolean query refinement for systematic review literature search. In *The web conference 2019 - proceedings of the world wide web conference, WWW 2019: Vol. 11* (pp. 1646–1656).

Tsafnat, G., Glasziou, P., Karystianis, G., & Coiera, E. (2018). Automated screening of research studies for systematic reviews using study characteristics. *Systematic Reviews*, *7*, 1–9. https://doi.org/10.1186/S13643-018-0724-7.

Wallace, B. C., Small, K., Brodley, C. E., & Trikalinos, T. A. (2010). Active learning for biomedical citation screening. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 173–181).