Active learning for efficient systematic reviews

Evaluating models across research areas

Gerbrich Ferdinands

20 July, 2020

Abstract

Background the context and purpose of the study

Methods how the study was performed and statistical tests used

Results the main findings

Conclusions brief summary and potential implications

keywords:

Background

Systematic reviews are top of the bill in research. A systematic review brings together all studies relevant to answer a specific research question [1]. Systematic reviews inform practice and policy [2] and are key in developing clinical guidelines [3]. However, systematic reviews are costly because they involve the manual screening of thousands of titles and abstracts to identify publications relevant to answering the research question.

Conducting a systematic review typically requires over a year of work by a team of researchers [4]. Nevertheless, systematic reviewers are often bound to a limited budget and timeframe. Currently, the demand for systematic reviews exceeds the available time and resources by far [5]. Especially when the need for guidelines is urgent it is extremely challenging to provide a review that is both timely and comprehensive.

To ensure a timely review, reducing workload in systematic reviews is essential. With advances in Machine Learning (ML), there has been wide interest in tools to reduce workload in systematic reviews [6]. Various learning models have been proposed, aiming to predict whether a given publication is relevant or irrelevant to the systematic review. Previous findings suggest that such models potentially reduce workload with 30-70% at the cost of losing 5% of relevant publications, i.e. 95% recall [7].

A well-established approach in increasing efficiency in title and abstract screening is screening prioritization [8,9]. In screening prioritization, the learning model presents the reviewer with the publications that are most likely to be relevant first, thereby expediting the process of finding all of the relevant publications. Such an approach allows for substantial time-savings in the screening process as the reviewer can decide to

stop screening after a sufficient number of relevant publications have been retrieved [10]. Moreover, the early retrieval of relevant publications facilitates a faster transition of those publications to the next steps in the review process [8].

Recent studies have demonstrated the effectiveness of screening prioritization by means of active learning models [10–16]. With active learning, the ML model can iteratively improve its predictions on unlabelled data by allowing the model to select the records from which it wants to learn [17]. The model queries these records to a human annotator who provides them with a label, from which the model then updates its predictions. The general assumption is that by letting the model select which records are labelled, the model can achieve higher accuracy more quickly while requiring the human annotator to label as few records as possible [18]. Active learning has proven to be an efficient strategy in large unlabelled datasets where labels are expensive to obtain [18]. This makes the screening phase in systematic reviewing an ideal candidate solution for such models because typically, labelling a large number of publications is very costly. When active learning is applied in the screening phase, the reviewer screens publications that are suggested by an active learning model. Subsequently, the active learning model learns from the reviewers' decision ('relevant', 'irrelevant') and uses this knowledge to update its predictions and to select the next publication to be screened by the reviewer.

The application of active learning models in reducing workload of systematic reviews has been extensively studied [10–12,15,16]. Whilst previous studies have evaluated active learning models in many forms and shapes [10–15,19–21], ready-to-use software tools implementing such active learning models (Abstrackr [22], Colandr [23], FASTREAD [11], Rayyan [24], and RobotAnalyst [25]) currently use the same classification technique to predict relevance of publications, namely Support Vector Machine (SVM). Findings show that different classification techniques can serve different needs in the retrieval of relevant publications, for example the desired balance between recall and precision [26,27]. Therefore, is essential to evaluate different classification techniques in the context of active learning models. The current study investigates active learning models adopting four classification techniques: Naive Bayes, Logistic Regression, Support Vector Machine, and Random Forest. These are widely adopted techniques in text classification [28] and are suitable for creating interpretable ML models [29]. Moreover, due to their relatively short computation time they are fit for software tools to be used in scientific practice, as they allow the reviewer to be timely presented with the next publication.

Another component that influences model performance is the way how the textual content of titles and abstracts are represented in a model, called the feature extraction strategy [19,20,30]. One of the more sophisticated feature extraction strategies is called Doc2vec (D2V), also known as Paragraph vector [31]. D2V

learns continuous distributed vector representations for pieces of text. The idea of distributed representation is that every word can be known by its neighborhood. D2V can overcome problems with more simplistic feature extraction strategies such as TF-IDF, which ignores the the semantics and the ordering of the words [31]. Therefore, it is of interest to compare models adopting D2V to models adopting TF-IDF.

Lastly, previous studies have mainly focussed on reviews from a single scientific field, like medicine [15,16] and computer science [11]. Model replications on reviews from varying research contexts are essential to draw conclusions about the general effectiveness of active learning models [7,32]. As far as known to the authors, Miwa et al [12] were the only researchers to make a direct comparison between two systematic reviews from different research areas, namely the social and the medical sciences. They found that active learning was more difficult on data from the social sciences due to the different nature of the vocabularies used. Thus, it is of interest to evaluate model performance across different research contexts, namely the social science, humanities, medical science and computer science.

Taken together, for a more comprehensive understanding of active learning models in the context of systematic reviewing, a methodical evaluation of such models is required. The current study aims to address this issue by answering the following research questions:

RQ1 What is the performance of active learning models across four different classification techniques?

RQ2 What is the performance of active learning models across two different feature extraction strategies?

RQ3 Does the performance of active learning models differ across six systematic reviews from four different research areas?

The purpose of the current paper is to increase the evidence base on active learning models for reducing workload in title and abstract screening in systematic reviews. We adopt four different classification techniques (NB, LR, SVM, and RF) and two different feature extraction strategies (TF-IDF and D2V) for the purpose of maximizing the number of identified relevant publications, while minimizing the number of publications needed to screen. Models were assessed by conducting a simulation on six systematic review datasets. Datasets were collected from the fields of medical science [33–35], computer science [11], humanities [36] and social science [37], to assess generalizability of the models across research contexts. The models, datasets and simulations are implemented in a pipeline of active learning for screening prioritization, called `ASReview` [38]. `ASReview` is an open source and generic tool in which users can adapt and add modules as they like, encouraging fellow researchers to replicate findings from previous studies. All scripts and data used are openly published to facilitate usability and acceptability of ML-assisted title and abstract screening in the field of systematic review.

The remaining part of this paper is organized as follows. The methods section describes the study that was designed to answer the research questions and elaborates on the technical details of active learning models for identifying relevant publications in the context of systematic reviews. The findings of the simulation study are reported in the results section. The implications of the findings in context of previous research are discussed in the discussion section, followed by this study's main conclusions in the conclusion section.

Methods

Technical details

What follows is a more detailed account of the active learning models. The structure and functions of the key components of the models are introduced to clarify the choices made in the design of the current study.

Task description

The screening process of a systematic review starts with all publications obtained in the search. The task is to identify which of these publications are relevant, by screening their titles and abstracts. In active learning for screening prioritization, the screening process proceeds as follows:

- Start with the set of all unlabelled records (titles and abstracts), $\mathcal{U}$.
- The reviewer provides a label for a few, for example 5-10 records $x \in \mathcal{U}$, creating a set of labelled records $x \in \mathcal{L}$ such that $x \notin \mathcal{U}$. The label can be either Relevant $\langle x, \texttt{R} \rangle$ or Irrelevant $\langle x, \texttt{I} \rangle$.
- The active learning cycle starts:
    1. A classifier C is trained on the labelled records $\mathcal{L}$, $C = \texttt{train}(\mathcal{L})$
    2. The classifier predicts relevancy scores for all unlabelled records $\mathcal{U}$, $C(\mathcal{U})$
    3. Based on the predictions by C, the model selects the most relevant record $x^* \in \mathcal{U}$
    4. The model queries the reviewer to screen this record, $\langle x^*, \texttt{?} \rangle$
    5. The reviewer screens the record and provides a label, $\langle x^*, \texttt{R} \rangle$ or $\langle x^*, \texttt{I} \rangle$
    6. The newly labelled record is added to the training data, such that $x^* \in \mathcal{L}$ and $x^* \notin \mathcal{U}$
    7. Back to step 1

8. Repeat this cycle until the reviewer decides to stop [10] or until all records have been labelled, $\mathcal{U} = \varnothing$.

In this active learning cycle, the model incrementally improves its predictions on the remaining unlabelled title and abstracts. Relevant titles and abstracts are identified as early in the process as possible.

This case is an example of pool-based active learning, as the next record to be queried is selected by predicting relevancy for all records in a fixed pool [17]. Another form of active learning is stream-based active learning, in which data is regarded as a stream instead of a fixed pool, in which the model selects one record at a time and then decides whether or not to query this record. This approach of active learning is preferred when it is expensive or impossible to exhaustively search the data for selecting the next query. A possible application of stream-based active learning is living systematic reviews, as the review is continually updated as new evidence becomes available. For an example see the study by Wynants et al., [39].

Class imbalance problem

There are two classes in the dataset: relevant and irrelevant publications. Typically, the inclusion rate is low as only a fraction of the publications belong to the relevant class (2.94%, [4]). The class imbalance causes the classifier to miss relevant publications, because there are far fewer examples of relevant than irrelevant publications to train on [7]. Moreover, classifiers can achieve high accuracy but still fail to identify any of the relevant publications [15]. This is evident in the case of a systematic review dataset where only three percent of publications are relevant.

Previous studies have addressed the class imbalance problem by rebalancing the training data in various ways [7]. To decrease the class imbalance in the training data, the models in the current study rebalance the training set by a technique we propose to call Dynamic Resampling (DR). DR undersamples the number of irrelevant publications in the training data, whereas the number of relevant publications are oversampled such that the size of the training data remains the same. The ratio between relevant and irrelevant publications in the rebalanced training data is not fixed, but dynamically updated depending on the number of publications in the available training data, the total number of publications in the dataset, and the ratio between relevant and irrelevant publications in the available training data.

Classification

To make predictions on the unlabelled publications, a classifier is trained on features from the training data. The performance of the following four classifiers is explored:

- Support Vector Machine (SVM) - SVMs separate the data into classes by finding a multidimensional hyperplane [40,41].
- L2-regularized Logistic Regression (LR) models the probabilities describing the possible outcomes by a logistic function. The coefficients are regularizes/model is regularized. A penalty is imposed on the size of the coefficients, shrinking coefficients of features with a minor contribution to the solution towards zero.
- Naive Bayes (NB) is a supervised learning algorithm often used in text classification. Based on Bayes' theorem, with the 'naive' assumption that all features are independent given the class value [42].
- Random Forests (RF) is a supervised learning algorithm where a large number of decision trees are fit on bootstrapped samples of the original data. All trees cast a vote on the class, which are aggregated into a class prediction for each instance [43].

Feature extraction

To predict relevance of a given publication, the classifier uses information from the publications in the dataset. Examples of such information are titles and abstracts. However, a model cannot make predictions from the titles and abstracts as they are; their textual content needs to be represented numerically. The textual information needs to be mapped to feature vectors. This process of numerically representing textual content is referred to as 'feature extraction'.

Term-Frequency Inverse Document Frequency (TF-IDF) stores the term frequency - the number of occurrences of each word - for each text in the dataset. The term frequency is adjusted with the inverse document frequency, the number of documents which contain the given word [44]. A disadvantage of TF-IDF and other bag-of-words methods is that they do not take the ordering of words into account, thereby ignoring semantics. An example of an approach that aims to overcome this weakness is D2V. D2V learns continuous distributed vector representations for pieces of text capable of grasping semantics by learning to predict the words in the texts [31].

Query strategy

The active learning model can adopt different strategies in selecting the next publication to be screened by the reviewer. A strategy mentioned before is selecting the publication with the highest probability of being relevant. In the active learning literature this is referred to as certainty-based active learning [17].

Another well-known strategy is uncertainty-based active learning, where the instances that are presented next are those instances on which the model's classifications are the least certain, i.e. close to 0.5 probability [17]. Traditionally, this strategy trains the most accurate model because the model can learn the most from instances it is uncertain about. However, a study comparing the performance of both strategies in detecting relevant publications found that the accuracy gain of uncertainty-based screening was not significant [12].

Certainty-based active learning is the preferred strategy for the task at hand. Firstly, the aim of the screening process is to find relevant publications, not to train a good model. Secondly, certainty-based active learning is far better equipped at dealing with imbalanced data in active learning, as it aims to present only records that belong to the relevant class [45].

Simulation study

The section below describes the simulation study that was carried out to answer the research questions.

Set-up

To address **RQ1**, four models combining every classifier with TF-IDF feature extraction were investigated:

1. SVM + TF-IDF
2. NB + TF-IDF
3. RF + TF-IDF
4. LR + TF-IDF

To address **RQ2**, the classifiers were combined with D2V feature extraction, leading to the following three models:

5. SVM + D2V
6. RF + D2V
7. LR + D2V

The combination NB + D2V could not be tested because the multinomial naive Bayes classifier[1] can only handle a feature matrix with positive values, whereas the D2V feature extraction approach[2] produces a

[1] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html#sklearn.naive_bayes.MultinomialNB

[2] https://radimrehurek.com/gensim/models/doc2vec.html

feature matrix that can also contain negative values. The performance of the seven models was evaluated by simulating every model on six systematic review datasets, addressing **RQ3**. Hence, 42 simulations were carried out, representing all model-dataset combinations.

Instead of having a human reviewer label publications manually, the screening process was simulated by retrieving the labels in the data. Each simulation started with an initial training set of one relevant and one irrelevant publication to represent a 'worst case scenario' where the reviewer has minimal prior knowledge on the publications in the data. The model was retrained every time after a publication had been labelled. A simulation ended after all publications in the dataset had been labelled. To account for sampling variance, each simulation was repeated 15 times. To account for bias due to the content of the initial publications, the initial training set was randomly sampled from the dataset for each of the 15 trials. Although varying over trials, the 15 initial training sets were kept constant over datasets to allow for a direct comparison of models within datasets. A seed value was set to ensure reproducibility. The simulation study was carried out using the `ASReview` simulation extension [46]. For each simulation, hyperparameters were optimized through a ... tpe to arrive at maximum model performance. todo Scripts for the hyperparaemter optimization and resulting hyperparameter values can be found at GitHub.

Simulations were carried out in ASReview, version 0.9.3 [46]. Analyses were carried out using `R`, version 3.6.1 [47]. Scripts and data are stored in the GitHub repository for this paper[3].The output resulting from the simulation was stored on the Open Science Framework page of this paper,[4] as their size exceeded the storage limit of GitHub by far. Due to their large number, the simulations were carried out on Cartesius, the Dutch national supercomputer. Access was granted by SURF via a grant (ID EINF-156).

Datasets

The models were simulated on a convenience sample of six systematic review datasets. The data selection process was driven by two factors. Firstly, datasets were collected from various research areas to assess generalizability of the models across research contexts (RQ3). Secondly, all original data files should be openly published with a CC-BY license.

todo to different section. and are available through the `ASReview` GitHub page.

The models are simulated on a convenience sample of six systematic review datasets. The data selection process was driven by two factors. Firstly, datasets are collected from various research areas to assess

[3]https://github.com/asreview/paper-evaluating-models-across-research-areas
[4]https://osf.io/7mr2g/

generalizability of the models across research contexts (RQ3). Secondly, all original data files have to be openly published with a CC-BY license. Datasets are available through this paper's GitHub page.

The Wilson dataset [48] - from the field of *medicine* - is on a review on the effectiveness and safety of treatments of Wilson Disease, a rare genetic disorder of copper metabolism [34]. From the same scientific field, the ACE dataset contains publications on the efficacy of Angiotensin-converting enzyme (ACE) inhibitors, a drug treatment for heart disease [33]. Additionally, the Virus dataset is from a systematic review on studies that performed viral Metagenomic Next-Generation Sequencing (mNGS) in farm animals [35]. From the field of *computer science*, the Software dataset contains publications from a review on fault prediction in software engineering [49]. The Nudging dataset [50] belongs to a systematic review on nudging healthcare professionals [36], stemming from the *humanities*. From *social science*, the PTSD dataset contains publications on studies applying latent trajectory analyses on posttraumatic stress after exposure to traumatic events [37]. Of these six datasets, ACE and Software have been used for model simulations in previous studies on ML-aided title and abstract screening [11,33].

Data were preprocessed from their original source into a dataset, containing title and abstract of the publications obtained in the initial search. Candidate studies with missing abstracts and duplicate instances were removed from the data. Preprocessing scripts and resulting datasets can be found at GitHub repository for this paper. Datasets were labelled to indicate which candidate studies were included in the systematic review, thereby indicating relevant publications. All datasets consisted of thousands of candidate studies, of which only a fraction was deemed relevant to the systematic review. For the Virus and the Nudging dataset, the inclusion rate was about 5 percent. For the remaining six datasets, inclusion rates were centered around 1-2 percent. (Table 1).

Table 1: Statistics on the datasets obtained from six original systematic reviews.

dataset	Candidate publications	Relevant publications	Inclusion rate (%)
Nudging	1,847	100	5.4
PTSD	5,031	38	0.8
Software	8,896	104	1.2
ACE	2,235	41	1.8
Virus	2,304	114	5.0
Wilson	2,333	23	1.0

Evaluating performance

Model performance was assessed by three different measures, Work Saved over Sampling (WSS), Relevant References Found (RRF), and Average Time to Discovery (ATD). WSS indicates the reduction in publications needed to be screened, at a given level of recall [33]. Typically measured at a recall level of 0.95, WSS@95 yields an estimate of the amount of work that can be saved at the cost of failing to identify 5% of relevant publications. In the current study, WSS is computed at 0.95 recall. RRF@10 represents the proportion of relevant publications that are found after screening 10% of all publications.

Both RRF and WSS are sensitive to the position of the cutoff value. Moreover, WSS makes assumptions about acceptable recall levels whereas this level might depend on the research question at hand [7]. Therefore, we introduce the ATD, the average proportion of publications needed to screen to find a relevant publication. The ATD is an indicator of performance throughout the entire screening process instead of performance at some arbitrary cutoff value.

Furthermore, model performance was visualized by plotting recall curves. Plotting recall as a function of the proportion of screened publications offers insight in model performance throughout the entire screening process [11,13]. The curves give information in two directions. On the one hand they display the number of publications that need to be screened to achieve a certain level of recall, but on the other hand they present how many relevant publications are identified after screening a certain proportion of all publications (RRF).

For each simulation, the RRF@10, WSS@95, and ATD are reported as means over 15 trials. To indicate the spread of performance within simulations, the means are accompanied by an estimated[5] standard devation $\hat{s}$. To compare the overall performance across datasets, median performance is reported for every dataset, accompanied by the Median Absolute Deviation (MAD), indicating variability between models within a certain dataset. Recall curves are plot for each simulation, representing the average recall over 15 trials $\pm$ the standard error of the mean.

Results

This section proceeds as follows: Firstly, the results of the Nudging dataset are discussed in detail to provide a basis for answering the research questions. Secondly, the results are presented for each research question over all datasets.

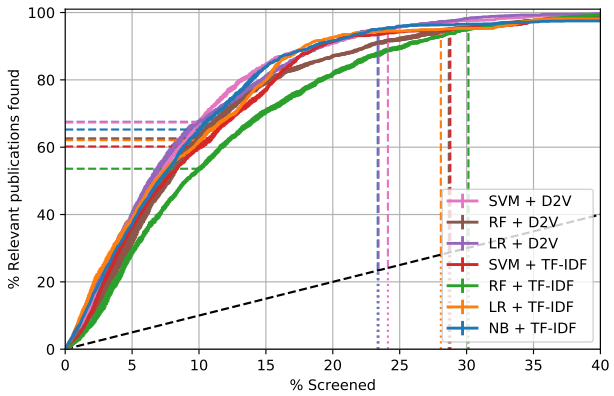
[5]The metrics for all individual 15 trials deviate slightly from the overal mean over 15 trials because of pre-averaging in the `ASReview` source code. As the analyses across all trials did not produce information on the 15 separate runs, the standard deviation of the mean, $\hat{s}$, was estimated by computing the standard deviation within the individual 15 trials.

Evaluation on the Nudging dataset

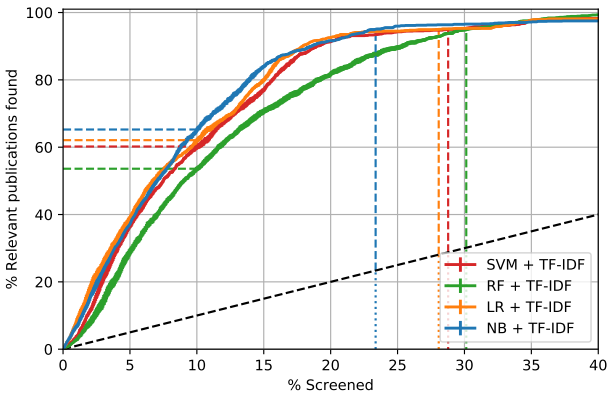
Figure 1a shows the recall curves for all simulations on the Nudging dataset. As desribed in the previous section, these curves plot recall as a function of the proportion of publications screened. The curves represent the average recall over 15 trials ± the standard error of the mean in the direction of the y-axis. The x-axis is cut off at 40% since at this point in screening all models had already reached 95% recall. The dashed horizontal lines indicate the RRF@10 values, the dashed vertical lines the WSS@95 values. The dashed grey diagonal line corresponds to the expected recall curve when publications are screened in a random order. Desirable model performance was defined as maximizing recall while minimizing the number of publications needed to screen.

The recall curves were used to examine model performance throughout the entire screening process and to make a visual comparison between models within datasets. For example in Figure 1a, after screening about 30% of the publications all models had already found 95% of the relevant publications. Moreover, after screening 5% the green curve - representing the RF + TF-IDF model - splits away from the others and remains to be the lowest of all curves until about 30% of publications have been screened. Hence, from screening 5 to 30 percent of publications, the RF + TF- IDF model was the slowest in finding the relevant publications. The ordering of the remaining recall curves changes throughout the screening process, but maintain to show relatively similar performance at face value.

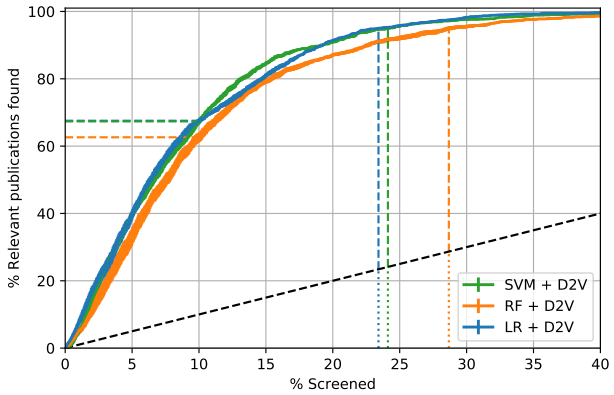
Figure 1b shows a subset of the recall curves in Figure 1a, namely the curves for the first four models only to allow for a visual comparison across classification techniques adopting the TF-IDF feature extraction strategy. Figure 1c shows recall curves for the remaining three models to compare the models using D2V feature extraction. Figures 1d to 1f plot recall curves for models adopting the TF-IDF feature extraction strategy to recall curves for their D2V-using counterparts to allow for a comparison between models adopting TF-IDF and models adopting D2V feature extraction.



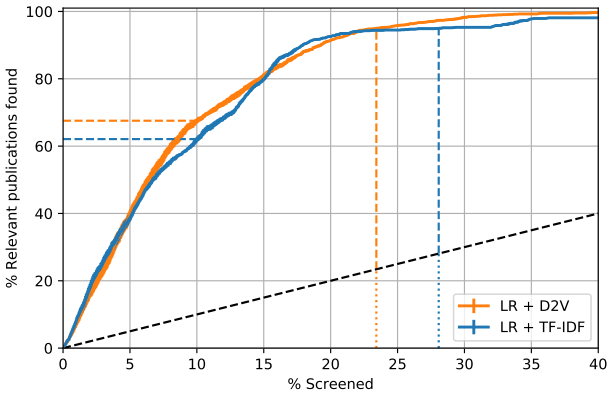
(a) All seven models



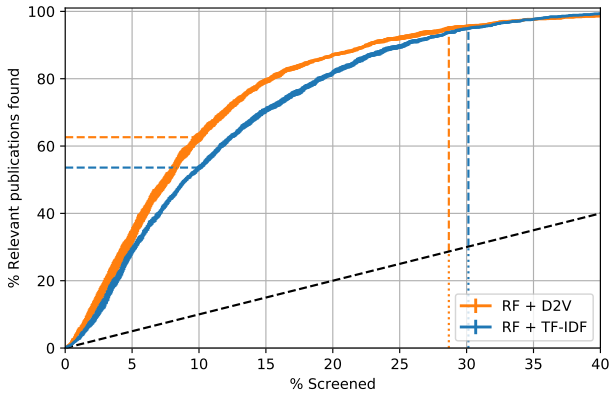
(b) Models adopting TF-IDF feature extraction



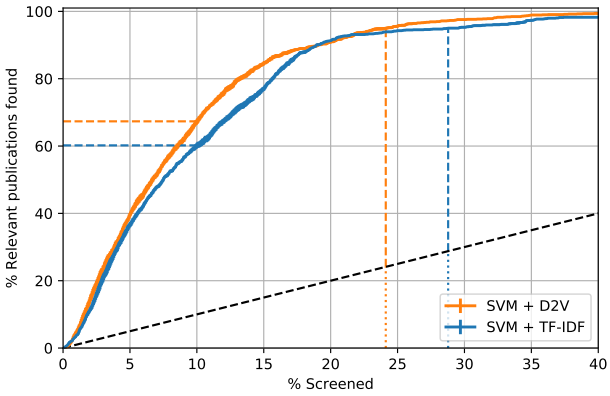
(c) Models adopting D2V feature extraction



(d) D2V vs TF-IDF for LR classifier



(e) D2V vs TF-IDF for RF classifier



(f) D2V vs TF-IDF for SVM classifier

Figure 1: Recall curves of different models for the Nudging dataset, indicating how fast the model finds relevant publications during the process of screening publications. Figure a displays curves for all seven models at once. The remaining figures display curves for several subsets of those models to allow for a more detailed inspection.

It can be seen from Table 2 that in terms of ATD, the best performing models on the Nudging dataset were SVM + D2V and LR + D2V, both with an ATD of 8.9%. This indicates that the average proportion of publications needed to screen to find a relevant publication was 8.9% for both models. In the SVM + D2V model, the standard deviation was 0.33 , whereas for the LR + D2V model $\hat{s} = 0.47$. This indicates that for the SVM + D2V model, the ATD values of individual trials were closer to the overal mean compared to the LR + D2V model, meaning that the SVM + D2V model performed more stable across different initial training datasets. Median ATD for this dataset was 9.6% with an MAD of 1.06, indicating that for half of the models, the ATD was within 1.06 distance from the median ATD.

Table 2: ATD values ($\bar{x}(\hat{s})$) for all model-dataset combinations, and median (MAD) for all datasets.

	Nudging	PTSD	Software	ACE	Virus	Wilson
SVM + TF-IDF	10.2 (0.19)	2.1 (0.13)	1.9 (0.04)	7.3 (1.18)	8.5 (0.17)	4.2 (0.33)
NB + TF-IDF	9.4 (0.29)	1.8 (0.11)	1.5 (0.03)	5.0 (0.53)	8.2 (0.22)	4.1 (0.37)
RF + TF-IDF	11.8 (0.44)	3.4 (0.27)	2.0 (0.09)	7.0 (0.76)	10.6 (0.42)	5.9 (1.20)
LR + TF-IDF	9.6 (0.19)	1.7 (0.10)	1.4 (0.02)	6.1 (1.20)	8.4 (0.24)	4.5 (0.34)
SVM + D2V	8.9 (0.33)	2.1 (0.15)	1.4 (0.05)	6.2 (0.34)	8.5 (0.21)	4.7 (0.31)
RF + D2V	10.4 (0.88)	3.1 (0.34)	1.6 (0.09)	7.3 (1.29)	9.3 (0.43)	7.5 (1.56)
LR + D2V	8.9 (0.47)	1.9 (0.17)	1.4 (0.04)	5.6 (0.18)	8.4 (0.41)	4.9 (0.32)
median (MAD)	9.6 (1.06)	2.1 (0.49)	1.5 (0.12)	6.2 (1.14)	8.5 (0.18)	4.7 (0.66)

As Table 3 shows, the highest WSS@95 value on the Nudging dataset was achieved by the NB + TF-IDF model with a mean of 71.7%, meaning that this model reduced the number of publications needed to screen with 71.7% at the cost of losing 5% of relevant publications. The estimated standard deviation of 1.37% indicates that in terms of WSS@95, this model performed the most stable across trials. The model with the lowest WSS@95 value was RF + TF-IDF ($\bar{x} = 64.9$, $\hat{s} = 2.50$). Median WSS@95 of these models was 66.9%, with a MAD of 3.05%, indicating that WSS@95 values of models varied the most within this dataset.

Table 3: WSS@95 values ($\bar{x}(\hat{s})$) for all model-dataset combinations, and median (MAD) for all datasets.

	Nudging	PTSD	Software	ACE	Virus	Wilson
SVM + TF-IDF	66.2 (2.90)	91.0 (0.41)	92.0 (0.10)	75.8 (1.95)	69.7 (0.81)	79.9 (2.09)
NB + TF-IDF	71.7 (1.37)	91.7 (0.27)	92.3 (0.08)	82.9 (0.99)	71.2 (0.62)	83.4 (0.89)
RF + TF-IDF	64.9 (2.50)	84.5 (3.38)	90.5 (0.34)	71.3 (4.03)	63.9 (3.54)	81.6 (3.35)
LR + TF-IDF	66.9 (4.01)	91.7 (0.18)	92.0 (0.10)	81.1 (1.31)	70.3 (0.65)	80.5 (0.65)
SVM + D2V	70.9 (1.68)	90.6 (0.73)	92.0 (0.21)	78.3 (1.92)	70.7 (1.76)	82.7 (1.44)
RF + D2V	66.3 (3.25)	88.2 (3.23)	91.0 (0.55)	68.6 (7.11)	67.2 (3.44)	77.9 (3.43)
LR + D2V	71.6 (1.66)	90.1 (0.63)	91.7 (0.13)	77.4 (1.03)	70.4 (1.34)	84.0 (0.77)
median (MAD)	66.9 (3.05)	90.6 (1.53)	92.0 (0.47)	77.4 (5.51)	70.3 (0.90)	81.6 (2.48)

As can be seen from the data in Table 4, LR + D2V was the best performing model in terms of RRF@10, with a mean of 67.5 indicating that after screening 10% of publications, on average 67.5% of all relevant publications had been identified, with a standard deviation of 2.59. The worst performing model was RF + TF-IDF ($\bar{x} = 53.6$, $\hat{s} = 2.71$). Median performance was 62.6, with an MAD of 3.89 indicating again that RRF@10 values were most dispersed for models within this dataset.

Table 4: RRF@10 values ($\bar{x}, (\hat{s})$) for all model-dataset combinations, and median (MAD) for all datasets.

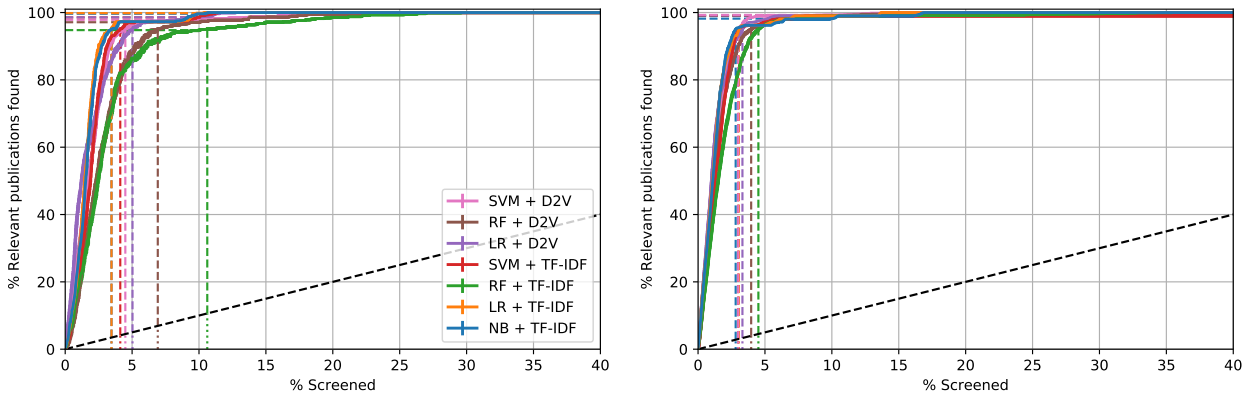
	Nudging	PTSD	Software	ACE	Virus	Wilson
SVM + TF-IDF	60.2 (3.12)	98.6 (1.40)	99.0 (0.00)	86.2 (5.25)	73.4 (1.62)	90.6 (1.17)
NB + TF-IDF	65.3 (2.61)	99.6 (0.95)	98.2 (0.34)	90.5 (1.40)	73.9 (1.70)	87.3 (2.55)
RF + TF-IDF	53.6 (2.71)	94.8 (1.60)	99.0 (0.00)	82.3 (2.75)	62.1 (3.19)	86.7 (5.82)
LR + TF-IDF	62.1 (2.59)	99.8 (0.70)	99.0 (0.00)	88.5 (5.16)	73.7 (1.48)	89.1 (2.30)
SVM + D2V	67.3 (3.00)	97.8 (1.12)	99.3 (0.44)	84.2 (2.78)	73.6 (2.54)	91.5 (4.16)
RF + D2V	62.6 (5.47)	97.1 (1.90)	99.2 (0.34)	80.8 (5.72)	67.3 (3.19)	75.5 (14.35)
LR + D2V	67.5 (2.59)	98.6 (1.40)	99.0 (0.00)	81.7 (1.81)	70.6 (2.21)	90.6 (5.00)
median (MAD)	62.6 (3.89)	98.6 (1.60)	99.0 (0.00)	84.2 (3.71)	73.4 (0.70)	89.1 (2.70)

Overall evaluation

Recall curves for the simulations on the five remaining datasets are presented in Figure 2. For the sake of conciseness, recall curves are only plotted once per dataset, like in Figure 1a. Please refer to Additional file 1 for figures presenting subsets of recall curves for the remaining datasets, like in Figure 1b-f.

First of all, for all datasets, the models were able to detect the relevant publications much faster compared to when screening publications at random order as the recall curves exceed the expected recall at screening at random order by far. Even the worst results outperform this reference condition. Across simulations, the ATD was at maximum 11.8% (in the Nudging dataset), the WSS@95 at least 63.9% (in the Virus dataset), and the lowest RRF@10 was 53.6% (in the Nudging dataset). Interestingly, all these values were achieved by the RF + TF-IDF model.

Similar to the simulations on the Nudging dataset (Figure 1b), the ordering of recall curves changes throughout the screening process, indicating that model performance is dependent on the number of publications that have been screened. Moreover, the ordering of models in the Nudging dataset (Figure 1b) does not replicate on the remaining five datasets (Figure 2).



(a) PTSD

(b) Software

(c) ACE

(d) Virus

(e) Wilson

Figure 2: Recall curves of all seven models for all indicating how fast the model finds relevant publications during the process of screening publications. on (a) the PTSD, (b) Software, (c) ACE, (d) Virus, and (e) Wilson dataset.

RQ1 - Comparison across classification techniques

The first reserach question was aimed at evaluating the first four models adopting either the NB, SVM, LR or RF classification technique, combined with TF-IDF feature extraction. When comparing ATD-values of the models (Table 2), the NB + TF-IDF model ranked first in the ACE, Nudging, PTSD, Virus and Wilson dataset, and second in the PTSD and the Software dataset, in which the LR + TF-IDF model achieved the lowest ATD value. The RF + TF-IDF ranked last in all of the datasets except in the ACE dataset, where the SVM + TF-IDF model achieved the highest ATD-value.

Additionally, in terms of WSS@95 (Table 3) the ranking of models was strikingly similar across all datasets. In the ACE, Nudging, Software, and Virus dataset, the highest WSS@95 value was always achieved by the NB + TF-IDF model, followed by LR + TF-IDF, SVM + TF-IDF, and RF + TF-IDF. In the PTSD dataset this ranking applied as well, except that the LR + TF-IDF and NB + TF-IDF showed equal WSS@95 values. The ordering of the models for the Wilson dataset was NB + TF-IDF, RF + TF-IDF, LR + TF-IDF and SVM + TF-IDF.

Moreover, in terms of RRF@10 (Table 4) the NB + TF-IDF model achieved the highest RRF@10 value in the ACE, Nudging, and Wilson dataset. LR + TF-IDF ranked first in the PTSD dataset, SVM + TF-IDF was the best performing model within the Wilson dataset. The RF + TF-IDF model was again the worst performing model within all datasets, with an exception for the Software dataset. In this dataset, NB + TF-IDF ranked fourth, the remaining three models achieved an equal RRF@10 score.

Taken together, these results show that while all four models perform quite well, the NB + TF-IDF shows high performance on all measures across all datasets, whereas the RF + TF-IDF model never performed best on any of the measures across all datasets.

RQ2 - Comparison across feature extraction techniques

This section is concerned with the question of how models using different feature extraction strategies relate to each other. The recall curves for the Nudging data (Figure 1d-f) show a clear trend of the models adopting D2V feature extraction outperforming their TF-IDF counterparts. This trend also shows from the WSS@95 and RRF@10 values indicated by the vertical and horizontal lines in the figure. Likewise, the ATD values (Table 2) indicate that for the models adopting a particular classification technique, the model adopting D2V feature extraction always achieved a smaller ATD-value than the model adopting TF-IDF feature extraction.

In contrast, this pattern of models adopting D2V outperforming their TF-IDF counterparts in the Nudging

dataset does not replicate across other datasets. Whether evaluated in terms of recall curves, WSS@95, RRF@10 or ATD, the findings were mixed. Neither one of the feature extraction strategies showed superior performance within certain datasets nor within certain classification techniques.

RQ3 - Comparison across research contexts

First of all, models showed much higher performance for some datasets than for others. Whilst performances on the PTSD (Figure 2a) and the Software dataset (Figure 2b) were quite high, performances were much lower across models for the Nudging (Figure 1a) and Virus (Figure 2d) datasets. There does not seem to be a clear distinction between the datasets from the biomedical sciences (ACE, Virus, and Wilson) and datasets from other fields (Nudging, PTSD, Software). The PTSD, Software and Nudging dataset also demonstrated high performances in terms of the median ATD, WSS@95 and RRF@10 values for these models (Table 2, 3, and 4).

Secondly, variability of between-model performance differed across datasets. For the PTSD (Figure 2a), Software (Figure 2b), and the Virus (Figure 2d) datasets, recall curves form a tight group meaning that within these datasets, the models perform relatively similar. For the Nudging (Figure 1a), ACE (Figure 2c), and Wilson (Figure 2e) dataset, the recall curves are much further apart, indicating that model performance is much more dependent on the classification technique and feature extraction strategy. The MAD values of the ATD, WSS@95 and RRF@10 confirm that within the PTSD, Software and Virus datasets, model performance is less spread out than within the Nudging, ACE and Wilson dataset.

Moreover, the curves for the ACE (Figure 2c) and Wilson (Figure 2e) datasets show a larger standard error of the mean compared to other the other datasets. For these datasets, model performance seemed to be more dependent on the initial training dataset compared to others.

Discussion

The current study set out to evaluate performance of active learning models for the purpose of identifying relevant publications in systematic review datasets. It has been one of the first attempts to examine different classification strategies and feature extraction strategies in active learning models for systematic reviews. Moreover, this study has provided a deeper insight into the performance of active learning models across research contexts.

Active learning-based screening prioritization

Overall, the findings confirm the great potential of active learning models in reducing workload for systematic reviewers. All models were able to detect 95% of the relevant publications after screening less than 40% of the total number of publications, indicating that active learning models can save more than half of the workload in the screening process. In a previous study, the ACE dataset was used to simulate a model that did not use active learning, finding a WSS@95 value of 56.61% [33], whereas the models in the current study achieved far superior WSS@95 values varying from 68.6% to 82.9% in this dataset. Active learning models clearly outperformed a model which did not use active learning. In addition, the Software dataset was used to simulate an active learning model [11] and reached WSS@95 of 91%, strikingly similar the WSS@95 values found in the current study which ranged from 90.5% to 92.3%.

Classification techniques

The first research question in this study sought to evaluate models adopting different classification techniques. The most obvious finding to emerge from these evaluations was that the NB + TF-IDF model consistently performed as one of the best models. The results suggest that whilst SVM - the classifier standardly used in software tools for active learning in systematic reviews - performed fairly well, the LR and NB classification techniques are good if not superior alternatives. Note that NB and LR were always good methods for text classifciation tasks [51].

Feature extraction strategy

The overall results on models adopting D2V versus TF-IDF feature extraction strategy remain inconclusive. According to these findings, adopting D2V instead of the well-established TF-IDF feature extraction strategy does not lead to better performing models. Given these results, although preliminary, preference goes out to the TF-IDF feature extraction technique as this relatively simplistic technique will lead to more interpretable model.

Research contexts

Simulating models on a heterogenous collection of systematic review datasets revealed that model performance is very data-dependent. Within some datasets, models achieved much higher overall performance than within other datasets. Moreover, for some datasets, differences between models were much larger than

for other datasets. It has been suggested that active learning is more difficult for datasets from the social sciences compared to data from the medical sciences [12]. This does not appear to be the case as performance within the biomedical datasets (Wilson, Virus, ACE) was not in any way superior to performance within the datasets from the social sciences (PTSD and Nudging). An issue that emerges from these findings is that difficulty of active learning was not confined to any particular research area. A possible explanation for this is that difficulty of active learning could be attributed to factors more directly related to the systematic review at hand, such as the inclusion rate and the complexity of inclusion criteria used to identify relevant publications [16,52]. Although the current study did not investigate the inclusion criteria of systematic reviews, the datasets on which the active learning models performed worst, Nudging and Virus, were interestingly also the datasets with the highest inclusion rates, 5.4% and 5.0%, respectively.

Limitations and future research

When applied in systematic review practice, the succes of active learning models stands or falls down with the generalizability of model performance across unseen datasets. It is important to bear in mind that model hyperparameters were optimized for each model-dataset combination. Thus, the observed results reflect maximum model performance for the datasets at hand. The question remains whether model performance generalizes to datasets for which the hyperparameters were not optimized. Further research should be undertaken to determine the sensitivity of model performance to the hyperparameter values.

Additionally, whilst the sample of datasets in the current study is diverse compared to previous studies, the sample size does not allow for investigating how characteristics of the data - such as inclusion rate - relate to model performance. To build confidence in active learning models for screening publications, it is essential to identify how characteristics of the data influence model performance. Such a study requires more data on systematic reviews. Therefore, we call systematic reviewers to make an effort to openly publish their datasets.

An unanticipated finding was that the runtime of simulations varied widely across models, indicating that some models need more time to retrain after a publication has been labelled than other models. This finding has important implications for the practical application of such models, as an efficient model should be able to keep up with the decision-making speed of the reviewer. Further studies taking into account retraining time will need to be undertaken.

Conclusions

Overall, the findings of this study confirm that active learning models show great potential of retrieving relevant publications in a systematic review dataset, while minimizing the number of publications needed to screen. The results shed new light on the performance of different classification techniques, indicating that the NB classification technique is superior to the widely used SVM As model performance differs vastly across datasets, this study raises the question what causes models to yield more workload savings for some systematic review datasets than for others. In order to facilitate the applicability of active learning models in systematic review practice, it is essential to identify how dataset characteristics relate to model performance.

Declarations

List of abbreviations

ATD

D2V LR MAD ML NB PTSD RF RRF SD SEM SVM TF-IDF WSS

Ethics approval and consent to participate

This study has been approved by the Ethics Committee of the Faculty of Social and Behavioural Sciences of Utrecht University, filed as an amendment under study 20-104.

Consent for publication

Not applicable.

Availability of data and materials

As reported in the main text, all data and materials are available through the GitHub repository for this paper, https://github.com/asreview/paper-evaluating-models-across-research-areas. This repository contains all systematic review datasets used during this study and their preprocessing scripts, scripts and data on the hyperparamter optimization, scripts on the simulations, scripts for analyzing the results of the simulations, and the source files for this manuscript. All output files of the simulation study are stored on the Open Science Framework page of this paper, https://osf.io/7mr2g/.

Competing interests

The author declares that they has no competing interests.

Funding

Computing hours on the Cartesius supercomputer were funded by SURFsara. SURFsara had no role whatsoever in the design of the current study, nor in the data collection, analysis and interpretation, nor in writing the manuscript.

Author's contributions

RS developed the DR balance strategy and ATD metric.

All authors read and approved the final manuscript.

Acknowledgements

I am grateful for all researchers who have made great efforts to openly publish the data on their systematic reviews, special thanks go out to Rosanna Nagtegaal. I would also like to thank dr. Caroline van Baal for supporting me in writing this paper, and prof. dr. René Eijkemans, for being the second grader of this paper Finally, I would like to express my appreciation to my supervisors prof. dr. Rens van de Schoot, Jonathan de Bruin, and dr. Raoul Schram. Your door was always open and your enthusiasm was contagious.

References

[1] PRISMA-P Group, Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015;4:1. https://doi.org/10.1186/2046-4053-4-1.

[2] Gough D, Elbourne D. Systematic Research Synthesis to Inform Policy, Practice and Democratic Debate. Soc Policy Soc 2002;1:225–36. https://doi.org/10/bdmp7h.

[3] Chalmers I. The lethal consequences of failing to make full use of all relevant evidence about the effects of medical treatments: The importance of systematic reviews. In:. Treating individualsfrom randomised trials to personalised medicine., Lancet; 2007, pp. 37–58.

[4] Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017;7:e012545. https://doi.org/10/f9tf57.

[5] Lau J. Editorial: Systematic review automation thematic series. Syst Rev 2019;8:70. https://doi.org/10/ggsmwf.

[6] Harrison H, Griffin SJ, Kuhn I, Usher-Smith JA. Software tools to support title and abstract screening for systematic reviews in healthcare: An evaluation. BMC Med Res Methodol 2020;20:7. https://doi.org/10.1186/s12874-020-0897-3.

[7] O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: A systematic review of current approaches. Syst Rev 2015;4:5. https://doi.org/10.1186/2046-4053-4-5.

[8] Cohen AM, Ambert K, McDonagh M. Cross-Topic Learning for Work Prioritization in Systematic Review Creation and Update. J Am Med Inform Assoc 2009;16:690–704. https://doi.org/10/c3shq2.

[9] Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O'Mara-Eves A, et al. Pinpointing needles in giant haystacks: Use of text mining to reduce impractical screening workload in extremely large scoping reviews. Res Synth Methods 2014;5:31–49. https://doi.org/10.1002/jrsm.1093.

[10] Yu Z, Menzies T. FAST2: An intelligent assistant for finding relevant papers. Expert Syst Appl 2019;120:57–71. https://doi.org/10.1016/j.eswa.2018.11.021.

[11] Yu Z, Kraft NA, Menzies T. Finding better active learners for faster literature reviews. Empir Softw Eng 2018;23:3161–86. https://doi.org/10.1007/s10664-017-9587-0.

[12] Miwa M, Thomas J, O'Mara-Eves A, Ananiadou S. Reducing systematic review workload through certainty-based screening. J Biomed Inform 2014;51:242–53. https://doi.org/10.1016/j.jbi.2014.06.005.

[13] Cormack GV, Grossman MR. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In:. Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, Gold Coast, Queensland, Australia: Association for Computing Machinery; 2014, pp. 153–62. https://doi.org/10.1145/2600428.2609601.

[14] Cormack GV, Grossman MR. Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review 2015.

[15] Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinform 2010;11:55. https://doi.org/10.1186/1471-2105-11-55.

[16] Gates A, Johnson C, Hartling L. Technology-assisted title and abstract screening for systematic reviews: A retrospective evaluation of the Abstrackr machine learning tool. Syst Rev 2018;7:45. https://doi.org/10/ggpsx4.

[17] Settles B. Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 2012;6:1–114. https://doi.org/10.2200/S00429ED1V01Y201207AIM018.

[18] Settles B. Active Learning Literature Survey. University of Wisconsin-Madison Department of Computer Sciences; 2009.

[19] Singh G, Thomas J, Shawe-Taylor J. Improving Active Learning in Systematic Reviews 2018.

[20] Carvallo A, Parra D. Comparing Word Embeddings for Document Screening based on Active Learning n.d.:8.

[21] Yimin M. Text Classification on Imbalanced Data: Application to Systematic Reviews Automation. University of Ottowa, 2007.

[22] Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. In:. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, Miami, Florida, USA: Association for Computing Machinery; 2012, pp. 819–24. https://doi.org/10.1145/2110363.2110464.

[23] Cheng SH, Augustin C, Bethel A, Gill D, Anzaroot S, Brun J, et al. Using machine learning to advance synthesis and use of conservation and environmental evidence. Conserv Biol 2018;32:762–4. https://doi.org/10.1111/cobi.13117.

[24] Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyana web and mobile app for systematic reviews. Syst Rev 2016;5:210. https://doi.org/10.1186/s13643-016-0384-4.

[25] Przybyła P, Brockmeier AJ, Kontonatsios G, Pogam M-AL, McNaught J, Erik von Elm, et al. Prioritising references for systematic reviews with RobotAnalyst: A user study. Res Synth Methods 2018;9:470–88. https://doi.org/10.1002/jrsm.1311.

[26] Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards Automatic Recognition of Scientifically Rigorous Clinical Research Evidence. J Am Med Inform Assn 2009;16:25–31. https://doi.org/10/bjkhh9.

[27] Aphinyanaphongs Y. Text Categorization Models for High-Quality Article Retrieval in Internal Medicine. J Am Med Inform Assoc 2004;12:207–16. https://doi.org/10/cpvp52.

[28] Aggarwal CC, Zhai C. A Survey of Text Classification Algorithms. In: Aggarwal CC, Zhai C, editors. Mining Text Data, Boston, MA: Springer US; 2012, pp. 163–222. https://doi.org/10.1007/978-1-4614-3223-4_6.

[29] Molnar C. Interpretable Machine Learning. Lulu.com; 2020.

[30] Zhang W, Yoshida T, Tang X. A comparative study of TF*IDF, LSI and multi-words for text classification. Expert Syst Appl 2011;38:2758–65. https://doi.org/10/dp7268.

[31] Le QV, Mikolov T. Distributed Representations of Sentences and Documents 2014.

[32] Marshall IJ, Johnson BT, Wang Z, Rajasekaran S, Wallace BC. Semi-Automated evidence synthesis in health psychology: Current methods and future prospects. Health Psychol Rev 2020;14:145–58. https://doi.org/10/ggjv98.

[33] Cohen AM, Hersh WR, Peterson K, Yen P-Y. Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. J Am Med Inform Assoc 2006;13:206–19. https://doi.org/10.1197/jamia.M1929.

[34] Appenzeller-Herzog C, Mathes T, Heeres MLS, Weiss KH, Houwen RHJ, Ewald H. Comparative effectiveness of common therapies for Wilson disease: A systematic review and meta-analysis of controlled studies. Liver Int 2019;39:2136–52. https://doi.org/10.1111/liv.14179.

[35] Kwok KTT, Nieuwenhuijse DF, Phan MVT, Koopmans MPG. Virus Metagenomics in Farm Animals: A Systematic Review. Viruses 2020;12:107. https://doi.org/10.3390/v12010107.

[36] Nagtegaal R, Tummers L, Noordegraaf M, Bekkers V. Nudging healthcare professionals towards evidence-based medicine: A systematic scoping review. J Behav Public Adm 2019;2. https://doi.org/doi.org/10.30636/jbpa.22.71.

[37] van de Schoot R, Sijbrandij M, Winter SD, Depaoli S, Vermunt JK. The GRoLTS-Checklist: Guidelines for reporting on latent trajectory studies. Struct Equ Model Multidiscip J 2017;24:451–67. https://doi.org/10/gdpcw9.

[38] van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. ASReview: Open source software for efficient and transparent active learning for systematic reviews 2020.

[39] Wynants L, Calster BV, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ 2020;369. https://doi.org/10/ggr2qk.

[40] Tong S, Koller D. Support vector machine active learning with applications to text classification. J Mach Learn Res 2001;2:45–66.

[41] Kremer J, Steenstrup Pedersen K, Igel C. Active learning with support vector machines. WIREs Data Min Knowl Discov 2014;4:313–26. https://doi.org/10/f6fss7.

[42] Zhang H. The Optimality of Naive Bayes. In:. vol. 2, 2004.

[43] Breiman L. Random Forests. Machine Learning 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.

[44] Ramos J, others. Using tf-idf to determine word relevance in document queries. In:. Proceedings of the first instructional conference on machine learning, vol. 242, Piscataway, NJ; 2003, pp. 133–42.

[45] Fu JH, Lee SL. Certainty-Enhanced Active Learning for Improving Imbalanced Data Classification. In:. 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada: IEEE; 2011, pp. 405–12. https://doi.org/10.1109/ICDMW.2011.43.

[46] van de Schoot R, de Bruin J, Schram R, Zahedi P, Kramer B, Ferdinands G, et al. ASReview: Active learning for systematic reviews 2020. https://doi.org/10/ggssnj.

[47] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.

[48] Appenzeller-Herzog C. Data from Comparative effectiveness of common therapies for Wilson disease: A systematic review and meta-analysis of controlled studies 2020.

[49] Hall T, Beecham S, Bowes D, Gray D, Counsell S. A Systematic Literature Review on Fault Prediction Performance in Software Engineering. IEEE Trans Softw Eng 2012;38:1276–304. https://doi.org/10.1109/TSE.2011.103.

[50] Nagtegaal R, Tummers L, Noordegraaf M, Bekkers V. Nudging healthcare professionals towards evidence-based medicine: A systematic scoping review 2019.

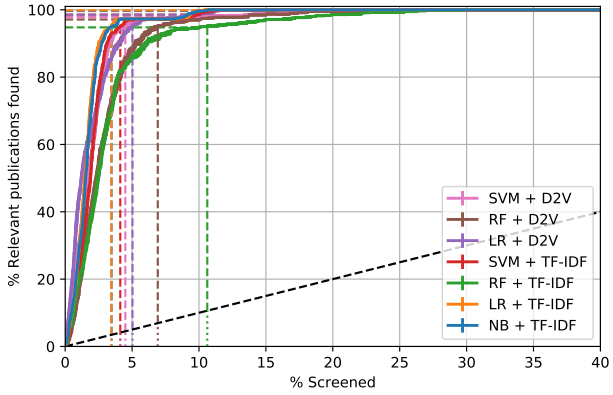
[51] Mitchell TM. Does Machine Learning Really Work? AI Mag 1997;18:11–1. https://doi.org/10/gg4g34.

[52] Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. Systematic Reviews 2015;4:80. https://doi.org/10/f7ms4w.

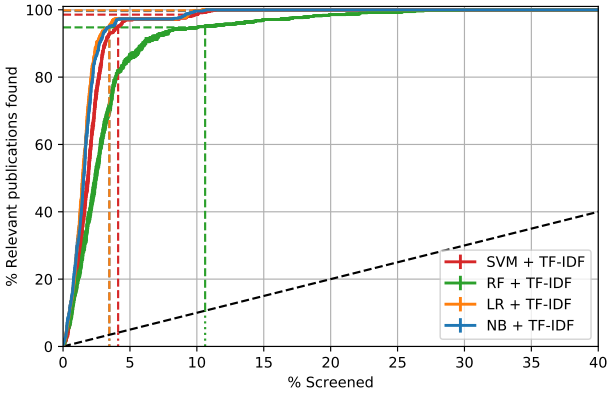
Additional file 1

Recall curves plot separately for the PTSD, Software, ACE, Virus and Wilson datasets.

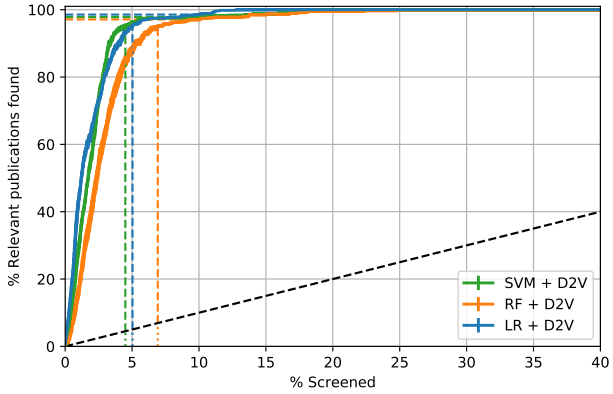
PTSD



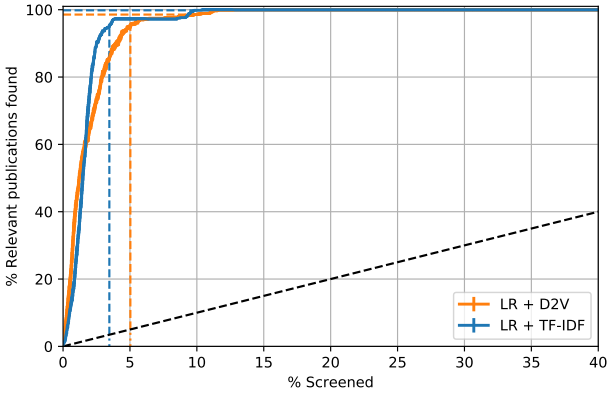
(a) All seven models



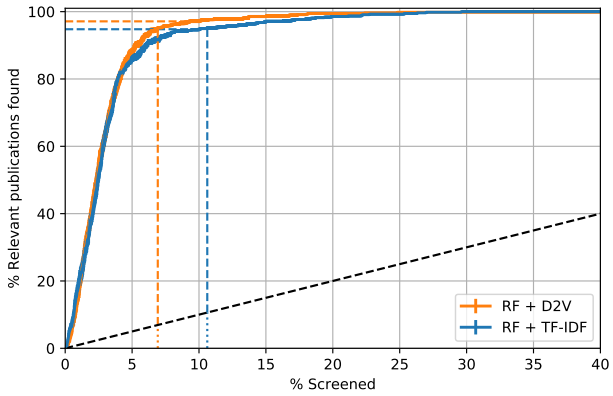
(b) Models adopting TF-IDF feature extraction



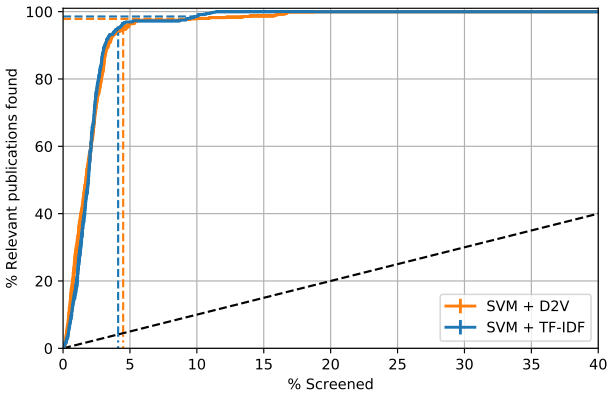
(c) Models adopting D2V feature extraction



(d) D2V vs TF-IDF for LR classifier



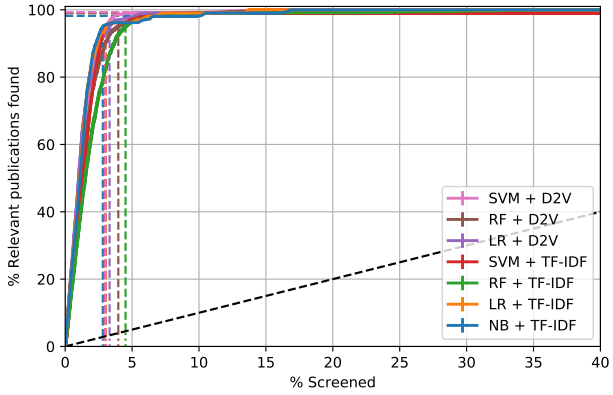
(e) D2V vs TF-IDF for RF classifier



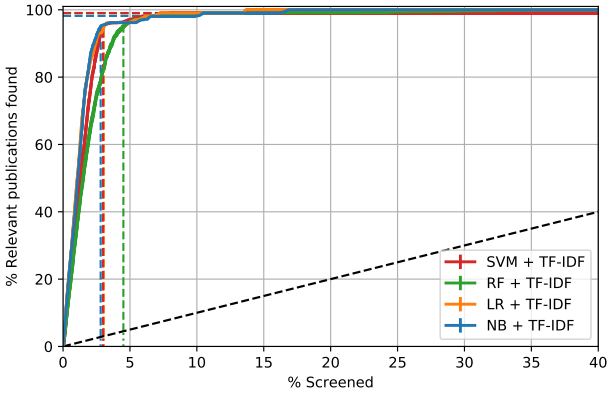
(f) D2V vs TF-IDF for SVM classifier

Figure 3: Recall curves for the PTSD dataset.

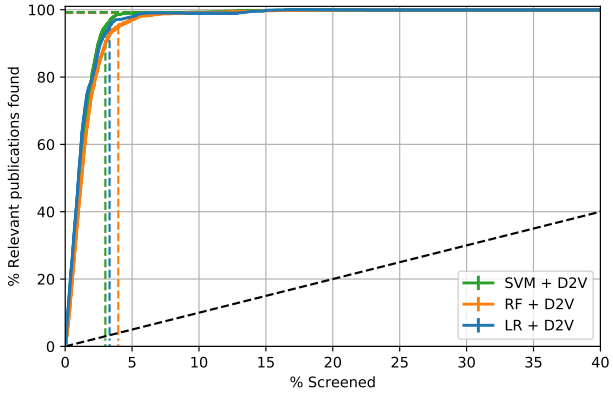
Software



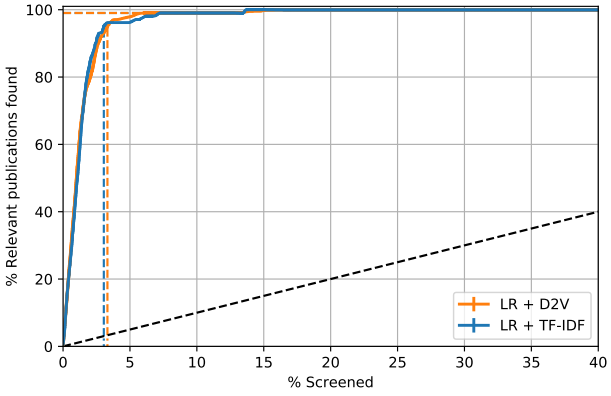
(a) All seven models



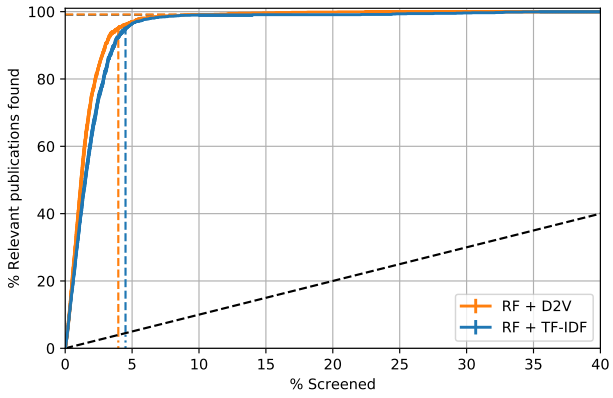
(b) Models adopting TF-IDF feature extraction



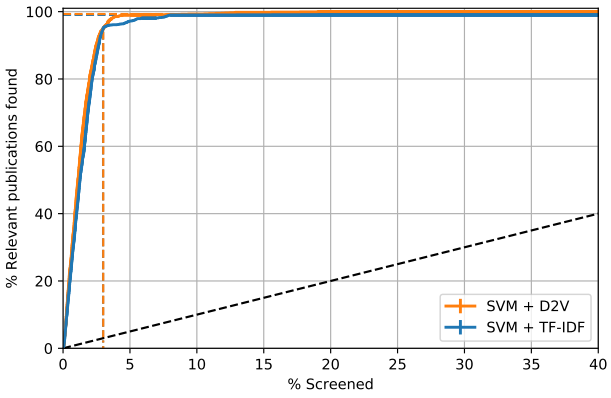
(c) Models adopting D2V feature extraction



(d) D2V vs TF-IDF for LR classifier



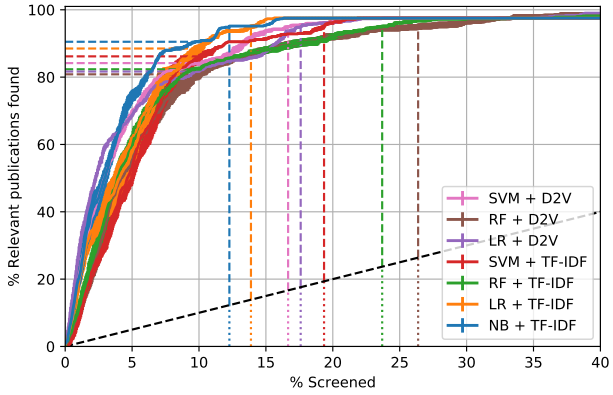
(e) D2V vs TF-IDF for RF classifier



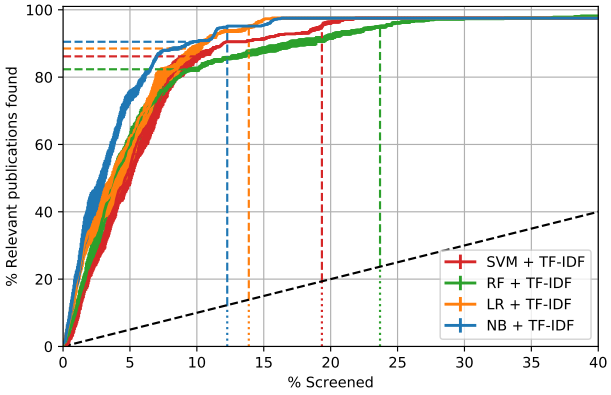
(f) D2V vs TF-IDF for SVM classifier

Figure 4: Recall curves for the Software dataset.

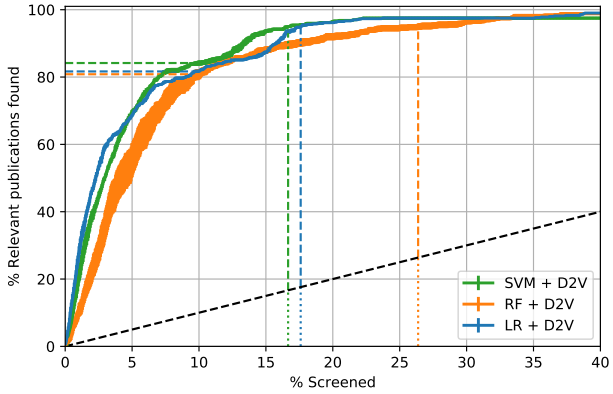
ACE



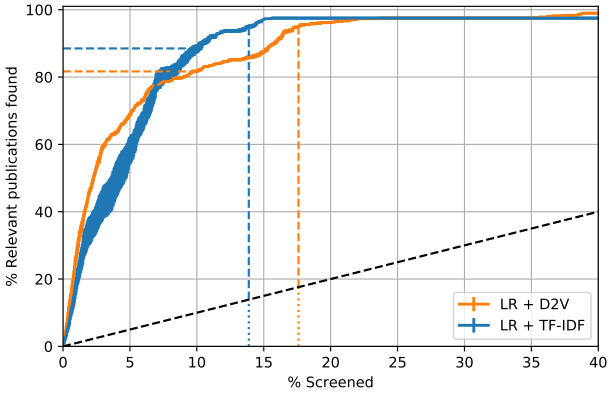
(a) All seven models



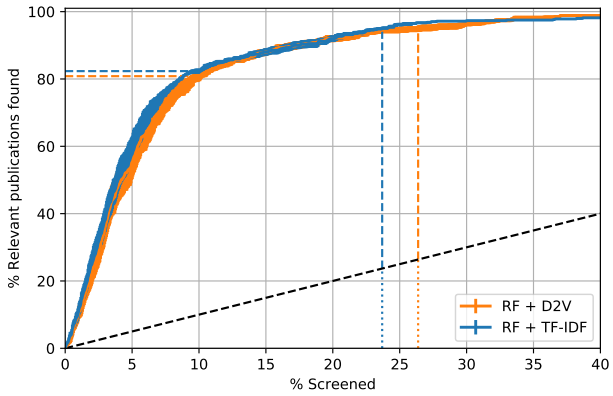
(b) Models adopting TF-IDF feature extraction



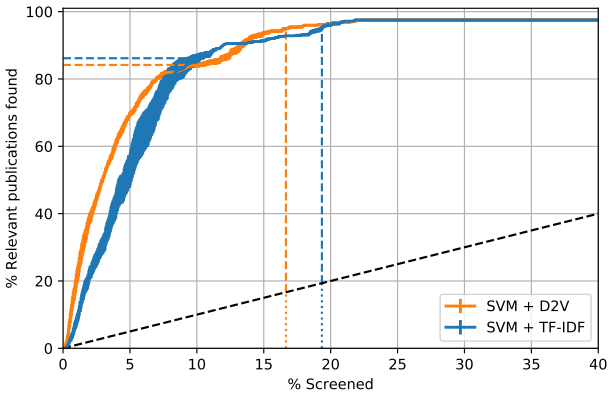
(c) Models adopting D2V feature extraction



(d) D2V vs TF-IDF for LR classifier



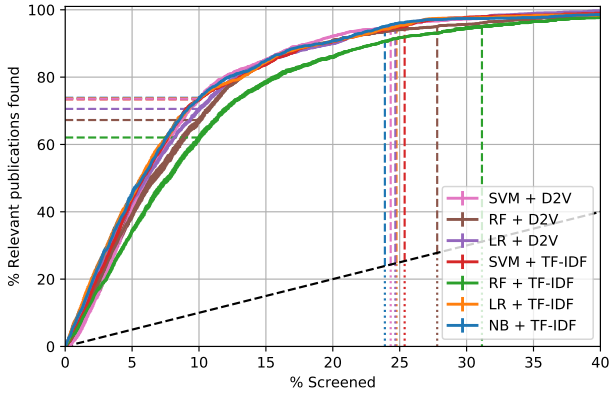
(e) D2V vs TF-IDF for RF classifier



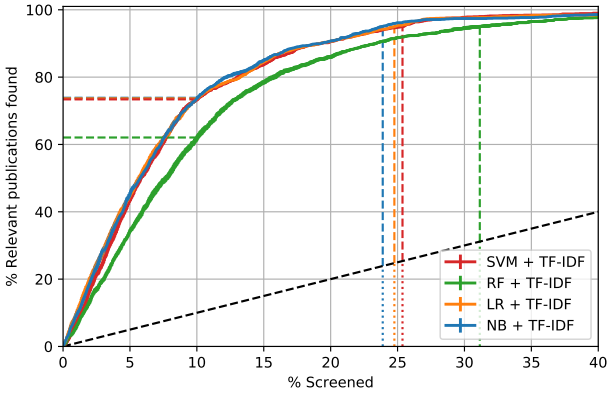
(f) D2V vs TF-IDF for SVM classifier

Figure 5: Recall curves for the ACE dataset.

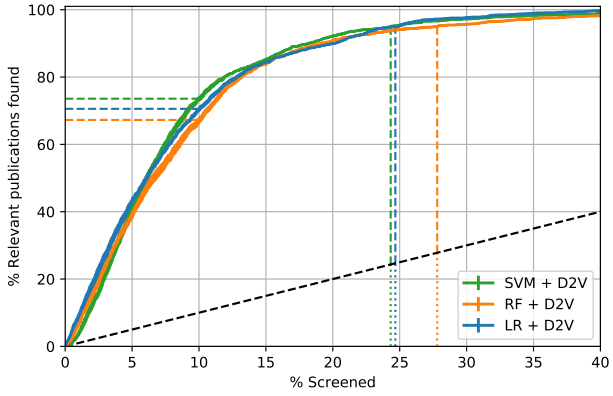
Virus



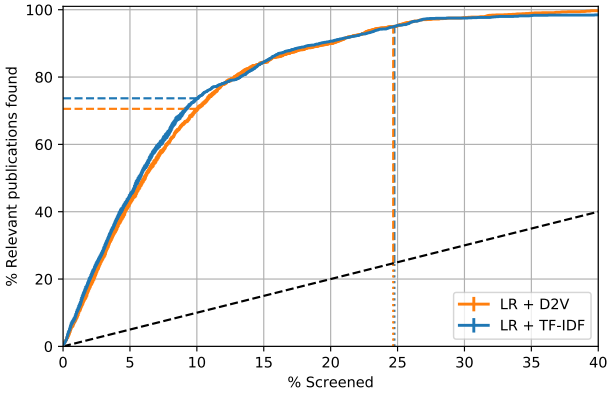
(a) All seven models



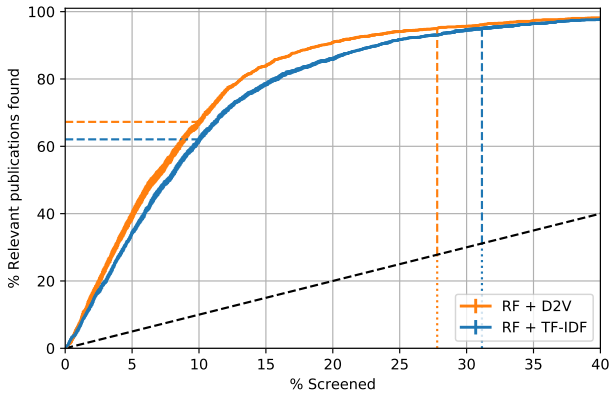
(b) Models adopting TF-IDF feature extraction



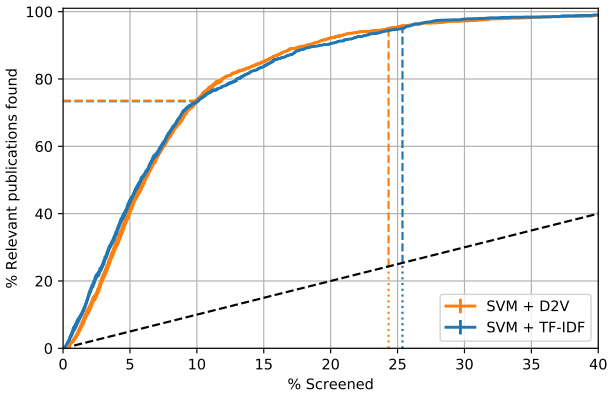
(c) Models adopting D2V feature extraction



(d) D2V vs TF-IDF for LR classifier



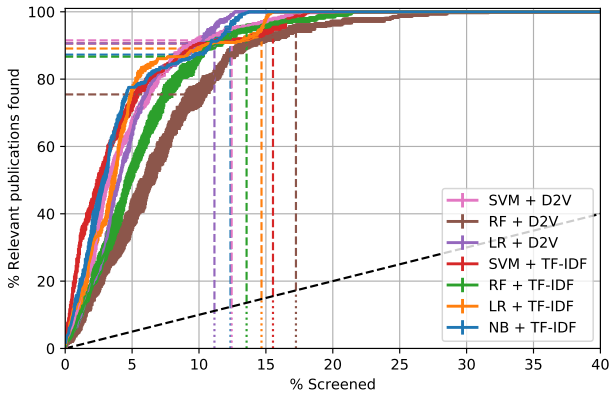
(e) D2V vs TF-IDF for RF classifier



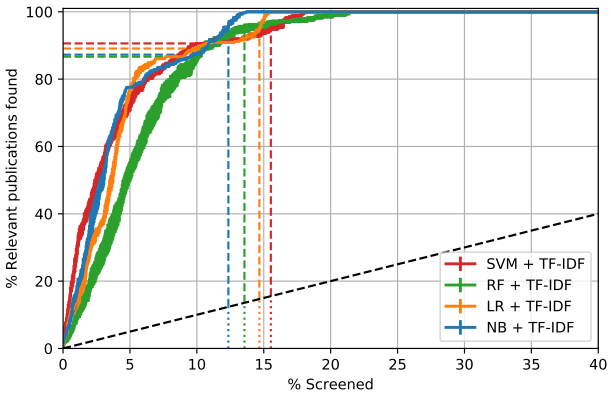
(f) D2V vs TF-IDF for SVM classifier

Figure 6: Recall curves for the Virus dataset.

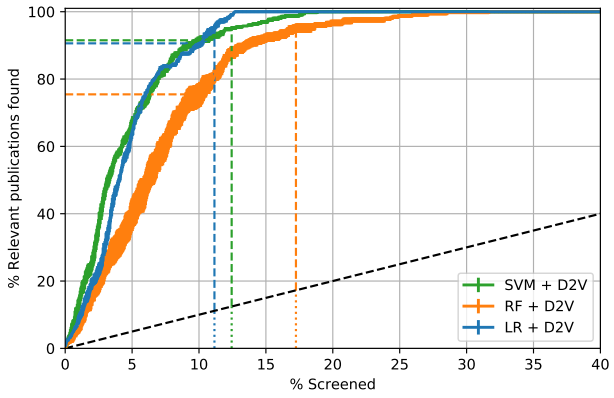
Wilson



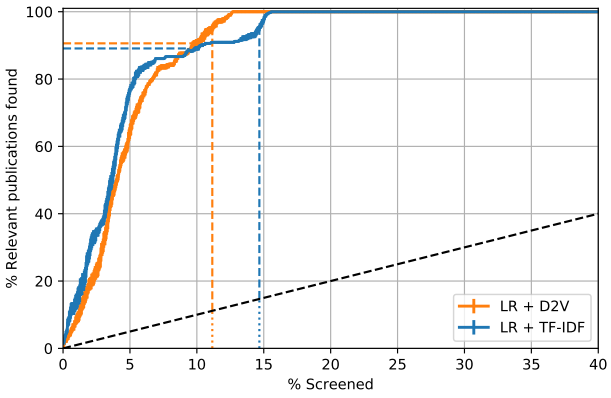
(a) All seven models



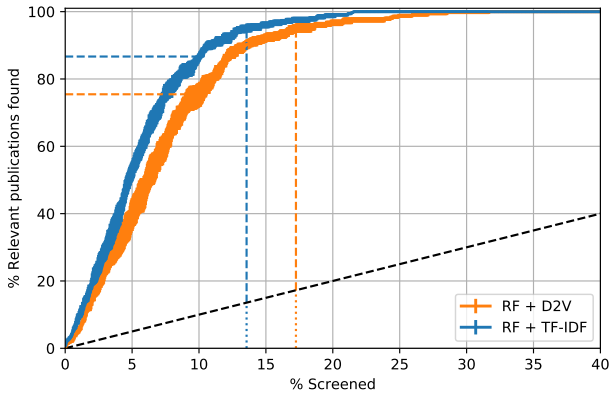
(b) Models adopting TF-IDF feature extraction



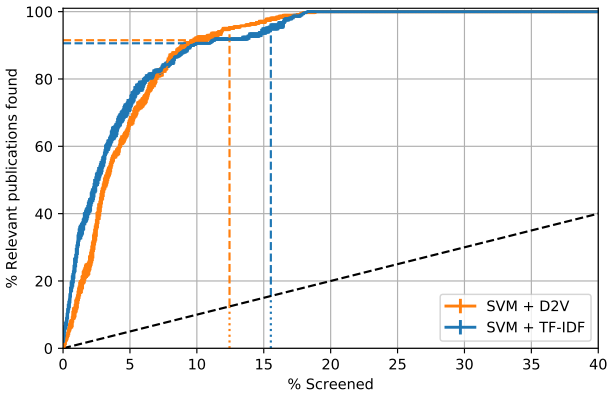
(c) Models adopting D2V feature extraction



(d) D2V vs TF-IDF for LR classifier



(e) D2V vs TF-IDF for RF classifier



(f) D2V vs TF-IDF for SVM classifier

Figure 7: Recall curves for the Wilson dataset.