Active learning for screening prioritization in systematic reviews

A simulation study

Gerbrich Ferdinands Raoul Schram Jonathan de Bruin Ayoub Bagheri
Daniel Oberski Lars Tummers Rens van de Schoot

30 July, 2020

Abstract

Background Conducting a systematic review requires great screening effort. Various tools have been proposed to speed up the process of screening thousands of titles and abstracts by engaging in active learning. In active learning, the reviewer interacts with machine learning software to identify relevant publications as early as possible. To gain a comprehensive understanding of active learning models for reducing workload in systematic reviews, the current study provides an methodical overview of such models. Active learning models were evaluated across four different classification techniques (naive Bayes, logistic regression, support vector machines, and random forest) and two different feature extraction strategies (TF-IDF and Doc2vec). Moreover, models were evaluated across six systematic review datasets from various research areas to assess generalizability of active learning models across different research contexts.

Methods Performance of the models were assessed by conducting simulations on six systematic review datasets. We defined desirable model performance as maximizing recall while minimizing the number of publications needed to screen. Model performance was evaluated by recall curves, WSS@95, RRF@10, and ATD.

Results Within all datasets, the model performance exceeded screening at random order to a great degree, reducing the number of publications needed to screen by 91.7% to 63.9%.

Conclusions Active learning models for screening prioritization show great potential in reducing the workload of systematic reviewers. Overall, the Naive Bayes + TF-IDF model performed the best.

keywords: systematic reviews, active learning, screening prioritization, researcher-in-the-loop, title-and-abstract screening

Background

Systematic reviews are top of the bill in research. A systematic review brings together all available 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 to identify publications relevant to answering the research question, they among else involve the manual screening of thousands of titles and abstracts.

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 answering the research question at hand is urgent, it is extremely challenging to provide a review that is both timely and comprehensive.

To ensure a timely review, reducing the workload in systematic reviews is essential. With advances in machine learning (ML), there has been wide interest in tools to reduce the workload in systematic reviews [6]. Various ML 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 the workload with 30-70% at the cost of losing 5% of relevant publications, in else, a 95% recall [7].

A well-established approach to increase the efficiency of title and abstract screening is screening prioritization [8,9]. In screening prioritization, the ML 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 found [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 unlabeled data by allowing the model to select the records from which it wants to learn [17]. The model proposes these records to a human annotator who provides the records with labels, which the model then uses to update 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 unlabeled 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 systematic reviews has been extensively studied [10–12,15,16]. While previous studies have evaluated active learning models in many forms and shapes [10–15,19–21], ready-to-use software tools implementing such 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 machines (SVM). It was found [26,27] that different classification techniques can serve different needs in the retrieval of relevant publications, for example the desired balance between recall and precision. Therefore, it 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 (NB), logistic regression (LR), SVM, and random forest (RF). These are widely adopted techniques in text classification [28] and are fit for software tools to be used in scientific practice due to their relatively short computation time.

Another component that influences model performance is how the textual content of titles and abstracts is represented in a model, called the feature extraction strategy [19,20,29]. One of the more sophisticated feature extraction strategies is Doc2vec (D2V), also known as paragraph vectors [30]. D2V learns continuous distributed vector representations for pieces of text. In distributed text representations, words are assumed to appear in the same context when they are similar in terms of a latent space, the "embedding". A word embedding is simply a vector of scores estimated from a corpus for each word; D2V is an extension of this idea to document embeddings. Embeddings can sometimes outperform simpler feature extraction strategies such as term frequency-inverse document frequency (TF-IDF). They can be trained on large corpora to capture wider semantics and subsequently applied in a specific systematic reviewing application [30]. Therefore, it is interesting 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] or computer science [11]. To draw conclusions about the general effectiveness of active learning models, it is essential to evaluate models on reviews from varying research contexts [7,31]. To our knowledge, Miwa et al [12] were the only researchers to make a direct comparison between systematic reviews from different research areas, such as the social and the medical sciences. They found that the application of active learning to systematic reviews was more difficult on a systematic review 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 social science, 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 classification techniques? **RQ2** What is the performance of active learning models across two feature extraction strategies? **RQ3** Does the performance of active learning models differ across six systematic reviews from four research areas?

The purpose of this paper is to show the usefulness of 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. To assess generalizability of the models across research contexts, datasets containing previous systeamtic reviews were collected from the fields of medical science [32–34], computer science [11], and social science [35,36]. The models, datasets and simulations are implemented in a pipeline of active learning for screening prioritization, called `ASReview` [37]. `ASReview` is a generic open source tool, encouraging fellow researchers to replicate findings from previous studies. To facilitate usability and acceptability of ML-assisted title and abstract screening in the field of systematic review our scripts and data used are openly available.

Methods

Technical details

What follows is a more detailed account of the active learning models 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 unlabeled records (titles and abstracts), $\mathcal{U}$.
- The reviewer provides a label for a few, e.g. 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 unlabeled 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 requests 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 unlabeled 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 the 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 [38].

Class imbalance problem

Typically, only a fraction of the publications belong to the relevant class (2.94%, [4]). To some extent, this fraction is under the control of the researcher through the search criteria: if the researcher narrows the search query, it will generally result in a higher proportion of relevant publications. However, in most applications this practice would yield an unacceptable number of false negatives (erroneously excluded papers) in the querying phase of the review process. For this reason, the query phase in most practical applications would

typically yield a very low percentage of relevant publications. Because there are genearlly far fewer examples of relevant than irrelevant publications to train on, the class imbalance causes the classifier to miss relevant publications [7]. Moreover, classifiers can achieve high accuracy but still fail to identify any of the relevant publications [15].

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, we 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 and depends 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. Additional file 1 provides a detailed script to perform DR.

Classification

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

- Support vector machines (SVM) - SVMs separate the data into classes by finding a multidimensional hyperplane [39,40].
- L2-regularized logistic regression (LR) - models the probabilities describing the possible outcomes by a logistic function. The classifier uses regularization, shrinking coefficients of features with small contributions 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 [41].
- Random forests (RF) is a supervised learning algorithm where a large number of decision trees are fit on samples obtained from the original data by sampling both rows (bootstrapped samples) and columns (feature samples). In prediction mode, each tree casts a vote on the class, and the final prediction is the class that received the most votes [42].

Feature extraction

To predict relevance of a given publication, the classifier uses information from the publications in the dataset. Examples of 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 as feature vectors. This process of numerically representing textual content is referred to as 'feature extraction'.

TF-IDF is a specific way of assigning scores to the cells of the "document-term matrix" used in all bag-of-words representations. That is, the rows of the document-term matrix represent the documents (titles and abstracts) and the columns represent all words in the dictionary. Instead of simply counting the number of times each word occurred in the given document, TF-IDF assigns a score to a word relative to the number of documents the word occurs. The idea behind weighting words by their rarity is that surprising word choices should subsequently make for more discriminative features [43]. 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 syntax. However, in practice, TF-IDF is often found to be a strong baseline [44].

In recent years, a range of modern methods have been developed that often outperform bag-of-words approaches. Here, we consider doc2vec, an extension of the classic word2vec embedding [30]. In word embedding models, whether a word did or did not happen to appear in a specific context is predicted by its similarity to that context in a latent space - the "embedding". The context is usually a sliding window across training sentences. For example, if the window "child ate cookies" occurs in the training data, this might be compared with a random 'negative' window that did not occur, such as "child lovely cookies". The tokens "child" and "cookies" are then assigned scores (vectors) that give a higher inner product with the "child" vector, and a smaller product with "lovely". The word vectors of "ate" and "lovely" are similarly updated. Typically the embedding dimension is a few hundred, i.e. each word vector contains some two hundred scores. Note that if "cookies" previously co-occurred frequently with "spinach", then the above also indirectly makes "ate" more similar to "spinach", even if these two words have not been observed yet in the same context. Thus, the distributed representation learns something of the meaning of these words through their occurrence in similar contexts. D2V performs the above procedure while including a paragraph identifier, allowing for paragraph embeddings - or, in our case, embeddings for titles and abstracts. In short, D2V converts each abstract into a vector of a few hundred scores, which can be used to predict inclusion or exclusion.

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]. Further strategies include selecting the next instance to optimize for various criteria, including: model fit (MLI), model change (MMC), parameter estimate accuracy (EVR), and expected (EER) or worst-case (MER) prediction accuracy [45]. Although uncertainty sampling is not explicitly motivated by the optimization of any particular criterion, intuitively it can be seen as attempting to improve the model's accuracy by reducing uncertainty about its parameter estimates.

Simulation-based comparisons of these methods across different domains have yielded an ambiguous picture of their relative strengths [12,45]. What has become clear from such studies is that the features of the task at hand determine the effectiveness of active learning strategies ("no free active lunch"). For example, if a linear classifier is used for a task that also happens to have a Bayes optimal linear decision boundary, a model-based approach such as Fisher information reduction can be expected to perform well, whereas the same technique can be disastrous when the model is misspecified - a fact that cannot be known in advance. Furthermore, the criteria mentioned above differ from the task of semi-automated abstract screening: here, the aim is not to obtain an accurate model, but rather to end up with a list of records belonging to the relevant class [46]. This is the criterion corresponding intuitively to certainty-based sampling. For this reason, we choose to focus on certainty-based sampling strategies as the baseline strategy for active learning in systematic reviewing. However, different strategies may outperform our baseline in specific applications.

Simulation study

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

Set-up

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

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 combinations:

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

The combination NB + D2V could not be tested because the multinomial naive Bayes classifier[1] requires a feature matrix with positive values, whereas the D2V feature extraction approach[2] produces a feature matrix that can 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 challenging scenario where the reviewer has very little prior knowledge on the publications in the data. The model was retrained each time after a publication had been labelled. A simulation ended after all publications in the dataset had been labelled. To account for sampling variance, every 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 for each dataset 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 [47]. For each simulation, hyperparameters were optimized through a Tree of Parzen Estimators (TPE) algorithm [48] to arrive at maximum model performance.

Simulations were carried out in ASReview version 0.9.3 [47]. Analyses were carried out using `R` version 3.6.1 [49]. The simulations were carried out on Cartesius, the Dutch national supercomputer.

[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

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 are collected from various research areas to assess 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 ASReview's systematic review datasets GitHub[3].

The Wilson dataset [50] - from the field of medicine - is from a review on the effectiveness and safety of treatments of Wilson Disease, a rare genetic disorder of copper metabolism [33]. From the same field, the ACE dataset contains publications on the efficacy of Angiotensin-converting enzyme (ACE) inhibitors, a treatment drug for heart disease [32]. Additionally, the Virus dataset is from a systematic review on studies that performed viral Metagenomic Next-Generation Sequencing (mNGS) in farm animals [34]. From the field of computer science, the Software dataset contains publications from a review on fault prediction in software engineering [51]. The Nudging dataset [52] belongs to a systematic review on nudging healthcare professionals [35], stemming from the social sciences. From the same research area, the PTSD dataset contains publications on studies applying latent trajectory analyses on posttraumatic stress after exposure to traumatic events [36]. Of these six datasets, ACE and Software have been used for model simulations in previous studies on ML-aided title and abstract screening [11,32].

Data were preprocessed from their original source into a dataset, containing title and abstract of the publications obtained in the initial search. Duplicates and publications with missing abstracts were removed from the data. Datasets were labelled to indicate which candidate publications were included in the systematic review, thereby denoting relevant publications. All datasets consisted of thousands of candidate publications, of which only a fraction was deemed relevant to the systematic review. For the Virus and the Nudging dataset, this proportion was about 5 percent. For the remaining six datasets, the proportions of relevant publications were centered around 1-2 percent. (Table 1).

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

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

[3]https://github.com/asreview/systematic-review-datasets

dataset	Candidate publications	Relevant publications	Proportion relevant (%)
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

needed to be screened, at a given level of recall [32]. Typically measured at a recall level of 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 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 and the distribution of the data. Moreover, WSS makes assumptions about the acceptable recall level whereas this level might depend on the research question at hand [7]. Therefore, we introduce the ATD, the average fraction of non-reviewed relevant publications during the review (except the relevant publications in the initial training set). The ATD is an indicator of performance throughout the entire screening process instead of performance at some arbitrary cutoff value. The ATD is computed by taking the average of the Time to Discovery (TD) of all relevant publications. The TD for a given relevant publication i is computed as the fraction of publications needed to screen to detect i. Additional file 2 provides a detailed script to compute the ATD.

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 standard deviation $\hat{s}$. 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 plotted for each simulation, representing the average recall over 15 trials $\pm$ the standard error of the mean.

Results

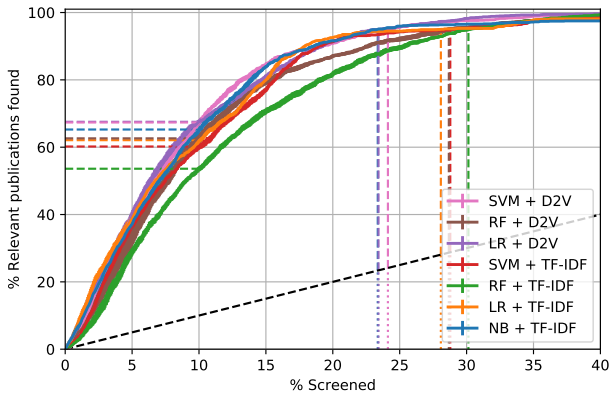
This section proceeds as follows: Firstly, as an example the results of the Nudging dataset are discussed in detail to provide a basis for answering the research questions. Secondly, the results are presented for each research question over all datasets.

Evaluation on the Nudging dataset

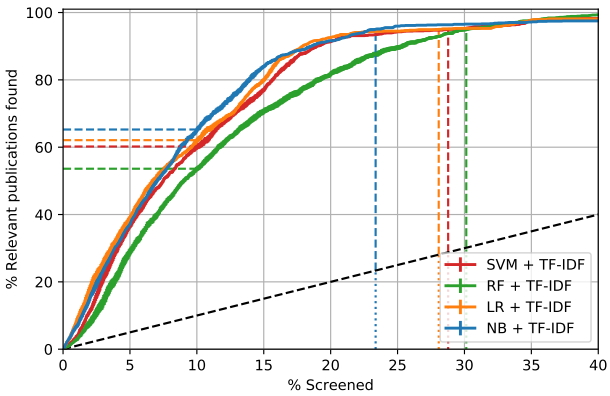
Figure 1a shows the recall curves for all simulations on the Nudging dataset. As described in the previous section, these curves plot recall as a function of the proportion of publications screened. The curves represent the average recall over 15 trials ± the standard error of the mean in the direction 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 black diagonal line corresponds to the expected recall curve when publications are screened in a random order.

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 relatively similar performance at face value.

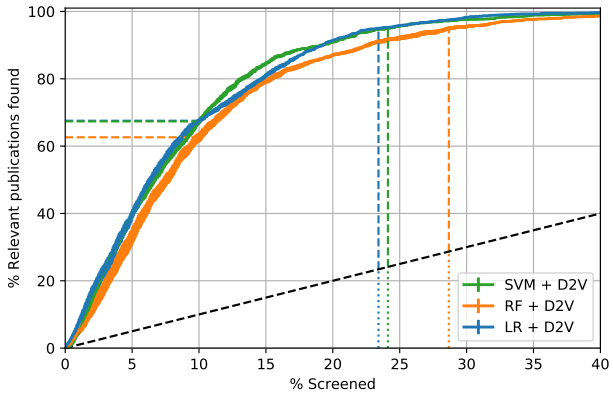
Figure 1b shows a subset of the recall curves in Figure 1a, namely the curves of the first four models to allow for a visual comparison across classification techniques adopting the TF-IDF feature extraction strategy. Figure 1c shows recall curves for the remaining three models to compare the models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



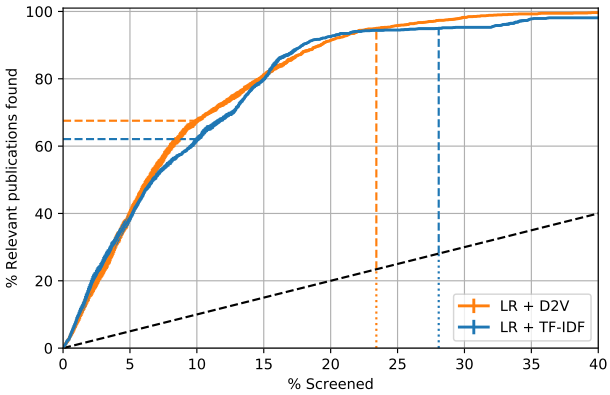
(a) All seven models



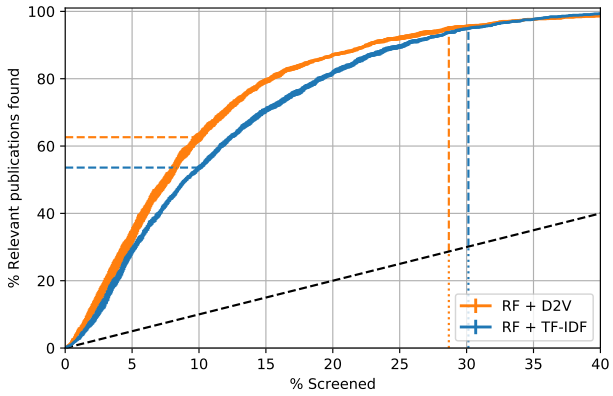
(b) Models adopting TF-IDF feature extraction



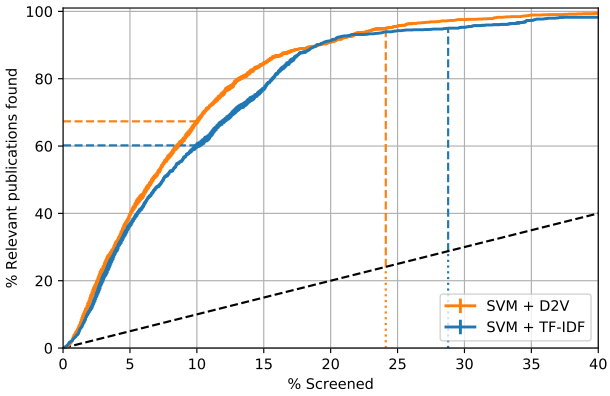
(c) Models adopting D2V feature extraction



(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. Figures b to f display curves for several subsets of those models to allow for a more detailed inspection of model performance.

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 overall 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 percentage point distance from the median ATD.

Table 2: ATD values ($\bar{x}(\hat{s})$) for all model-dataset combinations. For every dataset, the best results are in bold. Median (MAD) is given 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 by 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 of all datasets, the WSS@95 values of the models simulated on the Nudging dataset varied the most within the Nudging dataset.

Table 3: WSS@95 values ($\bar{x}(\hat{s})$) for all model-dataset combinations. For every dataset, the best results are in bold. Median (MAD) is given for all datasets.

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 of all datasets, the RRF@10 values were most dispersed for models simulated on the Nudging dataset.

	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)

Table 4: RRF@10 values $(\bar{x}, (\hat{s}))$ for all model-dataset combinations. For every dataset, the best results are in bold. Median (MAD) is given for all datasets.

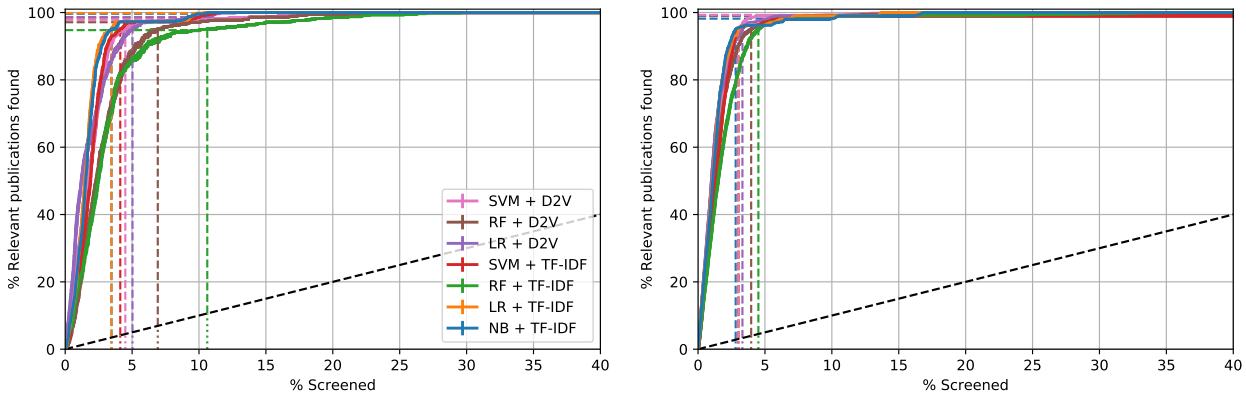
	Nudging	PTSD	Software	ACE	Virus	Wilson
SVM + TF-IDF	60.2 (3.12)	98.6 (1.40)	99.0 (0.00)	86.2 (5.25)	73.4 (1.62)	90.6 (1.17)
NB + TF-IDF	65.3 (2.61)	99.6 (0.95)	98.2 (0.34)	**90.5 (1.40)**	**73.9 (1.70)**	87.3 (2.55)
RF + TF-IDF	53.6 (2.71)	94.8 (1.60)	99.0 (0.00)	82.3 (2.75)	62.1 (3.19)	86.7 (5.82)
LR + 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 for the Nudging dataset. Please refer to Additional file 3 for figures presenting subsets of recall curves for the remaining datasets, like in Figure 1b-f.

First of all, as the recall curves exceed the expected recall at screening at random order by far for all datasets, the models were able to detect the relevant publications much faster compared to when screening publications at random order. 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 1a), the ordering of recall curves changes throughout the screening process, indicating that some models perform better at the start of the screening phase whereas others models take the lead later on. Moreover, the ordering of models in the Nudging dataset (Figure 1a) is not replicated in 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 (a) the PTSD, (b) Software, (c) ACE, (d) Virus, and (e) Wilson dataset.

RQ1 - Comparison across classification techniques

The first research question was aimed at evaluating the 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 for the ACE dataset, in which 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 models showed the same WSS@95 value. 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 model demonstrates 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 dataset (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 models adopting D2V feature extraction always achieved a lower 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 is not replicated 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 recall curves for some datasets than for others. While performance of the PTSD (Figure 2a) and Software datasets (Figure 2b) was quite high, performance was much lower across models for the Nudging (Figure 1a) and Virus (Figure 2d) datasets. The models simulated on the PTSD and Software datasets 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 performed similarly In contrast, for the Nudging (Figure 1a), ACE (Figure 2c), and Wilson (Figure 2e) dataset, the recall curves are much further apart, indicating that model performance was more dependent on the adopted classification technique and feature extraction strategy. The MAD values of the ATD, WSS@95 and RRF@10 confirm that model performance is less spread out within the PTSD, Software, and Virus datasets than within the Nudging, ACE, and Wilson datasets. Moreover, the curves for the ACE (Figure 2c) and Wilson (Figure 2e) datasets show a larger standard error of the mean compared the other datasets.

Taken together, although model performance is very data-dependent, there does not seem to be a distinction in performance between the datasets from the biomedical sciences (ACE, Virus, and Wilson) and datasets from other fields (Nudging, PTSD, and Software).

Discussion

The current study evaluates the 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

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% [32], whereas the models in the current study achieved far superior WSS@95 values varying from 68.6% to 82.9% in this dataset. In another study [11] that did use active learning, the Software dataset was used for simulation and a WSS@95 value of 91% was reached, strikingly similar to the 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 important finding to emerge from these evaluations was that the NB + TF-IDF model consistently performed as one of the best models. Our results suggest that while SVM performed fairly well, the LR and NB classification techniques are good if not superior alternatives to this default classifier in software tools. Note that LR and NB were always good methods for text classification tasks [53].

Feature extraction strategy

The overall results on models adopting D2V versus TF-IDF feature extraction strategy remain inconclusive. According to our findings, models adopting D2V do not outperform models adopting the well-established TF-IDF feature extraction strategy. Given these results, preference goes out to the TF-IDF feature extraction technique as this relatively simple technique will lead to a model that is easier to interpret. Another advantage of this technique is its short computation time.

Research contexts

Difficulty of applying active learning is not confined to any particular research area. The suggestion that active learning is more difficult for datasets from the social sciences compared to data from the medical sciences [12] does not seem to be the case. A possible explanation for this is that this difficulty has to be attributed to factors more directly related to the systematic review at hand, such as the proportion of relevant publications or the complexity of inclusion criteria used to identify relevant publications [16,54]. 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 proportion of relevant publications, 5.4% and 5.0%, respectively.

Limitations and future research

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

Additionally, while the sample of datasets in the current study is diverse compared to previous studies, the sample size (n=6) does not allow for investigating how model performance relates to characteristics of the data, such as the proportion of relevant publications. To build more confidence in active learning models for screening publications, it is essential to identify how data characteristics affect model performance. Such a study requires more data on systematic reviews. Thus, a more thorough study depends on researchers to openly publish their systematic review datasets.

Moreover, the runtime of simulations varied widely across models, indicating that some models take longer to retrain after a publication has been labelled than other models. This 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 should take into account the retraining time of models.

Conclusions

Overall, the findings confirm the great potential of active learning models to reduce the workload for systematic reviews. 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 which factors cause 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 - Average Time to Discovery

D2V - Doc2vec

LR - Logistic regression

MAD - Median Absolute Deviation

ML - Machine Learning

NB - Naive Bayes

PTSD - Post Traumatic Stress Disorder

RF - Random forest

RRF - Relevant References Found

SD - Standard Deviation

SEM - Standard Error of the Mean

SVM - Support vector machines

TF-IDF - Term Frequency - Inverse Document Frequency

TPE - Tree of Parzen Estimators

TD - Time to Discovery

WSS - Work Saved over Sampling

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

All data and materials are stored in the GitHub repository for this paper[4]. This repository contains all systematic review datasets used during this study and their preprocessing scripts, scripts for the hyperparameter optimization, the simulations, the processing and analysis of the results of the simulations, and for the figures and tables in this paper. The raw output files of the simulation study are stored on the Open Science Framework page of this paper, https://osf.io/7mr2g/ and https://osf.io/ag2xp/.

Competing interests

The authors declare that they have no competing interests.

Funding

Access to the Cartesius supercomputer was granted by SURFsara (ID EINF-156). 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

RvdS, RS, JdB and GF designed the study. RS developed the DR balance strategy and ATD metric, and wrote the programs required for hyperparameter optimization and cloud computation. RS, JdB, and DO designed the architecture required for the simulation study. GF extracted and analyzed the data and drafted the manuscript. RvdS, AB, RS, DO, LT, and JdB assisted with writing the paper. LT, DO, AB, and RvdS provided domain knowledge. All authors read and approved the final manuscript.

Acknowledgements

We are 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.

[4]https://github.com/asreview/paper-evaluating-models-across-research-areas

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 [Internet]. 2015 [cited 2020 Feb 4];4:1. Available from: https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/2046-4053-4-1

2. Gough D, Elbourne D. Systematic Research Synthesis to Inform Policy, Practice and Democratic Debate. Soc Policy Soc [Internet]. Cambridge University Press; 2002 [cited 2020 Apr 21];1:225–36. Available from: http://www.cambridge.org/core/journals/social-policy-and-society/article/systematic-research-synthesis-to-inform-policy-practice-and-democratic-debate/0A84765767F7ED99E55EA1B30C642D73

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. Treating individuals - from 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 [Internet]. British Medical Journal Publishing Group; 2017 [cited 2020 Apr 21];7:e012545. Available from: https://bmjopen-bmj-com.proxy.library.uu.nl/content/7/2/e012545

5. Lau J. Editorial: Systematic review automation thematic series. Syst Rev [Internet]. 2019 [cited 2020 Apr 21];8:70. Available from: https://doi.org/10.1186/s13643-019-0974-z

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 [Internet]. 2020 [cited 2020 Jan 16];20:7. Available from: 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 [Internet]. 2015 [cited 2020 Feb 11];4:5. Available from: 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 [Internet]. Oxford Academic; 2009 [cited 2020 Apr 24];16:690–704. Available from: https://academic-oup-com.proxy.library.uu.nl/jamia/article/16/5/690/804676

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 [Internet]. 2014 [cited 2020 Feb 11];5:31–49. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1093

10. Yu Z, Menzies T. FAST2: An intelligent assistant for finding relevant papers. Expert Syst Appl [Internet]. 2019 [cited 2020 Mar 3];120:57–71. Available from: http://www.sciencedirect.com/science/article/pii/S0957417418307413

11. Yu Z, Kraft NA, Menzies T. Finding better active learners for faster literature reviews. Empir Softw Eng [Internet]. Springer Science and Business Media LLC; 2018;23:3161–86. Available from: http://dx.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 [Internet]. 2014 [cited 2020 Mar 4];51:242–53. Available from: http://www.sciencedirect.com/science/article/pii/S1532046414001439

13. Cormack GV, Grossman MR. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval [Internet]. Gold Coast, Queensland, Australia: Association for Computing Machinery; 2014 [cited 2020 Mar 4]. pp. 153–62. Available from: https://doi.org/10.1145/2600428.2609601

14. Cormack GV, Grossman MR. Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review. 2015 [cited 2020 Apr 29]; Available from: http://arxiv.org/abs/1504.06868

15. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinform [Internet]. 2010 [cited 2020 Mar 4];11:55. Available from: 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 [Internet]. 2018 [cited 2020 May 4];7:45. Available from: https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-018-0707-8

17. Settles B. Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning [Internet]. 2012 [cited 2020 Feb 4];6:1–114. Available from: http://www.morganclaypool.com/doi/abs/10.2200/S00429ED1V01Y201207AIM018

18. Settles B. Active Learning Literature Survey [Internet]. University of Wisconsin-Madison Department of Computer Sciences; 2009. Available from: https://minds.wisconsin.edu/handle/1793/60660

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

2020 May 6]; Available from: http://arxiv.org/abs/1801.09496

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

21. Yimin M. Text Classification on Imbalanced Data: Application to Systematic Reviews Automation [Internet]. University of Ottawa; 2007. Available from: http://dx.doi.org/10.20381/ruor-12126

22. Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium [Internet]. Miami, Florida, USA: Association for Computing Machinery; 2012 [cited 2020 Feb 26]. pp. 819–24. Available from: 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 [Internet]. 2018 [cited 2020 Feb 5];32:762–4. Available from: https://conbio.onlinelibrary.wiley.com/doi/abs/10.1111/cobi.13117

24. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyana web and mobile app for systematic reviews. Syst Rev [Internet]. 2016;5:210. Available from: http://dx.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 [Internet]. 2018 [cited 2020 Feb 5];9:470–88. Available from: https://onlinelibrary.wiley.com/doi/abs/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 [Internet]. 2009 [cited 2020 Mar 10];16:25–31. Available from: https://academic.oup.com/jamia/article-lookup/doi/10.1197/jamia.M2996

27. Aphinyanaphongs Y. Text Categorization Models for High-Quality Article Retrieval in Internal Medicine. J Am Med Inform Assoc [Internet]. 2004 [cited 2020 May 2];12:207–16. Available from: https://academic.oup.com/jamia/article-lookup/doi/10.1197/jamia.M1641

28. Aggarwal CC, Zhai C. A Survey of Text Classification Algorithms. In: Aggarwal CC, Zhai C, editors. Mining Text Data [Internet]. Boston, MA: Springer US; 2012 [cited 2020 May 2]. pp. 163–222. Available from: http://link.springer.com/10.1007/978-1-4614-3223-4_6

29. Zhang W, Yoshida T, Tang X. A comparative study of TF*IDF, LSI and multi-words for text classification. Expert Syst Appl [Internet]. 2011 [cited 2020 May 6];38:2758–65. Available from: http://www.sciencedirect.com/science/article/pii/S0957417410008626

30. Le QV, Mikolov T. Distributed Representations of Sentences and Documents. 2014 [cited 2020 Feb 4]; Available from: http://arxiv.org/abs/1405.4053

31. 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 [Internet]. Routledge; 2020;14:145–58. Available from: https://doi.org/10.1080/17437199.2020.1716198

32. Cohen AM, Hersh WR, Peterson K, Yen P-Y. Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. J Am Med Inform Assoc [Internet]. 2006 [cited 2020 Feb 24];13:206–19. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1447545/

33. 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 [Internet]. 2019 [cited 2020 Feb 20];39:2136–52. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/liv.14179

34. Kwok KTT, Nieuwenhuijse DF, Phan MVT, Koopmans MPG. Virus Metagenomics in Farm Animals: A Systematic Review. Viruses [Internet]. Multidisciplinary Digital Publishing Institute; 2020 [cited 2020 Mar 24];12:107. Available from: https://www.mdpi.com/1999-4915/12/1/107

35. 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.

36. 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 [Internet]. Routledge; 2017;24:451–67. Available from: https://doi.org/10.1080/10705511.2016.1247646

37. 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; Available from: http://arxiv.org/abs/2006.12166

38. 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 [Internet]. British Medical Journal Publishing Group; 2020 [cited 2020 Jun 29];369. Available from: http://www.bmj.com/content/369/bmj.m1328

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

40. Kremer J, Steenstrup Pedersen K, Igel C. Active learning with support vector machines. WIREs Data

Min Knowl Discov [Internet]. 2014;4:313–26. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1132

41. Zhang H. The Optimality of Naive Bayes. 2004.

42. Breiman L. Random Forests. Machine Learning [Internet]. 2001 [cited 2020 Feb 10];45:5–32. Available from: https://doi.org/10.1023/A:1010933404324

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

44. Shahmirzadi O, Lugowski A, Younge K. Text Similarity in Vector Space Models: A Comparative Study. 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA) [Internet]. 2019. pp. 659–66. Available from: 10.1109/ICMLA.2019.00120

45. Yang Y, Loog M. A benchmark and comparison of active learning for logistic regression. Pattern Recognition [Internet]. 2018 [cited 2020 Jul 27];83:401–15. Available from: http://www.sciencedirect.com/science/article/pii/S0031320318302140

46. Fu JH, Lee SL. Certainty-Enhanced Active Learning for Improving Imbalanced Data Classification. 2011 IEEE 11th International Conference on Data Mining Workshops [Internet]. Vancouver, BC, Canada: IEEE; 2011 [cited 2020 Apr 28]. pp. 405–12. Available from: http://ieeexplore.ieee.org/document/6137408/

47. van de Schoot R, de Bruin J, Schram R, Zahedi P, Kramer B, Ferdinands G, et al. ASReview: Active learning for systematic reviews. Zenodo; 2020;

48. Bergstra J, Yamins D, Cox D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. International Conference on Machine Learning [Internet]. 2013 [cited 2020 Jul 23]. pp. 115–23. Available from: http://proceedings.mlr.press/v28/bergstra13.html

49. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2019. Available from: https://www.R-project.org/

50. Appenzeller-Herzog C. Data from Comparative effectiveness of common therapies for Wilson disease: A systematic review and meta-analysis of controlled studies [Internet]. Zenodo; 2020. Available from: https://doi.org/10.5281/zenodo.3625931

51. 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.

52. Nagtegaal R, Tummers L, Noordegraaf M, Bekkers V. Nudging healthcare professionals towards evidence-based medicine: A systematic scoping review [Internet]. Harvard Dataverse; 2019. Available from: https:

//doi.org/10.7910/DVN/WMGPGZ

53. Mitchell TM. Does Machine Learning Really Work? AI Mag [Internet]. 1997 [cited 2020 Jul 10];18:11–1. Available from: https://www.aaai.org/ojs/index.php/aimagazine/article/view/1303

54. Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. Systematic Reviews [Internet]. 2015 [cited 2020 May 10];4:80. Available from: https://doi.org/10.1186/s13643-015-0067-6

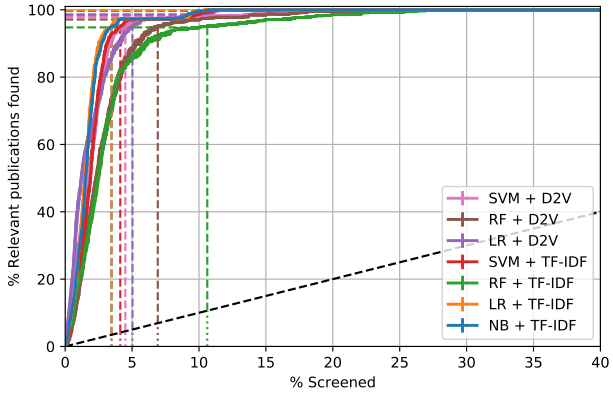
Additional file 1

Additional file 2

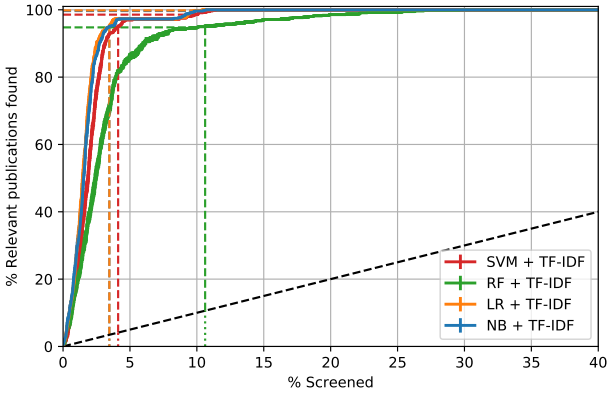
Additional file 3

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

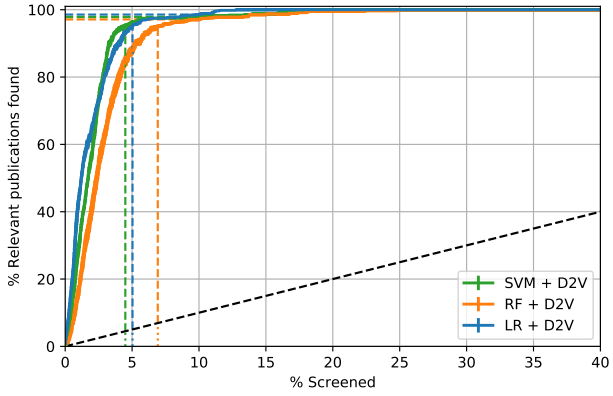
PTSD



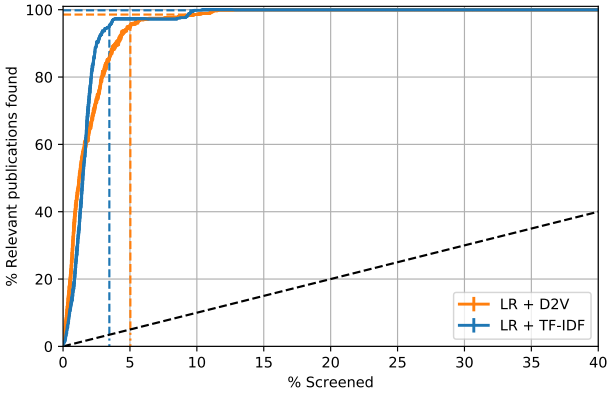
(a) All seven models



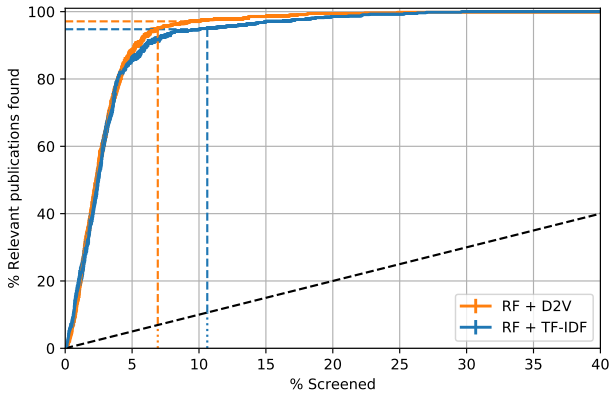
(b) Models adopting TF-IDF feature extraction



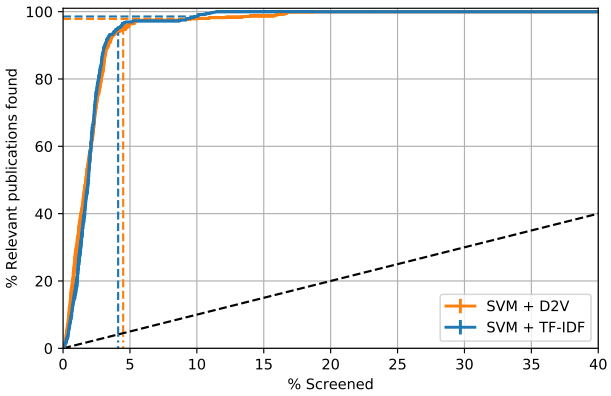
(c) Models adopting D2V feature extraction



(d) D2V vs TF-IDF for LR classifier



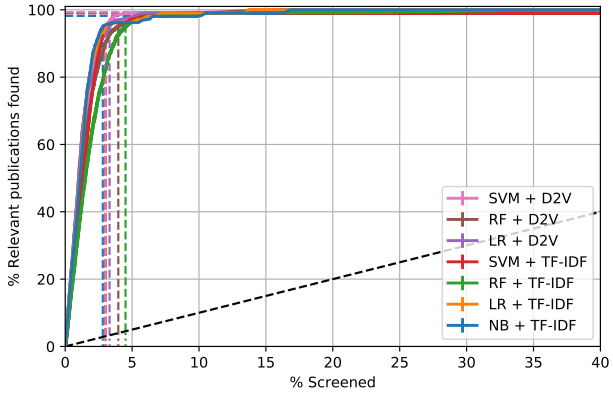
(e) D2V vs TF-IDF for RF classifier



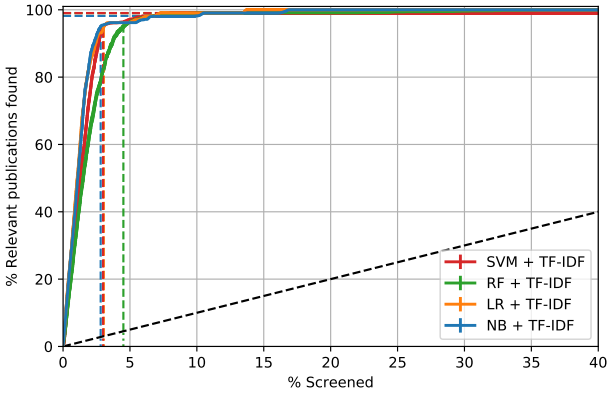
(f) D2V vs TF-IDF for SVM classifier

Figure 3: Recall curves for the PTSD dataset.

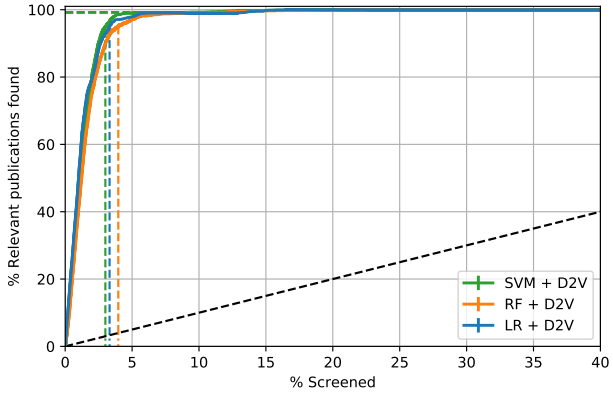
Software



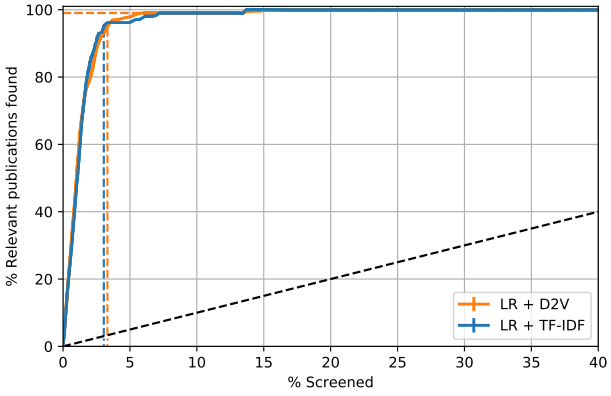
(a) All seven models



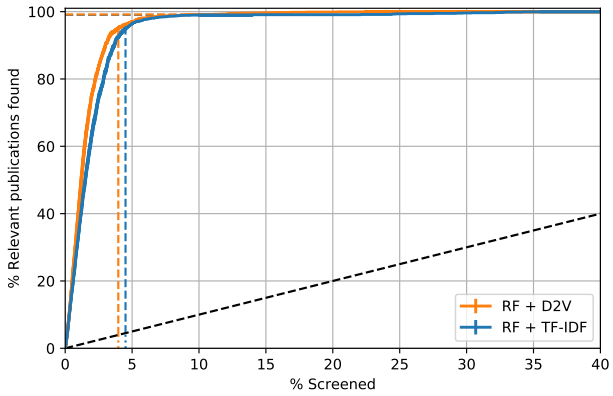
(b) Models adopting TF-IDF feature extraction



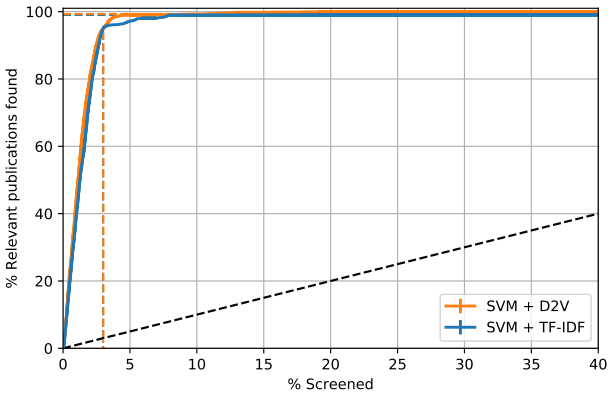
(c) Models adopting D2V feature extraction



(d) D2V vs TF-IDF for LR classifier



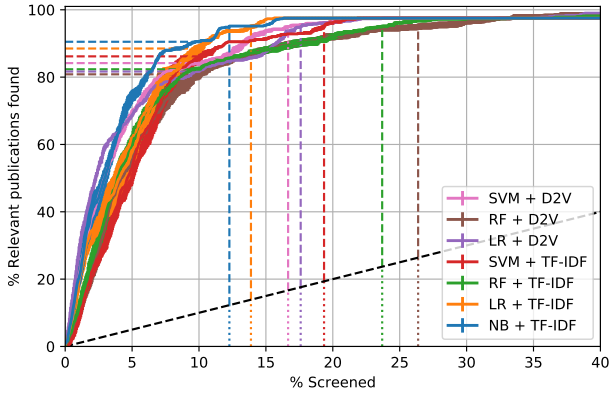
(e) D2V vs TF-IDF for RF classifier



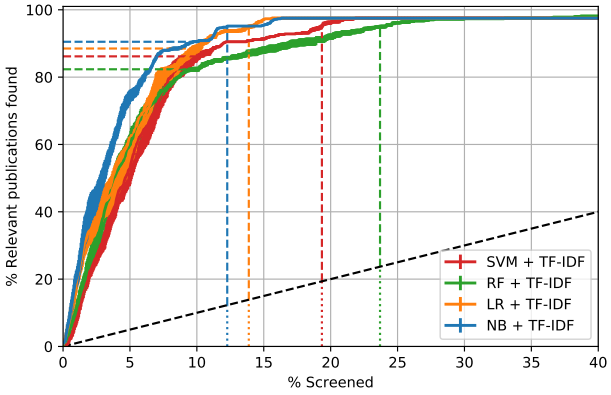
(f) D2V vs TF-IDF for SVM classifier

Figure 4: Recall curves for the Software dataset.

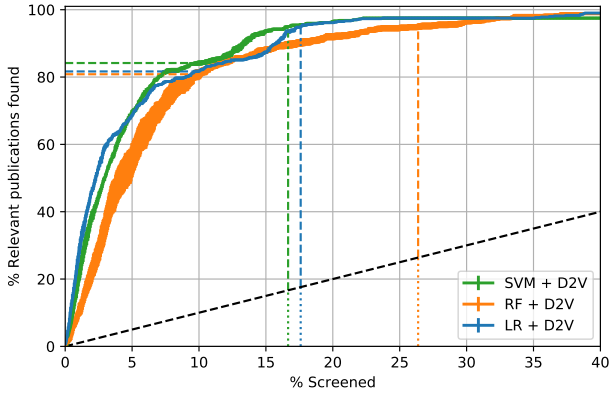
ACE



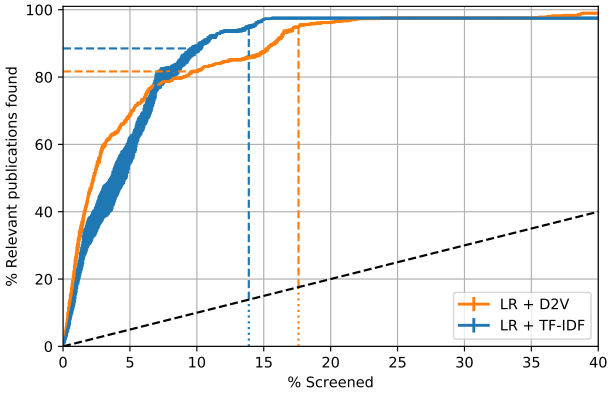
(a) All seven models



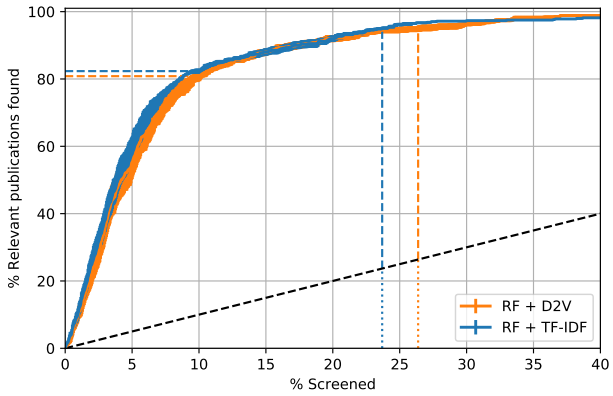
(b) Models adopting TF-IDF feature extraction



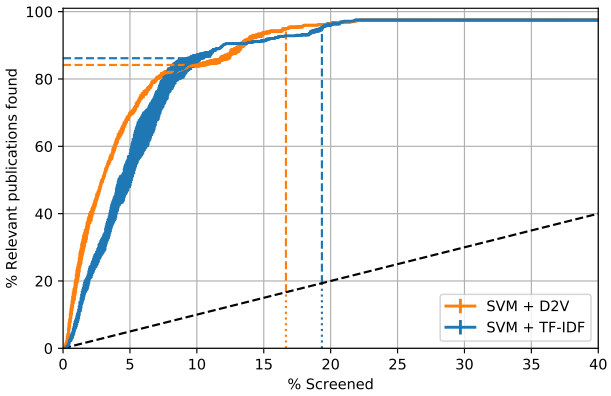
(c) Models adopting D2V feature extraction



(d) D2V vs TF-IDF for LR classifier



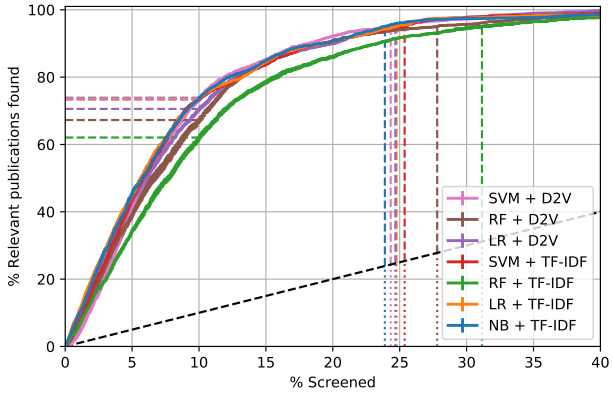
(e) D2V vs TF-IDF for RF classifier



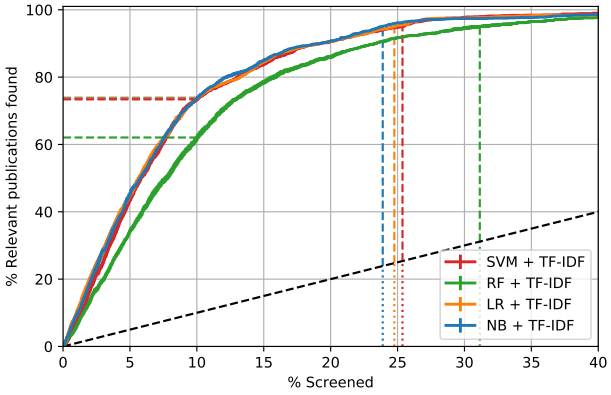
(f) D2V vs TF-IDF for SVM classifier

Figure 5: Recall curves for the ACE dataset.

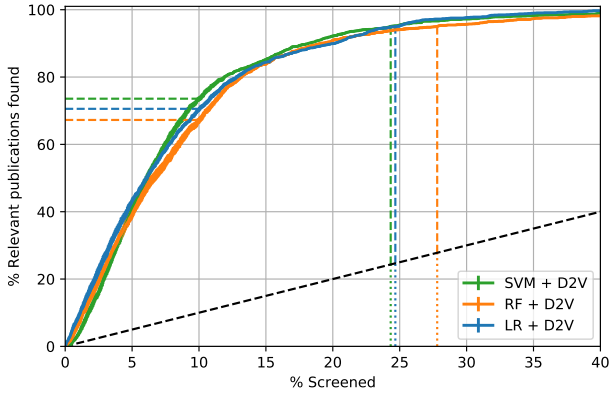
Virus



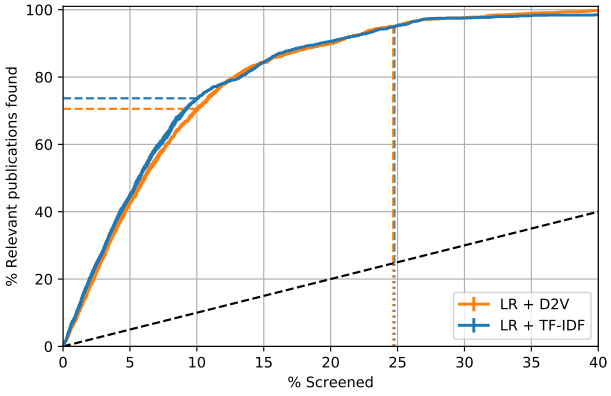
(a) All seven models



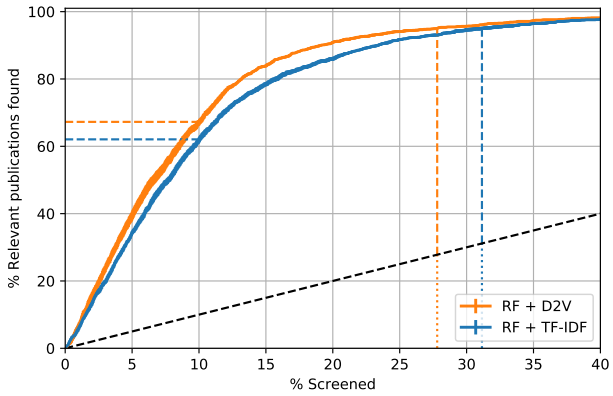
(b) Models adopting TF-IDF feature extraction



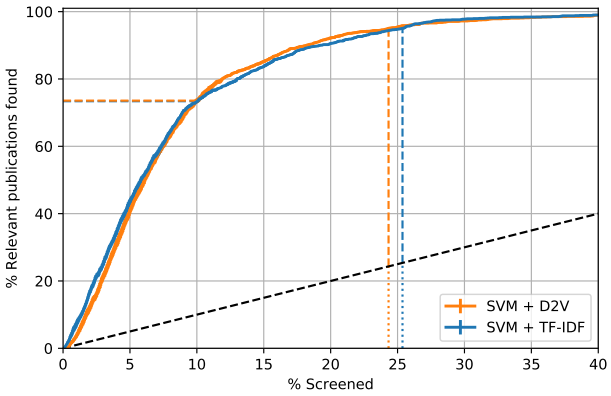
(c) Models adopting D2V feature extraction



(d) D2V vs TF-IDF for LR classifier



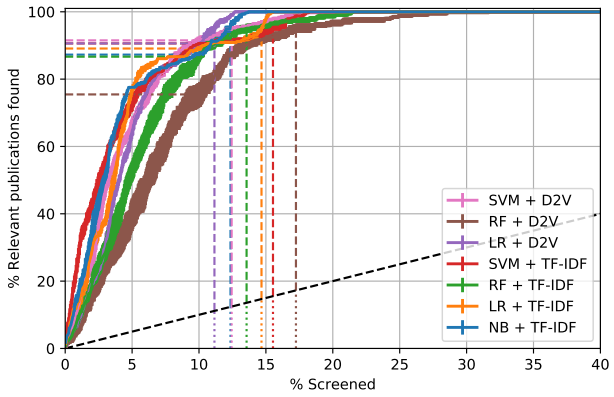
(e) D2V vs TF-IDF for RF classifaier



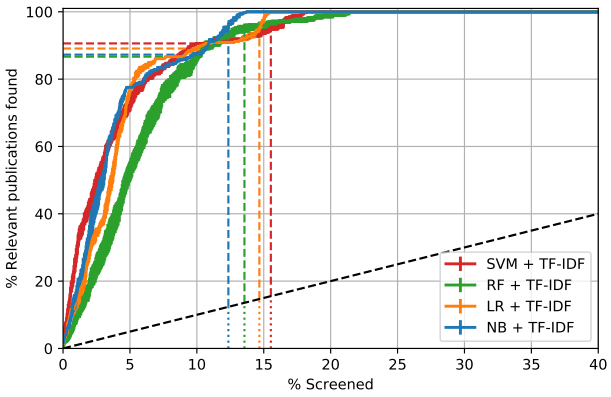
(f) D2V vs TF-IDF for SVM classifier

Figure 6: Recall curves for the Virus dataset.

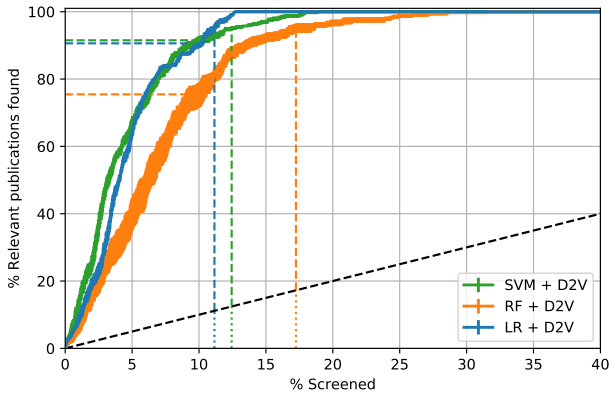
Wilson



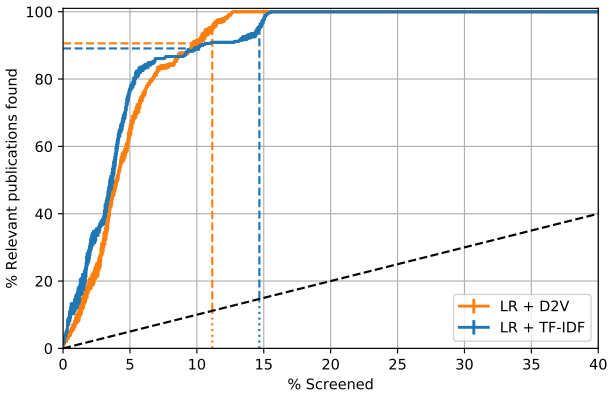
(a) All seven models



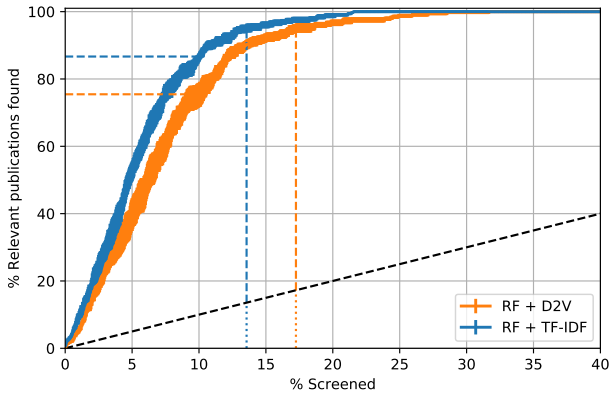
(b) Models adopting TF-IDF feature extraction



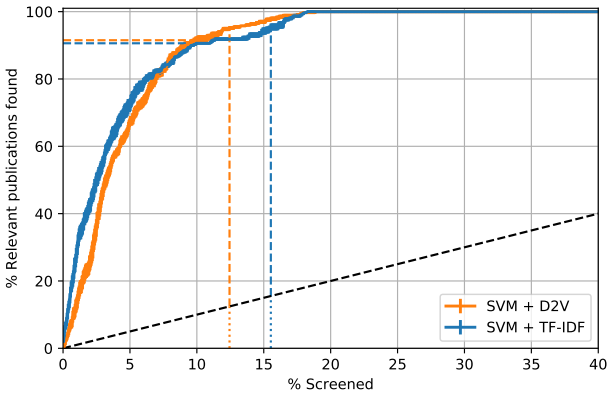
(c) Models adopting D2V feature extraction



(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.