


## RESEARCH ARTICLE

# Automating document classification with distant supervision to increase the efficiency of systematic reviews: A case study on identifying studies with HIV impacts on female sex workers

Xiaoxiao Li<sup>1</sup>✉, Amy Zhang<sup>1</sup>✉, Rabah Al-Zaidy<sup>2</sup>, Amrita Rao<sup>3</sup>, Stefan Baral<sup>3</sup>, Le Bao<sup>1\*</sup> , C. Lee Giles<sup>4</sup>

**1** Department of Statistics, Pennsylvania State University, University Park, PA, United States of America, **2** Information and Computer Science Department, King Fahad University of Petroleum and Minerals, Dhahran, Saudi Arabia, **3** Department of Epidemiology, Johns Hopkins University, Baltimore, MD, United States of America, **4** College of Information Sciences and Technology, Pennsylvania State University, University Park, PA, United States of America

✉ These authors contributed equally to this work.

\* [lebao@psu.edu](mailto:lebao@psu.edu)



## OPEN ACCESS

**Citation:** Li X, Zhang A, Al-Zaidy R, Rao A, Baral S, Bao L, et al. (2022) Automating document classification with distant supervision to increase the efficiency of systematic reviews: A case study on identifying studies with HIV impacts on female sex workers. *PLoS ONE* 17(6): e0270034. <https://doi.org/10.1371/journal.pone.0270034>

**Editor:** Andrej M Kielbassa, Danube Private University, AUSTRIA

**Received:** August 28, 2021

**Accepted:** June 2, 2022

**Published:** June 30, 2022

**Copyright:** © 2022 Li et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The dataset can be downloaded from <https://github.com/lebao0215/TextDataAbstracts>.

**Funding:** This research was supported by National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number R01AI136664. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Abstract

There remains a limited understanding of the HIV prevention and treatment needs among female sex workers in many parts of the world. Systematic reviews of existing literature can help fill this gap; however, well-done systematic reviews are time-demanding and labor-intensive. Here, we propose an automatic document classification approach to a systematic review to significantly reduce the effort in reviewing documents and optimizing empiric decision making. We first describe a manual document classification procedure that is used to curate a pertinent training dataset and then propose three classifiers: a keyword-guided method, a cluster analysis-based method, and a random forest approach that utilizes a large set of feature tokens. This approach is used to identify documents studying female sex workers that contain content relevant to either HIV or experienced violence. We compare the performance of the three classifiers by cross-validation in terms of area under the curve of the receiver operating characteristic and precision and recall plot, and found random forest approach reduces the amount of manual reading for our example by 80%; in sensitivity analysis, we found that even trained with only 10% of data, the classifier can still avoid reading 75% of future documents (68% of total) while retaining 80% of relevant documents. In sum, the automated procedure of document classification presented here could improve both the precision and efficiency of systematic reviews and facilitate live reviews, where reviews are updated regularly. We expect to obtain a reasonable classifier by taking 20% of retrieved documents as training samples. The proposed classifier could also be used for more meaningfully assembling literature in other research areas and for rapid documents screening with a tight schedule, such as COVID-related work during the crisis.

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

We are at a pivotal time in the global HIV response as progress towards ending the HIV pandemic by 2030 is off-track. A major reason is that not all have benefited from the advances in treatment and pre-exposure prophylaxis (PrEP) [1, 2]. An estimated 8% of new adult infections globally were among sex workers who are at 30 times greater risk of acquiring HIV than other reproductive-aged people [3]. Unfortunately, HIV-related data among female sex workers are still limited due to sustained individual and structural stigmas, including criminalization [4]. In both the Global Fund 2017–2021 strategy and President's Emergency Plan for AIDS Relief (PEPFAR) 3.0, the need for empirical data-driven responses was highlighted as central to informing an effective HIV response. As such, we are building data repositories for particular communities disproportionately affected by HIV, generally called key populations.

A comprehensive database that collects existing findings of sex workers can be leveraged to develop better strategies for optimizing the impact of HIV prevention and treatment programs [5]. Such a database could be created through a systematic literature review (SR), a key step of evidence-based public health programs, which distinguishes itself from ad hoc literature reviews and selection by its explicit and systematic approach [6]. Studies for specific diseases or interventions are collected, summarized, and extensively reported via SR to aid empiric decision-making by physicians, policymakers, and patients. Examples include the global burden of disease attributable to mental and substance use disorders [7], the long-term health consequences of child abuse [8], the effect of antibiotic prescribing on antimicrobial resistance [9], and so on. As such, guidelines for conducting and reporting SRs have been developed to improve the quality of SRs and increase their transparency. One notable example is the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), which consists of a 27-item checklist and four-phase diagram to help improve the reporting of systematic reviews [10]. There are five steps in forming an SR study [11]: (1) Formulating the research questions; (2) Identifying relevant work through an exhaustive search of the literature; (3) Assessing the quality of studies using; (4) Summarizing the evidence through tabulating study characteristics, quality, and effects as well as statistical methods for meta-analysis; (5) Interpreting the findings. Among those, Step (2) is particularly time-demanding and labor-intensive. To complete high-quality systematic reviews, one must invest significant efforts into developing and implementing appropriate search strategies, including searching through tens of thousands of research articles to include in an SR.

Here, we provide automatic document classification (ADC) procedures to reduce the manual labor involved in the systematic review process. We use a SR of identifying studies with HIV impacts on female sex workers as a case study, and this study represents the first attempt of automating the HIV and key population systematic reviews.

We first systematically reviewed all published literature related to female sex workers under the PRISMA guidelines, and then assembled data from studies published in all low and middle-income countries, including all Sub-Saharan African countries except for Seychelles. To compare with a traditional systematic review, we used ADC to identify journal articles that characterize key features of female sex workers as our baseline.

Multiple document classification models were implemented and evaluated for their performance by different metrics and sensitivity analysis. Our proposed method for document classification refines the high-dimensional encoding associated with bag-of-words to the  $N$  most important words for classification purposes, reducing the dimension of the feature space and improving performance. We demonstrated this using the random forest algorithm because of its ability to handle a large number of noisy predictors, to allow for high-order interaction effects, and to prevent over-fitting. The feature screening and extraction explored in the paper

can be generalized for different machine learning algorithms. We showed additional examples of Support Vector Machine (SVM), Boosting, Neural Network and ElasticNet, and compared their performance in our case study.

## Materials and methods

In this section, we first introduce the procedure for manually reviewing and labeling the status of articles. The results of manual document classification will be used as a gold standard and as a means to evaluate the performance of automated data classification methods. We then describe the natural language processing (NLP) tools for preprocessing the text data. The extracted tokens will serve as binary predictors for the automated classification algorithms. Finally, we propose a series of document classification models for automatically classifying documents into the relevant class versus the irrelevant class.

### Manual document classification and traditional systematic review methods

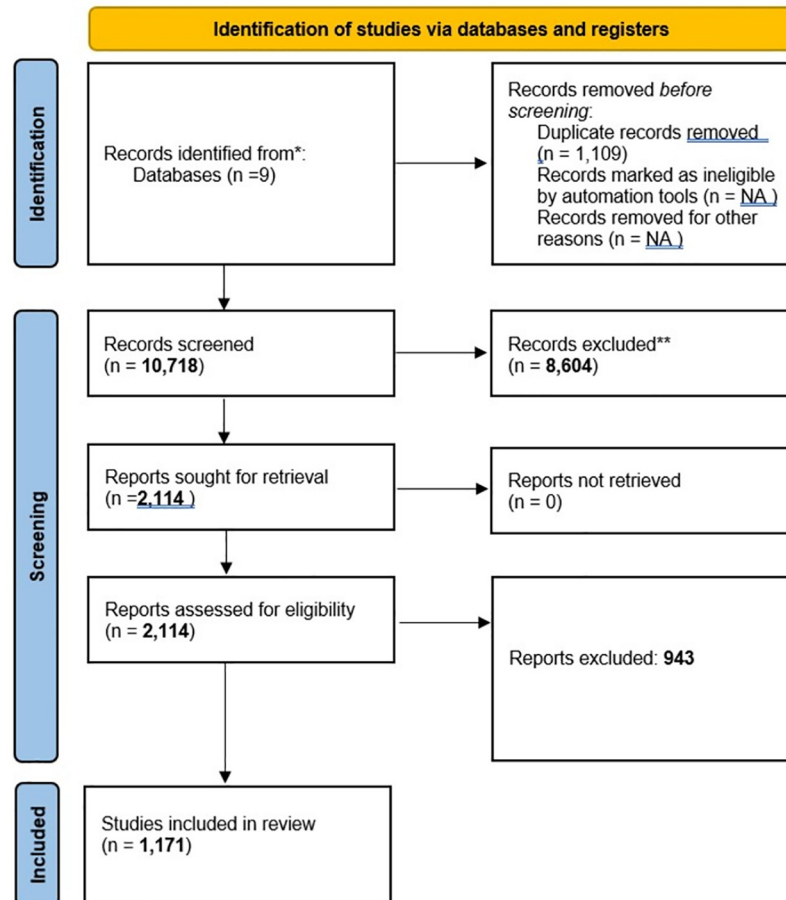
We obtained 10,718 journal articles which contained relevant keywords from PubMed, EMBASE, Global Health, SCOPUS, PsycINFO, Sociological Abstracts, CINAHL (Cumulative Index to Nursing and Allied Health Literature), Web of Science, and POPLINE. The search strategy and keywords used are detailed in the keyword-guided document classification model. The identified articles were labeled as relevant if they included HIV prevalence, the HIV treatment cascade (HIV testing, linkage to care, treatment, viral suppression), or experienced violence (physical, sexual, and intimate partner) among female sex workers and labeled as irrelevant otherwise. The labeling process consisted of an initial screening stage and a final selection stage. In the screening stage, a team of reviewers, two independent reviewers per article, using standardized processes reviewed titles, abstracts, and citation information to determine the article's relevance. The full text of articles marked as potentially relevant was then extracted and reviewed by the same team of reviewers, again using two independent reviewers per article, using a standardized approach to determine the final selection. Differences were resolved through consensus and referral to a senior study team member when necessary. Both the initial screening and final reviews were conducted using Covidence, a tool designed to help facilitate the systematic review process. The systematic review protocol has been published elsewhere [12] and is registered in the PROSPERO database (CRD42016047259; 28 September 2016). The flow diagram describing the screening and the review process and which follows PRISMA guidelines is presented in Fig 1.

In addition to subject matter constraints, all articles and reports used in the study must have been published in a peer-reviewed journal, presented as an abstract at a scientific conference, or available on the web from governmental or non-governmental sources between 2006 and 2017. Works published in languages other than English and studies where the sample size was less than 50 were not included. Some documents in the database had no abstract information and were also excluded during analysis.

### Data preprocessing for automatic document classification

Data preprocessing was conducted using the tidytext [13], quanteda [14], and tm [15] packages in R [16]. We combined the titles and abstracts of the documents and normalized the raw text data into tokens as follows:

**Tokenization:** We split the text into individual word tokens based on white space and segmented the text into basic linguistic units. All stop words, punctuation, numbers, special characters, and languages other than English were removed.



**Fig 1. Flow diagram for systematic reviews of HIV prevalence, the HIV treatment cascade (HIV testing, linkage to care, treatment, viral suppression) and experienced violence (physical, sexual, and intimate partner) among female sex workers.**

<https://doi.org/10.1371/journal.pone.0270034.g001>

**Lemmatization:** Following Mechura’s [17] English lemmatization list, we grouped together various derivative forms of a word which share similar semantic meaning such that they could be analyzed as a single item.

**Stemming:** We removed word suffixes and conflated the resulting morphemes with the Porter stemmer [18] which leads to a crude affix chopping. For example, “automates” and “automation” all reduce to “automat” using the Porter stemmer.

The tokens obtained after pre-processing for each document were translated into a numerical vector representation called the document-term-matrix (DTM). A typical representation of the DTM is a  $D \times P$  matrix, where  $D$  is the number of documents and  $P$  is the number of unique tokens after pre-processing. The  $(d, i)^{th}$  entry in the DTM records the frequency of token  $i$  in document  $d$ . Note that many tokens are common across documents but may not be relevant to the SR search criteria, such as articles, prepositions, and certain common verbs (e.g., “give”, “perform”). Thus we scaled token frequency within a document by the inverse of a token’s frequency across all documents, which is called term-frequency-inverse-document-frequency (TFIDF) [19]:

$$tfidf_{i,d} = tf_{i,d} \cdot idf_i, \quad (1)$$

where  $tf_{i,d}$  is the term-frequency for token  $i$  in document  $d$  and  $idf_i$  is the inverse document frequency for token  $i$ . For our study, we used the logarithmically scaled inverse fraction of documents containing token  $i$ :  $idf_i = \log \frac{N}{df(i)}$  where  $N$  is the total number of documents and  $df(i)$  is the frequency of token  $i$  in documents. Using TF-IDF, tokens which are common across all documents are treated as less informative and thus less important. This measures how much information the word provides.

## Document classification models

We compared three classifiers that automatically identify articles of interest based on query searching, clustering, and a broadly used machine learning algorithm. The first model is a keyword-guided filtering approach and serves as the baseline model. The second model refines the tokens extracted from the abstracts of the documents into related clusters and uses the clusters as features in a random forest (RF) model, which we refer to as “RF with refined clusters”. The third model includes both the refined clusters and additional tokens screened for relevance in the random forest model, which we refer to as “RF with top tokens”. These are described in detail in the following subsections.

**Keyword-Guided approach.** Our goal was to identify documents with studies of female sex workers, which included HIV or violence data. As a baseline model, we only examined document classification based on a keyword-guided filtering approach. We retrieved articles based on both their Medical Subject Headings (MeSH) and the occurrence of any related terms within the body of the article, referred to as a Text Word (TW) query. MeSH is a specific dictionary of keywords maintained by the National Library of Medicine to index and catalog biomedical information. A MeSH query retrieves articles that the author has listed as belonging to that category. We identified keywords across three relevant categories: female sex workers (FSW), human immunodeficiency virus (HIV), and violence (Violence). The three categories were combined in a query with the structure “FSW AND (HIV OR Violence)” under a Boolean search algorithm [20]. Table 1 lists MeSH and TW key words for each category.

**Cluster refinement.** The baseline keyword-guided model can be viewed as a tree-based classifier with three major predefined clusters of terms. We further divided those three large categories into finer clusters based on word roots, and allowed more flexible structures for classification. Constructing the finer clusters from the terms we extracted was a nontrivial process. Popular stemming methods truncate the ends of words, which often includes the removal of derivational affixes. Although the stemming process performs well in terms of reducing dimensionality, it also has the potential of creating ambiguity [21]. Due to the “crude chopping” of the stemming procedure, many outputs are not recognizable words. For example, “stay” becomes “stai” using the Porter stemmer. Also, affixes are essential in meaning in

**Table 1. Medical Subject Headings (MeSH) and Text Words (TW) used to retrieve relevant articles.**

Category	MeSH	TW
Female Sex Workers (FSW)	Prostitution, Sex Worker	prostitut*, commercial sex, transactional sex, sw, fsw, csw, sex trade, trade sex
HIV	HIV, acquired Immunodeficiency Syndrom, HIV Infections	human immunodeficiency virus*, acquired immunodeficiency syndrome*, HIV*, AIDS
Violence	Violence, Domestic Violence, Workplace Violence, Crime Victims, Battered Women, Rape, Homicide, Coercion	Violen*, crime*, offense*, abuse*, victim*, rape*, assault*, batter*, extort*, intimidat*, exploit*, IPV, IPSV

\* Represents usage of a wildcard operator in the database query, e.g. “violen\*” returns both “violence” and “violent”.

<https://doi.org/10.1371/journal.pone.0270034.t001>

English, but stemmers fail to capture this effect extensively; for example, “recondition” shares a stem but not the root meaning of “recondite.” These issues may result in inaccurate stemming of words that share the same root but have different meanings, such as “absolutely” and “absolution.” To mitigate this problem, we grouped words of similar semantic meaning to guarantee that the same root can represent major words in the three categories. As a result, we partitioned the three major categories of tokens into the following 15 “finer” clusters: “hiv”, “fsw”, “violence”, “offense”, “abuse”, “torture”, “rape”, “victim”, “assault”, “harass”, “extort”, “homicide”, “coercion”, “ipv”, “exploit”.

To obtain the vector representation of those 15 clusters, we created a document-cluster-matrix and calculated TF-IDF based on the combined frequencies of the terms in each cluster. Each document was represented by a 15-dimension vector of TF-IDFs. We then used random forest [22] to classify the documents with 15 features. Random forest uses a bagging ensemble method and decision trees constructed by a subset of data to provide more stable and accurate classification results.

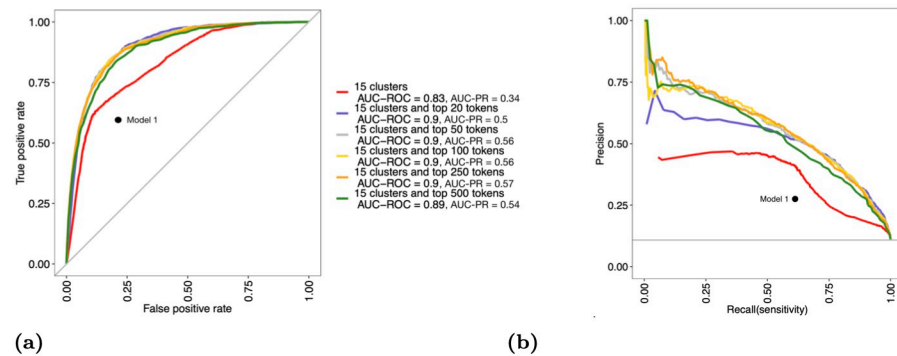
The two classes of documents were highly imbalanced: around one relevant document for every ten irrelevant ones. To relieve the distortion, we randomly down-sampled observations from the irrelevant class and kept similar sizes between the two classes when training each branch of the classification tree. We generated 500 classification trees, counted the votes for document  $d$  being relevant,  $Y_d = 1$ , among the 500 constructed trees, and used the proportion of votes as the fitted probability of  $\hat{P}(Y_d = 1)$ . The R Package we use for implementing the random forest algorithm is **RandomForest** [23].

**Top N tokens model.** In addition to the three major categories for tokens used in the baseline and cluster-refinement models, there still may have been some tokens that contained important information for classification purposes. We introduced a feature screening method to identify other significant tokens and included these tokens as additional covariates in the classification model. For each token  $i$ , we used a two-sample t-test to determine whether its mean TF-IDF differed statistically between the relevant class and the irrelevant class. The test statistic was  $\bar{x}_{i1} - \bar{x}_{i2}$  divided by the un-pooled variance, where  $\bar{x}_{i1}$  is the mean TF-IDF value for token  $i$  within the relevant documents and  $\bar{x}_{i2}$  the mean TF-IDF value for token  $i$  within the irrelevant documents. If a token  $i$  was useful for distinguishing between the two groups/classes, the t-statistic was expected to have a larger absolute value. We ranked tokens based on their absolute t-statistics and picked the top 20, 50, 100, 250, and 500 tokens as new features in addition to the 15 clusters defined in the cluster-refinement model and constructed a set of models using these top N features.

In addition to RF with refined clusters and RF with top tokens, we also explored a few other classifiers that could handle a large number of predictors. They are Support Vector Machine (SVM) implemented in the R Package **e1071** [24], Boosting implemented in the R Package **ada** [25], Neural Network implemented in the R Package **RSNNS** (Stuttgart Neural Network Simulator) [26], and ElasticNet implemented in the R Package **glmnet** [27]. All of the R packages were implemented in R environment, version 4.0.3 [16].

We trained each model using 5-fold cross-validation and evaluated their performance using the manually-assigned labels, described in the Data section, as ground truth. We used the receiver operating characteristic (ROC) curve and precision and recall (PR) curve as evaluation methods. ROC curve plots the true positive rate against the false positive rate; PR curve plots the precision rate against recall rate. Both ROC and PR curve illustrate the diagnostic ability of a binary classifier with varying discrimination threshold. All models except the baseline model produce probabilistic classifications; thus, we could additionally compare ROC and PR curves by varying the cut-off probability for classifying document  $d$  as relevant. We also





**Fig 2. (a) ROC and (b) PR curves for the baseline model (keyword-guided approach), RF with 15 refined clusters and RF with top tokens (20, 50, 100, 250, and 500 tokens in addition to 15 clusters).** As a summary metric, area under the curve (AUC) is provided in the legend.

<https://doi.org/10.1371/journal.pone.0270034.g002>

reported the area under the curve (AUC) as a summary metric for these models, i.e. the integration of the area under ROC curve or PR curve. We used the caret [28] package in R, which includes convenient methods for re-sampling, model training, parameter tuning, and cross-validation for a large variety of models.

## Results

Following the manual document classification procedure, we initially obtained 10,718 documents through keyword search, and identified 1,171 papers as relevant and 9,547 as irrelevant. We then evaluated the performance of automatic document classification (ADC) procedures by taking the manual classification results as the gold standard.

Fig 2(a) shows the ROC curves of all models for comparison. RF with 15 refined clusters outperforms the keyword-guided approach of the baseline model as the ROC curve of RF (red) is well above the point that corresponds to the deterministic classification of the baseline model. Table 2 shows that RF with top tokens further improves the area under the ROC curve (AUC-ROC) from 0.83 to 0.90. Varying the number of additional tokens (from  $N = 20$  to  $N = 500$ ) in RF does not obviously change its performance.

As our study contains two classes of data that are imbalanced, AUC-ROC is not capable of fully reflecting model performance [29]. Fig 2(b) shows each model's precision scores (the fraction of relevant articles among the retrieved articles) and recall scores (the fraction of the total amount of relevant articles that were actually retrieved) which better describe each model's performance under data imbalance. The baseline model attained a precision of 0.268 and a recall of 0.645. This suggests that the linear Boolean search criterion did not well express the differences between the relevant and irrelevant documents. RF with refined clusters outperformed the baseline model at the same recall level and provided a precision of 0.268 and a recall of 0.731. Table 2 shows that RF with additional 20 important tokens improved AUC-PR from 0.34 to 0.50 compared to RF with refined clusters. Increasing the number of important

**Table 2. The area under the curve (AUC) for ROC and PR of random forest models.** All models include 15 refined clusters. The additional number of tokens is shown as the column names.

Metric	0 tokens	20 tokens	50 tokens	100 tokens	250 tokens
AUC-ROC	0.83	0.90	0.90	0.90	0.90
AUC-PR	0.34	0.50	0.56	0.56	0.57

<https://doi.org/10.1371/journal.pone.0270034.t002>

**Table 3. The area under the curve (AUC) for ROC and PR of Random Forest, ElasticNet, Support Vector Machine (SVM), Neural Network and Boosting, all with top 50 tokens.** They are presented in the decreasing order of AUC-PR.

Metric	Random Forest	ElasticNet	SVM	Neural Network	Boosting
AUC-ROC	0.90	0.89	0.88	0.85	0.86
AUC-PR	0.56	0.49	0.44	0.36	0.12

<https://doi.org/10.1371/journal.pone.0270034.t003>

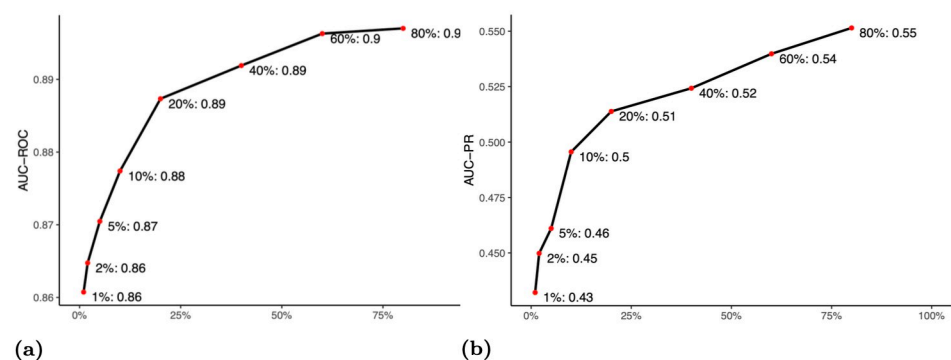
tokens to 50 and 100 improved AUC-PR to 0.56, while increasing to 250 tokens achieved a peak of 0.57.

Under RF with top tokens, a recall of 0.8 can be achieved while maintaining a precision score of 0.4. This reduces the amount of manual reading for our example by 80%. Our example consists of about 10,000 total documents, about 1,000 of which are truly relevant, thus for RF with top tokens to correctly identify 800 relevant articles (80% of 1,000 truly relevant cases),  $800/0.4 = 2,000$  documents will be labeled as potentially relevant. Researchers can then manually read the 2,000 labeled documents to identify the 800 relevant articles, reducing the amount of manual reading from 10,000 to 2,000. Other cut-off probabilities could be explored and would lead to different combination precision and recall scores as illustrated in Fig 2(b). In practice, people could rank the probability of relevance,  $\hat{P}(Y_d = 1)$ , and then prioritize manual reading of the highest probabilities. A stopping point for manual reading can be obtained through a heuristic, such as that used in SWIFT-Active Screening [30]. The above 5-fold cross-validation results reflect the model performance trained on 80% of the documents (10, 718 in total).

Finally, we compared random forest with other popular classifiers. As shown in Table 3, all with top 50 important tokens (features), random forest outperformed the SVM, Boosting (AdaBoost), Neural Network (multi-layer perceptron), and ElasticNet in terms of AUC-ROC and AUC-PR.

Next, we investigated the training data sample size needed to achieve a good model performance. We use report AUC values for both ROC and PR curves. We trained RF with the top 250 significant tokens on different proportions of pre-labeled documents (10, 718 in total) from 1% up to 80% and used cross-validation to estimate the prediction accuracy on the test data. Fig 3 shows that we can attain an AUC for ROC above 80% using only 1% of pre-labeled documents (107 samples) as training data. The improvements for AUC-ROC and AUC-PR are marginal when 2, 144 (20% among 10, 718) or more pre-labeled documents were used.

It is useful to understand model performance from the perspective of the percentage of future work saved. The percentage of work saved oversampling at random is a metric



**Fig 3. AUC for (a) ROC and (b) PR on testing data for RF with top 250 tokens with proportion of pre-labeled documents ranging from 1% to 80%.**

<https://doi.org/10.1371/journal.pone.0270034.g003>



commonly used to evaluate ADC models that screen documents for systematic reviews [31], often called  $WSS@R$ , where  $R$  is the desired recall level. If  $P$  is the total number of documents in the test data that were predicted relevant and  $N$  is the total number of test documents, then

$$WSS@R = R - \frac{P}{N}. \quad (2)$$

For RF with top tokens trained on 107 randomly selected training samples (1% of data), the average  $WSS@95$  is 37%, and the average  $WSS@80$  is 48%. With 1,070 training samples (10% of data), the average  $WSS@95$  is 42% and the average  $WSS@80$  is 55%. As such, researchers can avoid reading 75% of future documents (68% of total) while retaining 80% of relevant documents.

## Discussion

Automatic document classification (ADC) can help distinguish relevant work from others, and many techniques have been developed and applied in this field, such as the Naïve Bayes classifier [32], support vector machines [33], and deep neural network methods [34]. As such, many papers in recent years have reviewed ADC techniques, discussing how they may be incorporated in the systematic review process [35–38]. The majority of ADC methods for screening papers in systematic reviews use the bag-of-words encoding of titles and abstracts, with a variety of classification algorithms of which support vector machine (SVM) [39, 40], Naïve Bayes, and ensemble methods [41] are the most widely-used [42]. We did not intend to provide a comprehensive comparison of ADC techniques, but instead we presented a case study where the feature screening and extraction can be extended to other classifiers and SR of other domain fields. We also illustrated where improvement of prediction accuracy came from by incrementally increasing the model complexity. New advances in ADC techniques would provide future improvement potentials for similar tasks.

By comparing the AUC from both ROC and PR curves, we found RF with top tokens is outperforming the other models. The sensitive analysis on training size also indicates that even with 10% of data being trained, RF with top tokens can still greatly reduce the human reading while keeping the recall rate as high as 80%.

The trained classifiers can also be used as a sanity test, as mistakes are inevitable when manually screening thousands of documents. Any disagreements between manual and automatic classification would be addressed with an additional layer of review. For instance, in some cases, our model assigned high probabilities of relevant/irrelevant to documents manually labeled as the opposite. We provided 40 such documents to the manual screening team for verification: 18 documents labeled as irrelevant and 22 documents labeled as relevant. Among the 18 marked irrelevant documents, five are indeed relevant but missed during the manual labeling process; among the 22 marked relevant documents, 15 of them are identified as “likely should not have been included”.

Note that we excluded papers that were not published in English; these results on using an automatic document classifier may not be generalizable to non-English language studies.

## Conclusions

The need to extract information through a systematic literature review arises in a wide range of domains. Here, we proposed multiple models for automatically identifying documents that characterize key features of female sex workers. Empirical results showed that using a random forest on semantic clusters of key tokens and an additional set of tokens outperformed others.

RF with top tokens could be applied to identifying other populations most affected by HIV, including clients of female sex workers, gay men and other men who have sex with men, people who use drugs, transgender populations, and incarcerated populations. We expect to obtain a reasonable classifier by taking 20% of retrieved documents as training samples. The proposed classifier could also be used in other research areas and for rapid documents screening on a tight schedule, such as COVID-related work during the crisis. We feel methods proposed here will help other researchers with such reviews making them faster to implement and more complete and enable reviews to readily scale to much larger numbers of documents.

## Supporting information

### S1 Appendix.

(PDF)

### S1 Checklist. PRISMA 2020 checklist.

(PDF)

## Acknowledgments

Authors are grateful for helpful discussions with Jian Wu and Kunho Kim. Authors are also thankful to the editor, associate editor and two anonymous reviewers for their constructive suggestions and comments.

## Author Contributions

**Data curation:** Amrita Rao.

**Formal analysis:** Xiaoxiao Li, Amy Zhang.

**Methodology:** Rabah Al-Zaidy, Le Bao, C. Lee Giles.

**Project administration:** Le Bao.

**Resources:** Stefan Baral.

**Supervision:** Le Bao.

**Validation:** Xiaoxiao Li.

**Writing – original draft:** Xiaoxiao Li, Le Bao.

**Writing – review & editing:** Amy Zhang, Rabah Al-Zaidy, Amrita Rao, Stefan Baral, Le Bao, C. Lee Giles.

## References

1. Fauci AS, Marston HD. Ending AIDS—is an HIV vaccine necessary? *New England Journal of Medicine*. 2014; 370(6):495–498. <https://doi.org/10.1056/NEJMp1313771> PMID: 24499210
2. UNAIDS. Global HIV & AIDS statistics—2020 fact sheet. Geneva: UNAIDS.; 2020.
3. UNAIDS. UNAIDS Data 2020. Geneva: UNAIDS.; 2020.
4. Mishra S, Boily MC, Schwartz S, Beyrer C, Blanchard JF, Moses S, et al. Data and methods to characterize the role of sex work and to inform sex work programs in generalized HIV epidemics: evidence to challenge assumptions. *Annals of Epidemiology*. 2016; 26(8):557–569. Epub 2016 Jun 15. <https://doi.org/10.1016/j.annepidem.2016.06.004> PMID: 27421700
5. Rice B, Sanchez T, Baral S, Mee P, Sabin K, Garcia-Calleja JM, et al. Know your epidemic, strengthen your response: Developing a new HIV surveillance architecture to guide HIV resource allocation and target decisions. *JMIR Public Health and Surveillance*. 2018; 4(1):e18. <https://doi.org/10.2196/publichealth.9386> PMID: 29444766

6. Gough D, Oliver S, Thomas J. An introduction to systematic reviews. Sage; 2017.
7. Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The lancet*. 2013; 382(9904):1575–1586. [https://doi.org/10.1016/S0140-6736\(13\)61611-6](https://doi.org/10.1016/S0140-6736(13)61611-6) PMID: 23993280
8. Norman RE, Byambaa M, De R, Butchart A, Scott J, Vos T. The long-term health consequences of child physical abuse, emotional abuse, and neglect: a systematic review and meta-analysis. *PLoS Med*. 2012; 9(11):e1001349. <https://doi.org/10.1371/journal.pmed.1001349> PMID: 23209385
9. Costelloe C, Metcalfe C, Lovering A, Mant D, Hay AD. Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients: systematic review and meta-analysis. *Bmj*. 2010; 340:c2096. <https://doi.org/10.1136/bmj.c2096> PMID: 20483949
10. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009; 6(7):e1000097. <https://doi.org/10.1371/journal.pmed.1000097> PMID: 19621072
11. Khan KS, Kunz R, Kleijnen J, Antes G. Five steps to conducting a systematic review. *Journal of the royal society of medicine*. 2003; 96(3):118–121. <https://doi.org/10.1258/jrsm.96.3.118> PMID: 12612111
12. Rao A, Schwartz S, Sabin K, Wheeler T, Zhao J, Hargreaves J, et al. HIV-related data among key populations to inform evidence-based responses: protocol of a systematic review. *Systematic reviews*. 2018; 7(1):1–7. <https://doi.org/10.1186/s13643-018-0894-3> PMID: 30509317
13. Silge J, Robinson D. tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software*. 2016; 1(3):37. <https://doi.org/10.21105/joss.00037>
14. Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, et al. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*. 2018; 3(30):774. <https://doi.org/10.21105/joss.00774>
15. Feinerer I. Introduction to the tm Package Text Mining in R. Accessible en ligne: <http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf>. 2013.
16. R Core Team. R: A Language and Environment for Statistical Computing; 2021. Available from: <https://www.R-project.org/>.
17. Mechura M. Data Structures in Lexicography: from Trees to Graphs. 2016; p. 97–104.
18. Porter MF. An algorithm for suffix stripping. *Program*. 2006 <https://doi.org/10.1002/asi.4630260207>
19. Jones KS. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*. 1972.
20. Angione PV. On the equivalence of Boolean and weighted searching based on the convertibility of query forms. *Journal of the American Society for Information Science (pre-1986)*. 1975; 26(2):112. <https://doi.org/10.1002/asi.4630260207>
21. Schofield A, Mimno D. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*. 2016; 4:287–300. [https://doi.org/10.1162/tacl\\_a\\_00099](https://doi.org/10.1162/tacl_a_00099)
22. Liaw A, Wiener M. Classification and Regression by randomForest. *R News*. 2002; 2(3):18–22.
23. Lin N, Wu B, Jansen R, Gerstein M, Zhao H. Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*. 2004; 5(154). <https://doi.org/10.1186/1471-2105-5-154> PMID: 15491499
24. Karatzoglou A, Meyer D, Hornik K. Support Vector Machines in R. *Journal of Statistical Software*. 2006; 15(9). <https://doi.org/10.18637/jss.v015.i09>
25. Culp M, Johnson K, Michailidis G. ada: An R Package for Stochastic Boosting. *Journal of Statistical Software*. 2006; 17(2):1–27. <https://doi.org/10.18637/jss.v017.i02>
26. Bergmeir C, Benitez JM. Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNN. *Journal of Statistical Software*. 2012; 46(7):1–26. <https://doi.org/10.18637/jss.v046.i07>
27. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*. 2010; 33(1):1–22. <https://doi.org/10.18637/jss.v033.i01> PMID: 20808728
28. Kuhn M. Building predictive models in R using the caret package. *Journal of Statistical Software*. 2008; 28(5):1–26. <https://doi.org/10.18637/jss.v028.i05>
29. Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*. 2016; 49(2):1–50. <https://doi.org/10.1145/2907070>
30. Howard BE, Phillips J, Tandon A, Maharana A, Elmore R, Mav D, et al. SWIFT-Active Screener: Accelerated document screening through active learning and integrated recall estimation. *Environment International*. 2020; 138:105623. <https://doi.org/10.1016/j.envint.2020.105623> PMID: 32203803

31. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*. 2006; 13(2):206–219. <https://doi.org/10.1197/jamia.M1929> PMID: 16357352
32. Agrawal R, Bayardo R, Srikant R. Athena: Mining-based interactive management of text databases. In: *International Conference on Extending Database Technology*. Springer; 2000. p. 365–379.
33. Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: *European conference on machine learning*. Springer; 1998. p. 137–142.
34. Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: *Proceedings of the 25th international conference on Machine learning*; 2008. p. 160–167. <https://doi.org/10.1145/1390156.1390177>.
35. Bannach-Brown A, Przybyła P, Thomas J, Rice AS, Ananiadou S, Liao J, et al. The use of text-mining and machine learning algorithms in systematic reviews: reducing workload in preclinical biomedical sciences and reducing human screening error. *BioRxiv*. 2018; p. 255760. <https://doi.org/10.1197/jamia.M1929>.
36. Thomas J, Noel-Storr A, Marshall I, Wallace B, McDonald S, Mavergames C, et al. Living systematic reviews: 2. Combining human and machine effort. *Journal of clinical epidemiology*. 2017; 91:31–37. <https://doi.org/10.1016/j.jclinepi.2017.08.011> PMID: 28912003
37. Shemilt I, Khan N, Park S, Thomas J. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic reviews*. 2016; 5(1):140. <https://doi.org/10.1186/s13643-016-0315-4> PMID: 27535658
38. 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. *Systematic reviews*. 2015; 4(1):5. Erratum in: *Syst Rev*. 2015;4:59. <https://doi.org/10.1186/2046-4053-4-5> PMID: 25588314
39. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Systematic reviews*. 2016; 5(1):210. <https://doi.org/10.1186/s13643-016-0384-4> PMID: 27919275
40. Przybyła P, Brockmeier AJ, Kontonatsios G, Le Pogam MA, McNaught J, von Elm E, et al. Prioritising references for systematic reviews with RobotAnalyst: a user study. *Research synthesis methods*. 2018; 9(3):470–488. <https://doi.org/10.1002/jrsm.1311> PMID: 29956486
41. van de Schoot R., de Bruin J., Schram R. et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nature machine intelligence*. 2021; 3:125–133. <https://doi.org/10.1038/s42256-020-00287-7>
42. Beller E, Clark J, Tsafnat G, Adams C, Diehl H, Lund H, et al. Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). *Systematic reviews*. 2018; 7(1):1–7. <https://doi.org/10.1186/s13643-018-0740-7>