Manuscript drafts

Gerbrich Ferdinands

22/04/2020

Keywords: Human-Machine interaction, active learning, systematic reviews, text classification

#### Introduction

#### Methods

#### Results

#### Discussion

# Introduction

Systematic reviews are top of the bill in building evidence in research. A systematic review brings together all studies relevant to answer a specific research question (PRISMA-P Group et al. 2015). Systematic reviews inform practice and policy (Gough and Elbourne 2002) and are key in developing clinical guidelines (Chalmers 2007).

However, systematic reviews are costly because they involve the manual screening of thousands of titles and abstracts, identifying publications relevant to answering the research question. An experienced reviewer takes on average 30 seconds to screen one title and abstract, whereas an inexperienced reviewer takes even longer (Wallace et al. 2010). Conducting a systematic review typically requires over a year of work by a team of researchers (Borah et al. 2017).

Moreover, 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 (Lau 2019). Especially when the need for guidelines is urgent - such as in the context of the current COVID-19 crisis - it is almost impossible to provide a review that is both timely and comprehensive. To ensure a timely review, reducing workload in systematic reviews is imperative.

With advances in Artificial Intelligence (AI), there has been wide interest in tools to reduce workload in systematic reviews (Harrison et al. 2020). Various learning models have been proposed, aiming to predict whether a given publication is relevant or irrelevant to the systematic review. Findings suggest that such models potentially reduce workload with 30-70% at the cost of losing 5% of relevant publications (95% recall) (O’Mara-Eves et al. 2015).

Screening prioritization is a well-established approach in increasing efficiency in title and abstract screening (Cohen, Ambert, and McDonagh 2009; Shemilt et al. 2014). In screening prioritization, the learning model reorders publications to be screened by their likeliness to be relevant. The model presents the reviewer with the publications which 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. Reviewing relevant publications early facilitates a faster transition of those publications to the next steps in the review process (Cohen, Ambert, and McDonagh 2009). Additionally, several studies report increasing efficiency beyond saving time (O’Mara-Eves et al. 2015).

Recent studies have demonstrated the effectiveness of screening prioritization by means of active learning models (Yu and Menzies 2019; Yu, Kraft, and Menzies 2018b; Miwa et al. 2014). Active learning is when the model can iteratively improve its predictions by allowing the model to choose the data from which it can learn (Settles 2012). Active learning has proven to be an efficient strategy in large datasets where labels are scarce, which makes identifying publications an ideal candidate for such models. When applied in screening prioritization, the reviewer screens publications that are presented by an active learning model. Subsequently, the active learning model learns from the reviewers’ decision (‘relevant’, ‘irrelevant’) and uses this knowledge in selecting the next publication to be screened by the reviewer.

Although the application of learning models in reducing workload of systematic reviews has been extensively studied, the complex nature of the field is making it difficult to draw overarching conclusions about best practice.

Moreover, the lack of replication on data outside the biomedical sciences makes it impossible to draw conclusions about the general effectiveness of such technologies (O’Mara-Eves et al. 2015; Marshall et al. 2020).

Up to now only few studies have investigated active learning models for screening prioritization, with promising results (Yu, Kraft, and Menzies 2018b; Miwa et al. 2014). The question remains how active learning models for screening prioritization perform across different (1) classification techniques and (2) review contexts. Hence, additional evaluations of active learning models are required.

The purpose of the current paper is to increase the evidence base of active learning models for reducing workload in title and abstract screening in systematic reviews.

We want to maximize the number of identified relevant publications, while minimizing the number of publications needed to screen

Combining latest insights from this area, we propse seven different active learning models. a pipeline of active learning for prioritization screening, called ASReview (Van de Schoot et al. 2020).

ASReview is an open source and generic tool such that users can adapt and add modules as they like, encouraging fellow researchers to replicate findings from previous studies.

Working towards a general consensus in this emerging field, the current study compares the performance of 7 active learning models for 6 labelled datasets from the fields of …

implements various active learning models. Models are evaluated by performing a simulation on data from six existing systematic reviews from various research areas.

All scripts and data used are openly published to facilitate usability and acceptability of AI-assisted title and abstract screening in the field of systematic review.

The remaining part of this paper is organized as follows. The Technical Details section will cover the workings of active learning models for study selection in systematic reviews on a conceptual level. The method section describes … The results section reports … The discusssion section summarises the findings, comments on them and summarises the main findings, discusses discuss limitations, draw conclusion and

# Methods

In the current study, seven active learning models were developed for the purpose of identifying relevant publications in systematic review datasets. The models were designed to maximize the number of identified relevant publications, while minimizing the number of publications needed to screen. Models were assessed by conducting a retrospective simulation on six systematic review datasets. Datasets were collected from various research areas to assess generalizability of the models across research contexts.

## Technical details

The screening process starts with all publications obtained in the search. The task is to identify which of these publications are relevant, which is typically done by manually screening all titles and abstracts.

In the current study/active learning, the screening process proceeds as follows:

todo: add figure demonstrating the active learning cycles (maybe move to introduction?)

In this active learning cycle, the model can incrementally improve its predictions on the relevancy of the remaining unlabeled publications. Based on previous labelling decisions by the reviewer, the model constantly reorders the remaining publications in the dataset. The reviewer starts labeling some instances in , creating labeled data set . This is the starting point for the active learning model. The model trains a classifier who predicts … Hyperparameters were optimized separately for every model \* dataset combination[[1]](#footnote-1).

The reviewer and the are interacting in the active learning cycle.

In the remainder of this section, the key components of the active learning models will be further explained.

#### Classification

To make predictions on the unlabeled publications, a classifier is trained on the set of previously labeled publications. A technique widely used in classification tasks is the Support Vector Machine (SVM). SVMs attempt to find a multidimensional hyperplane that separates the data into classes (Tong and Koller 2001). SVMs have been proven to be effective in active learning models for screening prioritization [Yu, Kraft, and Menzies (2018b); Miwa2014]. Moreover, SVMs are the currently the only classifier implemented in ready-to-use software tools implementing active learning for screening prioritization [abstrackr, colandr, fastread, rayyan, robotanalyst].

Whilst several classifiers have been employed in the AI-aided title and abstract screening field in general, the relatively new subfield of active learning for screening prioritization has not yet determined the performance of classifiers other than SVMs (Yu, Kraft, and Menzies 2018a; Yu and Menzies 2019; Miwa et al. 2014).

The current study aims to address this gap by exploring performance of three classifiers besides SVM:

* To do: add sources to justify decisions for these classification techniques
* Logistic Regression (LR) - L2-regularized logistic regression.
* Naive Bayes (NB) - Naive Bayes assumes all features are independent given the class value. This is obviously not the case but still the algorithm performs impressively (Zhang 2004). Especially at … tasks. also used in …
* Random Forests (RF) is where a large number of decision trees are fit on bootstrapped samples of the original data. All trees cast a vote on the class, which are aggregated into a class prediction for each instance (Breiman 2001).

#### Class imbalance

There are two classes in the dataset: relevant and irrelevant publications. Typically, only a fraction of the publications in the data belong to the relevant class. This poses a problem for training a classifier as there are far fewer examples of relevant than irrelevant publications to train on (O’Mara-Eves et al. 2015). Moreover, classifiers are well-suited to separate data into classes, but not to correctly identifying one class (Wallace et al. 2010). This is evident in the case of a dataset where only one percent of publications are relevant. A model would achieve 99% accuracy when classifying all publications as irrelevant, even though none of the relevant papers would have been identified. Therefore, the class imbalance problem causes the classifier to miss relevant publications.

Previous studies have addressed the class imbalance problem by rebalancing the training data in different ways (O’Mara-Eves et al. 2015). To decrease the class imbalance in the training data, the models in the current study rebalance the training set by Dynamic Supersampling (DS). DS decreases the number of irrelevant publications in the training data, whereas the number of relevant publications are increased (by copy) such that the size of the training data remains the same. The ratio between relevant and irrelevant publications is not fixed, but dynamically updated and depends on the size of the training data, the total number of publications, and the ratio between total number of relevant and irrelevant publications.

#### Word embeddings

To predict publication class, the classifier uses information from the publications in the dataset. Examples of such information are titles and abstracts. However, a model cannot predict class from the titles and abstracts as they are; their textual content needs to be repressented numerically. The textual information needs to be mapped to feature vectors. This process of numerically representing textual content is called ‘word embeddings’.

A classical example of word embeddings is a ‘bag of words’ (bow) representation. For each text in the data set, the number of occurrences of each word is stored. This leads to features, where is the number of distinct words in the texts (Pedregosa et al. 2011). The bag-of-words method is simplistic and will highly value often occuring but otherwise meaningless words such as “and”. A more sophisticated approach is Term-frequency Inverse Document Frequency (TF-IDF). TF-IDF circumvents this problem by adjusting the term frequency in a text with the inverse docuement frequency, the frequency of a given word in the entire data set (Ramos and others 2003). A downside of TF-IDF and other bow methods is that they do not take into account the ordering of the words, thereby ignoring semantics. An example of an approach that aims to overcome this weakness is Doc2Vec, capable of grasping the relations between words by learning to predict the words in the texts (Le and Mikolov 2014).

Miwa et al. found that active learning was more difficult on data from the social sciences compared to data the medical sciences and were able to link this difficulty to a natural difference in text complexity between these research areas (Miwa et al. 2014). As the study by Miwa et al. adopted a bow approach (Miwa et al. 2014), we hypothesize that a more sophisticated word embeddings strategy has the potential to bridge the performance gap between these research areas.

#### 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 (Settles 2012). Another well-known strategy is uncertainty-based active learning, where the instances that will be presented next will be those instances on which on which the model’s classifications are the least certain, i.e. close to 0.5 probability (Settles 2012). Traditionally, this strategy trains the most accurate model because the model can learn the most from instances it is uncertain about.

Even though uncertainty-based active learning generally leads to a more accurate model in the end, certainty-based active learning is the preferred strategy for the task at hand. The main reason is that this strategy is far better suited to the goal of prioritizing the relevant publications. Moreover, a study comparing performance of both strategies in found that the accuracy gain of uncertainty-based screening was not significant (Miwa et al. 2014). Furthermore, uncertainty-based active learning is far better equipped at dealing with imbalanced data in active learning (Fu and Lee 2011). Therefore, certainty-based sampling is most equipped to the current scenario in which we are dealing with highly imbalanced datasets and where the goal is to identify all relevant publications as soon as possible.

## Models

Seven models were built from the components described in the Technical Details section, adopting four different classification techniques and two different word embeddings approaches:

maybe describe as two-stage approach? first classfication technique evaluation and second word embeddings

* SVM + TF-IDF
* Naive Bayes + TF-IDF[[2]](#footnote-2)
* Random Forest + TF-IDF
* Support Vector Machine + TF-IDF
* SVM + Doc2Vec
* Naive Bayes + Doc2Vec
* Random Forest + Doc2Vec
* Support Vector Machine + Doc2Vec

## Simulation study

For each of the seven models, performance was evaluated by simulating the model on the screening process of six systematic reviews. Performance of the seven models was evaluated by simulating their behaviour in the screening process of six systematic review datasets. Put differently, 42 simulations were carried out. For every model - dataset combination, hyperparameters were optimized[[3]](#footnote-3). To account for variance, every simulation was repeated for 15 trials. Simulations were run using ASReview’s simulation mode (Van de Schoot et al. 2020). There was no need for a human reviewer as the model could query the labels in the data instead.

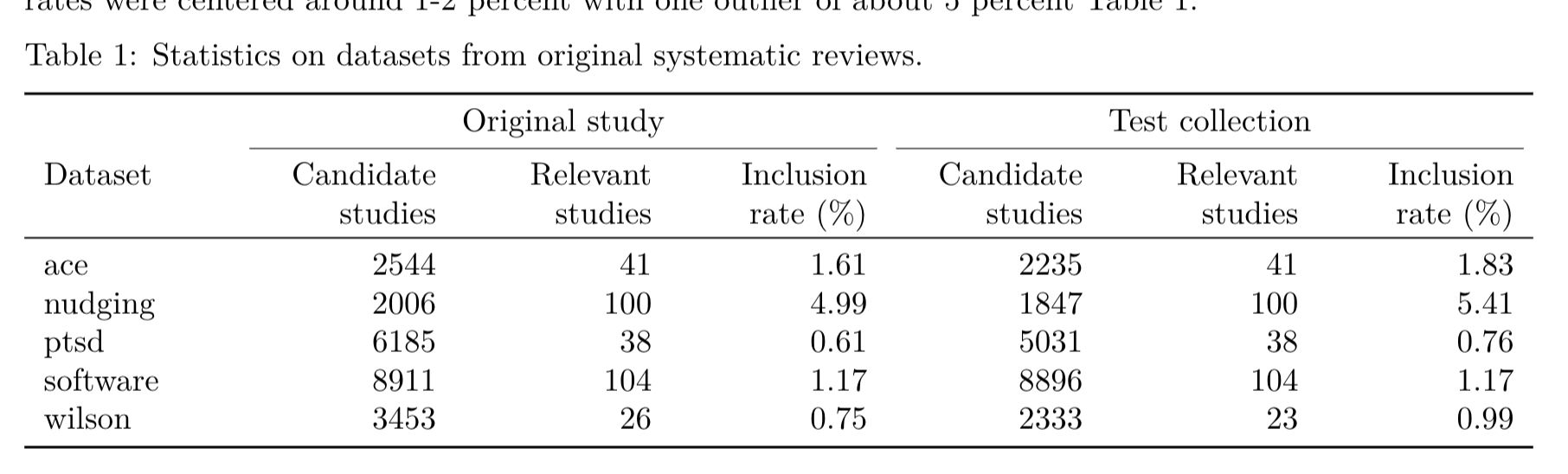
Every simulation started with an initial training set of one relevant and one irrelevant publication to represent a ‘worst case scenario’ where the reviewer has minimal prior knowledge on the publications in the data. To account for bias, the initial training set was randomly sampled from the dataset for every of the 15 runs. Although varying over runs, the initial training sets were kept constant over datasets to allow for a direct comparison of models within datasets. A seed value was set to ensure reproducability. The classifier was retrained every time after a publication had been labeled. The simulation ended after all publications in the dataset had been labeled.

## Datasets

The models were simulated on a convenience sample of six systematic review datasets. The data selection process was driven by two factors. Firstly, datasets were selected based on their background, given the need for datasets from diverse research areas. Secondly, datasets were selected by their availability, given the limited timespan of the current project. The datasets were retrieved from a collection of open systematic review datasets to be used for text mining purposes[[4]](#footnote-4).

Three out of six datasets originated from the *medical sciences*: Ace, Wilson, and Virus. The Wilson dataset (Appenzeller-Herzog 2020) is on a review on effectiveness and safety of treatments of Wilson Disease, a rare genetic disorder of copper metabolism (Appenzeller-Herzog et al. 2019). The Ace dataset contains publications on the efficacy of Angiotensin-converting enzyme (ACE) inhibitors, a drug treatment for heart disease (Cohen et al. 2006). The Virus dataset is from a systematic review on studies that performed viral Metagenomic Next-Generation Sequencing (mNGS) in farm animals (Kwok et al. 2020). From the field of *software engineering*, the Software dataset contains publications from a review on fault prediction in source code (Hall et al. 2012). The Nudging dataset (Nagtegaal et al. 2019a) belongs to a systematic review on nudging healthcare professionals (Nagtegaal et al. 2019b), stemming from the area of *behavioural public administration*, The PTSD dataset contains publications from the field of *psychology*. The corresponding systematic review is on studies applying latent trajectory analyses on posttraumatic stress after exposure to trauma (van de Schoot et al. 2017). Of these six datasets, Ace, and Software have been used for model simulations in previous studies on AI-aided title and abstract screening, respectively (Cohen et al. 2006) and (Yu, Kraft, and Menzies 2018a).

Data were preprocessed from their original source into a test dataset, containing title and abstract of the publications obtained in the initial search. Candidate studies with missing abstracts and duplicate instances were removed from the data. Test datasets were labelled to indicate which candidate studies were included in the systematic review, thereby indicating relevant publications. All test datasets consisted of thousands of candidate studies, of which only only a fraction was deemed relevant to the systematic review. Inclusion rates were centered around 1-2 percent with one outlier of about 5 percent Table 1.



## Evaluating performance

Model performance was visualized by plotting recall curves and further assessed by three different measures, Work Saved over Sampling (WSS), Relevant References Found (RRF), and Average Time to Discovery (ATD). Results were averaged over 15 trials for every simulation.

Plotting recall as a function of the number of screened publications offers insight in model performance throughout the screening process [Cormack and Grossman (2014); Yu2018a]. The curves give informations 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 (WSS), but on the other hand they present how many relevant publications are identified after screening a certain proportion of all publications (RRF).

WSS indicates the reduction in publications needed to be screened, at a given level of recall (Cohen et al. 2006). When measured at the typical recall level of 0.95 (Cohen et al. 2006), WSS yields an estimate of the amount of work that can have be saved at the cost of failing to identify 5% of relevant publications. In the current study, WSS is computed at 0.95 and 1.00 recall level. RRF statistics are computed at 10%, representing the proportion of relevant studies that were identified after screening 10% of all publications.

Both RRF and WSS are sensitive to random effects as these statistics are strongly dependent on the position of the cutoff value. Moreover, WSS makes assumptions about acceptable recall levels whereas this might depend on the research question at hand (O’Mara-Eves et al. 2015). A statistic that is not dependent on some arbitrary cutoff value is the ATD, which is the average number of publications that are screened to find a relevant publication, divided by the total number of publications in the data. The ATD is proportional to the area above the recall curve.

Analyses were carried out using R (R Core Team 2019), version 3.6.1. All datasets accompanying the systematic reviews are openly published. This study was approved by the Ethics Committee of the Faculty of Social and Behavioural Sciences of Utrecht University, filed as an amendement under study 20-104. All simulations were run using through Cartesius (EINF-156).

# References

Appenzeller-Herzog, Christian. 2020. “Data from Comparative Effectiveness of Common Therapies for Wilson Disease: A Systematic Review and Meta-Analysis of Controlled Studies.” Zenodo.

Appenzeller-Herzog, Christian, Tim Mathes, Marlies L. S. Heeres, Karl Heinz Weiss, Roderick H. J. Houwen, and Hannah Ewald. 2019. “Comparative Effectiveness of Common Therapies for Wilson Disease: A Systematic Review and Meta-Analysis of Controlled Studies.” *Liver International* 39 (11): 2136–52. <https://doi.org/10.1111/liv.14179>.

Borah, Rohit, Andrew W. Brown, Patrice L. Capers, and Kathryn A. Kaiser. 2017. “Analysis of the Time and Workers Needed to Conduct Systematic Reviews of Medical Interventions Using Data from the PROSPERO Registry.” *BMJ Open* 7 (2): e012545. <https://doi.org/10/f9tf57>.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32. <https://doi.org/10.1023/A:1010933404324>.

Chalmers, Iain. 2007. “The Lethal Consequences of Failing to Make Full Use of All Relevant Evidence About the Effects of Medical Treatments: The Importance of Systematic Reviews.” In *Treating Individualsfrom Randomised Trials to Personalised Medicine.*, 37–58. Lancet.

Cohen, Aaron M., Kyle Ambert, and Marian McDonagh. 2009. “Cross-Topic Learning for Work Prioritization in Systematic Review Creation and Update.” *Journal of the American Medical Informatics Association* 16 (5): 690–704. <https://doi.org/10/c3shq2>.

Cohen, A. M., W. R. Hersh, K. Peterson, and Po-Yin Yen. 2006. “Reducing Workload in Systematic Review Preparation Using Automated Citation Classification.” *Journal of the American Medical Informatics Association : JAMIA* 13 (2): 206–19. <https://doi.org/10.1197/jamia.M1929>.

Cormack, Gordon V., and Maura R. Grossman. 2014. “Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery.” In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 153–62. SIGIR ’14. Gold Coast, Queensland, Australia: Association for Computing Machinery. <https://doi.org/10.1145/2600428.2609601>.

Danka, Tivadar, and Peter Horvath. n.d. “modAL: A Modular Active Learning Framework for Python.”

Fu, Jui Hsi, and Sing Ling Lee. 2011. “Certainty-Enhanced Active Learning for Improving Imbalanced Data Classification.” In *2011 IEEE 11th International Conference on Data Mining Workshops*, 405–12. Vancouver, BC, Canada: IEEE. <https://doi.org/10.1109/ICDMW.2011.43>.

Gough, David, and Diana Elbourne. 2002. “Systematic Research Synthesis to Inform Policy, Practice and Democratic Debate.” *Social Policy and Society* 1 (3): 225–36. <https://doi.org/10/bdmp7h>.

Hall, Tracy, Sarah Beecham, David Bowes, David Gray, and Steve Counsell. 2012. “A Systematic Literature Review on Fault Prediction Performance in Software Engineering.” *IEEE Transactions on Software Engineering* 38 (6): 1276–1304. <https://doi.org/10.1109/TSE.2011.103>.

Harrison, Hannah, Simon J. Griffin, Isla Kuhn, and Juliet A. Usher-Smith. 2020. “Software Tools to Support Title and Abstract Screening for Systematic Reviews in Healthcare: An Evaluation.” *BMC Medical Research Methodology* 20 (1): 7. <https://doi.org/10.1186/s12874-020-0897-3>.

Kilicoglu, H., D. Demner-Fushman, T. C. Rindflesch, N. L. Wilczynski, and R. B. Haynes. 2009. “Towards Automatic Recognition of Scientifically Rigorous Clinical Research Evidence.” *Journal of the American Medical Informatics Association* 16 (1): 25–31. <https://doi.org/10/bjkhh9>.

Kwok, Kirsty T. T., David F. Nieuwenhuijse, My V. T. Phan, and Marion P. G. Koopmans. 2020. “Virus Metagenomics in Farm Animals: A Systematic Review.” *Viruses* 12 (1): 107. <https://doi.org/10.3390/v12010107>.

Lau, Joseph. 2019. “Editorial: Systematic Review Automation Thematic Series.” *Systematic Reviews* 8 (1): 70. <https://doi.org/10/ggsmwf>.

Le, Quoc V., and Tomas Mikolov. 2014. “Distributed Representations of Sentences and Documents.” *arXiv:1405.4053 [Cs]*, May. <http://arxiv.org/abs/1405.4053>.

Lewis, David D., and Jason Catlett. 1994. “Heterogeneous Uncertainty Sampling for Supervised Learning.” In *Machine Learning Proceedings 1994*, edited by William W. Cohen and Haym Hirsh, 148–56. San Francisco (CA): Morgan Kaufmann. <https://doi.org/10.1016/B978-1-55860-335-6.50026-X>.

Marshall, Iain J., Blair T. Johnson, Zigeng Wang, Sanguthevar Rajasekaran, and Byron C. Wallace. 2020. “Semi-Automated Evidence Synthesis in Health Psychology: Current Methods and Future Prospects.” *Health Psychology Review* 14 (1): 145–58. <https://doi.org/10/ggjv98>.

Miwa, Makoto, James Thomas, Alison O’Mara-Eves, and Sophia Ananiadou. 2014. “Reducing Systematic Review Workload Through Certainty-Based Screening.” *Journal of Biomedical Informatics* 51 (October): 242–53. <https://doi.org/10.1016/j.jbi.2014.06.005>.

Nagtegaal, Rosanna, Lars Tummers, Mirko Noordegraaf, and Victor Bekkers. 2019a. “Nudging Healthcare Professionals Towards Evidence-Based Medicine: A Systematic Scoping Review.” Harvard Dataverse.

———. 2019b. “Nudging Healthcare Professionals Towards Evidence-Based Medicine: A Systematic Scoping Review.” *Journal of Behavioral Public Administration* 2 (2). <https://doi.org/doi.org/10.30636/jbpa.22.71>.

O’Mara-Eves, Alison, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. 2015. “Using Text Mining for Study Identification in Systematic Reviews: A Systematic Review of Current Approaches.” *Systematic Reviews* 4 (1): 5. <https://doi.org/10.1186/2046-4053-4-5>.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12: 2825–30.

PRISMA-P Group, David Moher, Larissa Shamseer, Mike Clarke, Davina Ghersi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, and Lesley A Stewart. 2015. “Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 Statement.” *Systematic Reviews* 4 (1): 1. <https://doi.org/10.1186/2046-4053-4-1>.

Ramos, Juan, and others. 2003. “Using Tf-Idf to Determine Word Relevance in Document Queries.” In *Proceedings of the First Instructional Conference on Machine Learning*, 242:133–42. Piscataway, NJ.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reimers, Nils, and Iryna Gurevych. 2019. “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks.” *arXiv:1908.10084 [Cs]*, August. <http://arxiv.org/abs/1908.10084>.

Řehůřek, Radim, and Petr Sojka. 2010. “Software Framework for Topic Modelling with Large Corpora.” In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Settles, Burr. 2012. “Active Learning.” *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6 (1): 1–114. <https://doi.org/10.2200/S00429ED1V01Y201207AIM018>.

Shemilt, Ian, Antonia Simon, Gareth J. Hollands, Theresa M. Marteau, David Ogilvie, Alison O’Mara-Eves, Michael P. Kelly, and James Thomas. 2014. “Pinpointing Needles in Giant Haystacks: Use of Text Mining to Reduce Impractical Screening Workload in Extremely Large Scoping Reviews.” *Research Synthesis Methods* 5 (1): 31–49. <https://doi.org/10.1002/jrsm.1093>.

Tong, Simon, and Daphne Koller. 2001. “Support Vector Machine Active Learning with Applications to Text Classification.” *Journal of Machine Learning Research* 2 (Nov): 45–66.

Van de Schoot, Rens, Jonathan De Bruin, Raoul Schram, Parisa Zahedi, Bianca Kramer, Gerbrich Ferdinands, Albert Harkema, Qixiang Fang, and Daniel Oberski. 2020. “ASReview: Active Learning for Systematic Reviews,” April. <https://doi.org/10/ggssnj>.

van de Schoot, Rens, Marit Sijbrandij, Sonja D. Winter, Sarah Depaoli, and Jeroen K. Vermunt. 2017. “The GRoLTS-Checklist: Guidelines for Reporting on Latent Trajectory Studies.” *Structural Equation Modeling: A Multidisciplinary Journal* 24 (3): 451–67. <https://doi.org/10/gdpcw9>.

Wallace, Byron C., Thomas A. Trikalinos, Joseph Lau, Carla Brodley, and Christopher H. Schmid. 2010. “Semi-Automated Screening of Biomedical Citations for Systematic Reviews.” *BMC Bioinformatics* 11 (1): 55. <https://doi.org/10.1186/1471-2105-11-55>.

Yu, Zhe, Nicholas A. Kraft, and Tim Menzies. 2018a. “Finding Better Active Learners for Faster Literature Reviews.” *Empirical Software Engineering* 23 (6): 3161–86. <https://doi.org/10.1007/s10664-017-9587-0>.

Yu, Zhe, Nicholas Kraft, and Tim Menzies. 2018b. “Finding Better Active Learners for Faster Literature Reviews.” *Empirical Software Engineering*, March. <https://doi.org/10.1007/s10664-017-9587-0>.

Yu, Zhe, and Tim Menzies. 2019. “FAST2: An Intelligent Assistant for Finding Relevant Papers.” *Expert Systems with Applications* 120 (April): 57–71. <https://doi.org/10.1016/j.eswa.2018.11.021>.

Zhang, Harry. 2004. “The Optimality of Naive Bayes.” In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*. Vol. 2.

1. see the Appendix for more information [↑](#footnote-ref-1)
2. The combination Naive Bayes + Doc2Vec is not possible because nb takes positive input features [↑](#footnote-ref-2)
3. see the Appendix for more information [↑](#footnote-ref-3)
4. <https://github.com/asreview/systematic-review-datasets>, (Van de Schoot et al. 2020) [↑](#footnote-ref-4)