

Discussion

2022-02-27

Discussion

We propose a new integrated framework to help researchers collect and screen scientific publications characterised by high performance and versatility. This framework joins the growing field of systematic review automation (SRA) and helpers (SRH) tools (Ananiadou et al., 2009; A. M. Cohen et al., 2010, 2006; O’Mara-Eves et al., 2015). This framework implements standard approaches and uses ad-hoc solutions to common SRA issues. By freely sharing the tool as an open-source R package and by following a modular design, we sought to adopt some of the so-called Vienna Principles advocated by the International Collaboration for the Automation of Systematic Reviews (ICASR) (Beller et al., 2018).

The framework consists of four main components: 1) an integrated query-based citation search and management engine, 2) a Bayesian active machine learning-based citation classifier, and 3) a data-driven search query generation algorithm.

The search engine module used by the framework can automatically collect citation data from three well-known scientific databases (i.e., Pubmed, Web of Science, and the Institute of Electrical and Electronics Engineers) and process manually downloaded results from two more sources (SCOPUS, EMBASE). In comparison, most commercial or free SRH tools rely on internal databases (e.g., Mendeley <https://www.mendeley.com/>) sometimes focusing only on a particular topic (Visser, 2010) or a single external data source (Poulter et al., 2008; Soto et al., 2019; Thomas & Brunton, 2007).

Mixing different databases is essential to obtain a more comprehensive view of the literature (Bajpai et al., 2011; Wilkins et al., 2005; Woods & Trewheellar, 1998): in our results, 18.7% of the positive matches were found in only one of the different data sources, and no positive record was present in all the sources (data not shown).

The framework online search algorithms are efficient enough to manage tens of thousands of search results, using various solutions to overcome the limitations of citation databases in terms of traffic and download quotas. The results are then automatically organised, deduplicated and arranged by “simple query ordering” in a uniform corpus. The preliminary ordering increases the positivity rate in the initial training set (Wallace, Small, et al., 2010).

For the framework’s record screening module, we developed an active machine learning protocol (Miwa et al., 2014; Settles, 2009) based on the best practices from other SRA studies, bringing further improvements at various levels.

The feature extractor module uses modern NLP techniques (Ananiadou & McNaught, 2006; K. B. Cohen & Hunter, 2008) to transform free text into input data for machine learning. We did not include classical n-grams (Schonlau & Guenther, 2017); rather, we used network analysis to find non-consecutive, frequently associated terms, a generalisation of n-grams that relaxes the term adjacency assumption. This approach can also incorporate term connections across different parts of the records, e.g., terms having a different relevance when associated with a particular author. The same technique was used with different parameters to merge redundant terms, increasing estimation efficiency and reducing noise.

The use of concurrency network-driven text modelling is not new (Ohsawa et al., 1998; François Rousseau et al., 2015; François Rousseau, 2015; Violos et al., 2016) and is a valuable tool to extract semantic information that is not evident in one-word or consecutive n-gram models.

The automatic classification algorithm is based on Bayesian Additive Regression Trees (BART) (Chipman et al., 2010; Kapelner & Bleich, 2013). Like other boosted trees algorithms (Hastie et al., 2009), the BART method can explore complex non-linearities, perform variable selection, manage missing data while maintaining high predictive power.

However, the Bayesian foundation of the method provides further benefits: lower sensitivity to the choice of hyperparameters, natural regularisation through priors, and, most importantly, predictive distributions as output instead of point-wise predictions (Joo et al., 2020; Jospin et al., 2020; Soria-Olivas et al., 2011). By selecting relatively tight prior distributions, we discouraged overly deep trees, long tree sequences, and extreme predicted probabilities, thus reducing the risk of overfitting.

The algorithm runs multiple replications of the model and averages their predictive distributions creating an “ensemble”; this technique has been shown to improve out-of-sample predictive performance (Dietterich, 2000; Zhou, 2021), as confirmed during the hyperparameter evaluation (Supplemental Material S2). Ensembling reduces the uncertainty in the predictive distribution tails related to the randomness in the MCMC fit (Robert & Casella, 2004), generating a shift in the probability mass towards the distribution centre and stabilising it (i.e., reducing variance without impacting bias). On the other hand, simply imposing more robust uninformative priors against extreme predictions would have reduced variance but also shifted the distribution towards a non-decision zone, increasing bias (Hansen et al., 2000).

Since the number of model replications has a significant impact on computation times, we decided to use ten replicas, the lower value after which performance stabilised, as resulted from the evaluation of the hyperparameters (Supplemental Material S2, Fig. 2).

We also investigated whether bootstrapping between replications (Breiman, 1996) would improve performance; however, contrary to theory (Díez-Pastor et al., 2015), it appeared to be slightly detrimental in our case (Supplemental Material S2, Fig. 2) compared to simple ensembling.

A low proportion of relevant matches (class imbalance) is typical for literature reviews (O’Mara-Eves et al., 2015; Sampson et al., 2011; Wallace, Trikalinos, et al., 2010), and a strong imbalance between positive and negative records can affect sensitivity (Chawla et al., 2004; Khoshgoftaar et al., 2010).

To overcome this problem, we oversampled (Batista et al., 2004) the positive records ten times before model fitting. The hyperparameter analysis showed that the oversampling rate, together with model ensembling, was the parameter with the most significant impact on performance.

A known risk with positive oversampling is the misclassification of negative records (Ramezankhani et al., 2016). However, since all predicted positives in our approach are reviewed manually, we are always guaranteed to achieve 100% specificity/positive predictive value: the only price for the increased sensitivity due to oversampling is a larger number of records to be reviewed.

An alternative to oversampling would be to apply different weights and/or costs to the classes (Abd Elrahman & Abraham, 2013; Díez-Pastor et al., 2015), but the BART implementation we used did not have this feature; furthermore, using simple oversampling allows for a broader compatibility with different modelling engines (Galar et al., 2011; Roshan & Asadi, 2020).

Finally, sorting the records by query term frequency (simple query ordering) produces a much higher rate of relevant records in the initial training set (17.2%) compared to the overall data (0.11%), which boosts the sensitivity of the model.

One of the key innovations we have introduced is the concept of “uncertainty zone,” the implementation of which is possible thanks to the Bayesian foundation of the classification model.

This construct guides the selection of records to be manually reviewed and gets dynamically updated and reduced after each CR iteration, as more uncertain predictions are evaluated (Supplemental Material S2 Fig. 1).

The use of a dynamic uncertainty zone overcomes the usual requirement of dataset-specific hard thresholds in active machine learning and allows to review multiple items at once between iterations (Laws & Schütze, 2008; Miwa et al., 2014; Zhu et al., 2010). The hyperparameters required by our algorithm are general and non-task-specific, like the PPD intervals underlying the uncertainty zone and the maximum number of iterations without positive matches after which a session is concluded; the evaluation of the classification model hyperparameters shows that the algorithm is robust against variations in these parameters, and we

expect the default values to perform well on most datasets.

Since researchers are asked to review both records predicted as surely relevant and those inside the uncertainty zone, this method can be considered as a unifying synthesis of the “certainty” and “uncertainty” paradigms of active learning (Miwa et al., 2014).

We assessed performance as the ability of the screening procedure (automatic classification plus manual review) to find the largest number of relevant records while requiring manual reviewing for as few of them as possible (i.e., sensitivity \times efficiency).

We avoided the classical out-of-sample approaches such as train-test sampling, out-of-bag bootstrapping or cross-validation (James et al., 2013; Kohavi et al., 1995). Such methods primarily assume that the rate of positivity is the same on average in every possible random subset of the data (Tashman, 2000); this uniformity is broken by how the initial training set and the subsequent reviewed records are selected by the query-based ordering and active learning algorithm, resulting in a lower positivity rate in the unlabelled records (Fig. 2). Moreover, a literature corpus is unique per search query/database combination, and therefore any out-of-sample performance estimate is not replicable since no new data can be acquired related to the current corpus.

To estimate overall sensitivity, we instead applied simple Bayesian regression (surrogate model) to the manually reviewed data to abstract the classification model predictions and generate a maximum entropy (Harremoës & Topsøe, 2001) estimate of the number of missed positive matches among the unreviewed records in the whole dataset. This simple surrogate model fitted the data very well (R^2 consistently above 97%) using only the lower 98% PrI boundary of the PPDs as predictor, indicating predictive consistency in the classification model. The posterior predictive distribution of the surrogate model could be used to explore worse case scenarios in terms of sensitivity.

Our framework achieves very high sensitivity by screening only a very small fraction of all records, bringing a meaningful reduction in workload.

Based on the surrogate model, we predicted a predicted median sensitivity of 100% [93.5%, 100%] in the first session (screening 4.29% of records) and of 97.3% [73.8%, 100%] in the second (screening 1.34% of records): efficiency increased significantly in the second session as only a few new positive matches were found; however, given the large number of records, uncertainty about sensitivity increased, as expected.

Both results are above the usual performance in this field (O’Mara-Eves et al., 2015) and are in line with the average sensitivity of 92% estimated after human-only screening (Edwards et al., 2002). In one interesting case, the model detected a human-caused misclassification error, demonstrating its robustness and value as a second screener, a role already suggested for SRA tools in previous studies (Bekhuis & Demner-Fushman, 2010, 2012; Frunza et al., 2010). Although “simple query ordering” concentrated most relevant matches in the first 20-25 thousand records, without the tool support, the remaining relevant records would have been missed without manually screening almost the entire dataset.

The model required ~5-20 minutes per iteration to perform the predictions in session 1 (17,755 documents) and 20-40 minutes in session 2 (98,371 documents) on an eight-core, 2.5 GHz, 16 GB RAM, 2014 laptop; including manual record review, one session required 1-3 days of work, for a total of 1-2 weeks for the whole process (including record collection). This is a considerable time saving compared to the several months typically required for the screening phase of systematic reviews (Allen & Olkin, 1999; Bannach-Brown et al., 2019; Borah et al., 2017). To our knowledge, the amount of data processed (~100,000 records) was larger than what is typical of most SRA studies (O’Mara-Eves et al., 2015; Olorisade et al., 2016), highlighting the scalability of the tool in real-world scenarios.

The last module of our framework is an algorithm for data-driven search query generation. Generating an efficient and effective search query is a complex task (Hammerstrøm et al., 2010; Lefebvre et al., 2011); it requires building a combination of positive and negative terms to maximise the number of relevant search results while minimising the total number of records to be reviewed. Our solution combines a sensitivity-driven subquery proposal engine based on concurrent decision trees (Blanco-Justicia & Domingo-Ferrer, 2019;

Moore et al., 2018) built on the BART ensemble PPD, with a human review step and an efficiency-driven query builder. The aim is to generate a new search query to help find records missed in the first search session. The generated query did indeed retrieve a few more relevant records not found in session 1 but at the cost of significantly increasing the number of documents.

An interesting aspect of this feature is that it provides a human-readable overview of the classification rules learned by the classification model, showing which combination of terms was particularly relevant and even spotting authors and geographical locations associated with the study topic. The generated query, therefore, served also as a means for machine learning explainability (Bhatt et al., 2020; Burkart & Huber, 2021), useful for understanding and detecting biases in black-box classification algorithms (Malhi et al., 2020); explainability is often required or even legally mandatory for high-stake machine learning applications (Bibal et al., 2020, 2021).

It is important to note that this process is entirely data-driven. The algorithm is only aware of the “world” defined by the dataset used as input, which is generated by a specific search query focused on a particular topic. Therefore, the new query may not be specific enough when applied to an unbounded search domain and may return an unmanageable amount of irrelevant results. The solution we found was to add another component to the query, specifying the general topic (antimicrobial resistance and healthcare-associated infections) of our research.

As mentioned early, our framework builds on modularity. We have designed so that each module can become fully independent in future iterations; it will be possible for users to add custom features such as citation search and parsing for other scientific databases, alternative text processing algorithms or machine learning modules. We consider such interoperability to be extremely relevant: the main strength of our tool lies in the composition of many independent solutions, such as the idea of Bayesian active machine learning and the exploit of the derived uncertainty in defining the records needing human review.

Each component could benefit considerably from the recent improvements in text mining and machine learning.

For example, the text processing approach based on the “boolean bag-of-words” paradigm is quite simple and could be improved by more nuanced text representations. It might be considered whether feature transformations such as TF-IDF (Ananiadou & McNaught, 2006; Baeza-Yates et al., 1999) could be advantageous, although we hypothesise that tree-based classification algorithms like BART are robust enough not to require such operations. Instead, it might be worth exploring the application of word embedding: this technique transforms terms into semantic vectors derived from the surrounding text (Bollegala et al., 2015; Minaee et al., 2021; Turian et al., 2010) and could be used to reduce noise by merging different terms that are semantically similar or enhance signal by distinguishing identical terms with different meaning given the context. Another option would be to employ unsupervised learning models like Latent Dirichlet Analysis and Latent Semantic Analysis, (Q. Chen et al., 2016; Landauer et al., 1998; Pavlinek & Podgorelec, 2017) or graph-of-word techniques (Ohsawa et al., 1998; Francois Rousseau, 2015) to extract topics that expand the feature space.

Our classification algorithm is applicable with any Bayesian supervised machine learning method that provides full PPDs; therefore, alternative classification models, such as Gaussian Processes, known for their flexibility (S.-H. Chen et al., 2015; Jayashree & Srijith, 2020), could be evaluated. It would be even more interesting to test advanced learning algorithms that go beyond the bag-of-words approach and take into consideration higher-level features in the text such as term context and sequences, long-distance term relationships, semantic structures, etc., (Cheng et al., 2019; Farkas, 1995; Lai et al., 2015; Li et al., 2020; Minaee et al., 2021; Yang et al., 2020), provided that a Bayesian implementation of such algorithms is available (for example C. Chen et al. (2018)).

Finally, a natural improvement would be to provide a graphical user interface to make the framework easy to use also for less technical users.

The field of literature review automation is evolving rapidly, and we anticipate an increasing use of such technologies to address the accelerating pace of scientific production. We believe it is encouraging that a wide variety of tools are being made available to let researchers and policymakers find the approach that best fits their needs.

We contribute to this field with an innovative framework that provides excellent performance and easy

integration with existing systematic review pipelines. The value of this work lies not only in the framework itself, which we make available as open-source software, but also in the set of methodologies we developed to solve various SRA issues and which can also be used to improve existing solutions.

- Abd Elrahman, S. M., & Abraham, A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing*, 1(2013), 332–340.
- Allen, I. E., & Olkin, I. (1999). Estimating time to conduct a meta-analysis from number of citations retrieved. *Jama*, 282(7), 634–635.
- Ananiadou, S., & McNaught, J. (2006). *Text mining for biology and biomedicine*. Citeseer.
- Ananiadou, S., Rea, B., Okazaki, N., Procter, R., & Thomas, J. (2009). Supporting systematic reviews using text mining. *Social Science Computer Review*, 27(4), 509–523.
- Baeza-Yates, R., Ribeiro-Neto, B. others. (1999). *Modern information retrieval* (Vol. 463). ACM press New York.
- Bajpai, A., Davuluri, S., Haridas, H., Kasliwal, G., Deepti, H., Sreelakshmi, K., Chandrashekar, D., Bora, P., Farouk, M., Chitturi, N. others. (2011). In search of the right literature search engine (s). *Nature Precedings*, 1–1.
- Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S., Ananiadou, S., Liao, J., & Macleod, M. R. (2019). Machine learning algorithms for systematic review: Reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews*, 8(1), 1–12.
- Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–29.
- Bekhuis, T., & Demner-Fushman, D. (2010). Towards automating the initial screening phase of a systematic review. *MEDINFO 2010*, 146–150.
- Bekhuis, T., & Demner-Fushman, D. (2012). Screening nonrandomized studies for medical systematic reviews: A comparative study of classifiers. *Artificial Intelligence in Medicine*, 55(3), 197–207.
- Beller, E., Clark, J., Tsafnat, G., Adams, C., Diehl, H., Lund, H., Ouzzani, M., Thayer, K., Thomas, J., Turner, T. others. (2018). Making progress with the automation of systematic reviews: Principles of the international collaboration for the automation of systematic reviews (ICASR). *Systematic Reviews*, 7(1), 1–7.
- Bhatt, U., Andrus, M., Weller, A., & Xiang, A. (2020). Machine learning explainability for external stakeholders. *arXiv Preprint arXiv:2007.05408*.
- Bibal, A., Lognoul, M., De Streel, A., & Frénay, B. (2021). Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, 29(2), 149–169.
- Bibal, A., Lognoul, M., Streel, A. de, & Frénay, B. (2020). Impact of legal requirements on explainability in machine learning. *arXiv Preprint arXiv:2007.05479*.
- Blanco-Justicia, A., & Domingo-Ferrer, J. (2019). Machine learning explainability through comprehensible decision trees. *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 15–26.
- Bollegala, D., Maehara, T., & Kawarabayashi, K. (2015). Embedding semantic relations into word representations. *Twenty-Fourth International Joint Conference on Artificial Intelligence*.
- Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (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.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.
- Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70, 245–317.
- Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1), 1–6.
- Chen, C., Lin, X., & Terejanu, G. (2018). An approximate bayesian long short-term memory algorithm for outlier detection. *2018 24th International Conference on Pattern Recognition (ICPR)*, 201–206.
- Chen, Q., Yao, L., & Yang, J. (2016). Short text classification based on LDA topic model. *2016 International Conference on Audio, Language and Image Processing (ICALIP)*, 749–753.
- Chen, S.-H., Lee, Y.-S., Tai, T.-C., & Wang, J.-C. (2015). Gaussian process based text categorization for healthy information. *2015 International Conference on Orange Technologies (ICOT)*, 30–33. <https://doi.org/10.1145/2788888.2788918>.

//doi.org/10.1109/ICOT.2015.7498487

- Cheng, Y., Ye, Z., Wang, M., & Zhang, Q. (2019). Document classification based on convolutional neural network and hierarchical attention network. *Neural Network World*, 29(2), 83–98.
- Chipman, H. A., George, E. I., McCulloch, R. E. others. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266–298.
- Cohen, A. M., Adams, C. E., Davis, J. M., Yu, C., Yu, P. S., Meng, W., Duggan, L., McDonagh, M., & Smalheiser, N. R. (2010). Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools. *Proceedings of the 1st ACM International Health Informatics Symposium*, 376–380.
- Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P.-Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2), 206–219.
- Cohen, K. B., & Hunter, L. (2008). Getting started in text mining. *PLoS Computational Biology*, 4(1), e20.
- Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, 1–15.
- Díez-Pastor, J. F., Rodríguez, J. J., García-Osorio, C. I., & Kuncheva, L. I. (2015). Diversity techniques improve the performance of the best imbalance learning ensembles. *Information Sciences*, 325, 98–117.
- Edwards, P., Clarke, M., DiGuseppi, C., Pratap, S., Roberts, I., & Wentz, R. (2002). Identification of randomized controlled trials in systematic reviews: Accuracy and reliability of screening records. *Statistics in Medicine*, 21(11), 1635–1640.
- Farkas, J. (1995). Document classification and recurrent neural networks. *Proceedings of the 1995 Conference of the Centre for Advanced Studies on Collaborative Research*, 21.
- Frunza, O., Inkpen, D., & Matwin, S. (2010). Building systematic reviews using automatic text classification techniques. *Coling 2010: Posters*, 303–311.
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4), 463–484.
- Hammerstrøm, K., Wade, A., Jørgensen, A.-M. K., & Hammerstrøm, K. (2010). Searching for studies. *Education*, 54(11.3).
- Hansen, L. K. others. (2000). Bayesian averaging is well-tempered. *Proceedings of NIPS*, 99, 265–271.
- Harremoës, P., & Topsøe, F. (2001). Maximum entropy fundamentals. *Entropy*, 3(3), 191–226.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). Boosting and additive trees. In *The elements of statistical learning* (pp. 337–387). Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.
- Jayashree, P., & Srijith, P. (2020). Evaluation of deep gaussian processes for text classification. *Proceedings of the 12th Language Resources and Evaluation Conference*, 1485–1491.
- Joo, T., Chung, U., & Seo, M.-G. (2020). Being bayesian about categorical probability. *International Conference on Machine Learning*, 4950–4961.
- Jospin, L. V., Buntine, W., Boussaid, F., Laga, H., & Bennamoun, M. (2020). Hands-on bayesian neural networks—a tutorial for deep learning users. *arXiv Preprint arXiv:2007.06823*.
- Kapelner, A., & Bleich, J. (2013). bartMachine: Machine learning with bayesian additive regression trees. *arXiv Preprint arXiv:1312.2171*.
- Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). Comparing boosting and bagging techniques with noisy and imbalanced data. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(3), 552–568.
- Kohavi, R. others. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, 14, 1137–1145.
- Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3), 259–284.
- Laws, F., & Schütze, H. (2008). Stopping criteria for active learning of named entity recognition. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 465–472.

- Lefebvre, C., Manheimer, E., Glanville, J., Higgins, J., & Green, S. (2011). Searching for studies (chapter 6). *Cochrane Handbook for Systematic Reviews of Interventions Version*, 510.
- Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2020). A survey on text classification: From shallow to deep learning. *arXiv Preprint arXiv:2008.00364*.
- Malhi, A., Knapic, S., & Främling, K. (2020). Explainable agents for less bias in human-agent decision making. *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, 129–146.
- Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3), 1–40.
- Miwa, M., Thomas, J., O’Mara-Eves, A., & Ananiadou, S. (2014). Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics*, 51, 242–253.
- Moore, A., Murdock, V., Cai, Y., & Jones, K. (2018). Transparent tree ensembles. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1241–1244.
- O’Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, 4(1), 1–22.
- Ohsawa, Y., Benson, N. E., & Yachida, M. (1998). KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. *Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries-ADL’98*, 12–18.
- Olorisade, B. K., Quincey, E. de, Brereton, P., & Andras, P. (2016). A critical analysis of studies that address the use of text mining for citation screening in systematic reviews. *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering*, 1–11.
- Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and LDA topic models. *Expert Systems with Applications*, 80, 83–93.
- Poulter, G. L., Rubin, D. L., Altman, R. B., & Seoighe, C. (2008). MScanner: A classifier for retrieving medline citations. *BMC Bioinformatics*, 9(1), 1–12.
- Ramezankhani, A., Pournik, O., Shahrabi, J., Azizi, F., Hadaegh, F., & Khalili, D. (2016). The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes. *Medical Decision Making*, 36(1), 137–144.
- Robert, C. P., & Casella, G. (2004). *Monte carlo statistical methods* (Vol. 2). Springer.
- Roshan, S. E., & Asadi, S. (2020). Improvement of bagging performance for classification of imbalanced datasets using evolutionary multi-objective optimization. *Engineering Applications of Artificial Intelligence*, 87, 103319.
- Rousseau, Francois. (2015). *Graph-of-words: Mining and retrieving text with networks of features* [PhD thesis]. Ph. D. dissertation.
- Rousseau, François, Kiagias, E., & Vazirgiannis, M. (2015). Text categorization as a graph classification problem. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1702–1712.
- Sampson, M., Tetzlaff, J., & Urquhart, C. (2011). Precision of healthcare systematic review searches in a cross-sectional sample. *Research Synthesis Methods*, 2(2), 119–125.
- Schonlau, M., & Guenther, N. (2017). Text mining using n-grams. *Schonlau, M., Guenther, N. Sucholutsky, I. Text Mining Using n-Gram Variables. The Stata Journal*, 17(4), 866–881.
- Settles, B. (2009). *Active learning literature survey*.
- Soria-Olivas, E., Gomez-Sanchis, J., Martin, J. D., Vila-Frances, J., Martinez, M., Magdalena, J. R., & Serrano, A. J. (2011). BELM: Bayesian extreme learning machine. *IEEE Transactions on Neural Networks*, 22(3), 505–509.
- Soto, A. J., Przybyła, P., & Ananiadou, S. (2019). Thalia: Semantic search engine for biomedical abstracts. *Bioinformatics*, 35(10), 1799–1801.
- Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4), 437–450.
- Thomas, J., & Brunton, J. (2007). *EPPI-reviewer: Software for research synthesis*.
- Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 384–394.

- Violos, J., Tserpes, K., Psomakelis, E., Psychas, K., & Varvarigou, T. (2016). Sentiment analysis using word-graphs. *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*, 1–9.
- Visser, E. (2010). Performing systematic literature reviews with researchr: Tool demonstration. *Technical Report Series TUD-SERG-2010-010*.
- Wallace, B. C., Small, K., Brodley, C. E., & Trikalinos, T. A. (2010). Active learning for biomedical citation screening. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 173–182.
- Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., & Schmid, C. H. (2010). Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics*, 11(1), 1–11.
- Wilkins, T., Gillies, R. A., & Davies, K. (2005). EMBASE versus MEDLINE for family medicine searches: Can MEDLINE searches find the forest or a tree? *Canadian Family Physician*, 51(6), 848–849.
- Woods, D., & Trewheellar, K. (1998). Medline and embase complement each other in literature searches. *BMJ: British Medical Journal*, 316(7138), 1166.
- Yang, J., Bai, L., & Guo, Y. (2020). A survey of text classification models. *Proceedings of the 2020 2nd International Conference on Robotics, Intelligent Control and Artificial Intelligence*, 327–334.
- Zhou, Z.-H. (2021). Ensemble learning. In *Machine learning* (pp. 181–210). Springer.
- Zhu, J., Wang, H., Hovy, E., & Ma, M. (2010). Confidence-based stopping criteria for active learning for data annotation. *ACM Transactions on Speech and Language Processing (TSLP)*, 6(3), 1–24.