The effect of dataset size on neural network performance within systematic reviewing

Govert Verberg

July 19, 2021

Keywords— Neural Networks, Active Learning, Systematic Reviewing, ASReview

Abstract

The study contributes to the theory behind applying neural networks to active learning. It is generally assumed that having more data available will lead to increased performance when using machine learning. This assumption was tested in a specific problem setting: using neural networks in combination with active learning to aid in systematic reviewing. A simulation study was performed using the neural network classifier in ASReview. The ASReview simulation mode was applied to different sized samples out of three different datasets, to measure the change in performance. The results from this study show that for active learning, increasing the dataset sample size does not always lead to increased performance.

1 Introduction

Systematic reviews are important contributions to science. Synthesizing the outcomes of multiple studies on the same topic leads to more reliable and objective results while also making research more accessible to both scholars and the general public (Boaz, Ashby, Young, et al., 2002). Systematic reviews are defined by being both transparent and reproducible (Liberati et al., 2009). Researchers should strive to identify as much of the relevant work as possible when performing a systematic review (Boaz et al., 2002). This is traditionally done by first doing a keyword search to identify candidate records that might be relevant, and then manually screening the candidate papers to exclude the irrelevant records from the review. This manual screening process is time consuming and has also been proven to be somewhat unreliable; Wang et al. report an error rate of 10.76% (Wang, Nayfeh, Tetzlaff, O'Blenis, & Murad, 2020).

Machine learning can be used to aid researchers during the screening process (van de Schoot et al., 2021). The downside of machine learning is that training an accurate classifier model generally requires a lot labeled data, which is not available at the start of the screening process. Active learning approaches circumvent this problem by letting the model select which data point should be labeled next. This reduces the amount of labeled data required (Kremer, Steenstrup Pedersen, & Igel, 2014). In a Human-in-the-loop approach, a human user is involved in training the model by labeling the data that is selected by the model (Holzinger, 2016). Generally in active learning, the goal is to train a model for future use. Active learning is then used to limit the amount of labeled data that is required to train the model, which saves labeling costs and computation time (Kremer et al., 2014). In systematic reviewing, the goal is different. The goal is to find the relevant records from the dataset, not to train a model for future use. In literature, this application of active learning is usually not considered, e.g., review studies conducted by Schröder et al. and Ren et al. (Schröder & Niekler, 2020)(Ren et al., 2020).

Van de Schoot et al. have demonstrated with a simulation study that traditional classification models (e.g., Naive Bayes) used in combination with active learning can achieve good results in systematic reviewing, reducing the amount of records that need to be screened by as much as 90% (van de Schoot et al., 2021). However, the state of the art models for document classification are all based on deep learning (Minaee et al., 2021). In theory, applying these state of the art models to systematic reviewing should reach even better results than the traditional models. The challenge here is that deep learning models are trained on datasets that are much larger than typical systematic review dataset sizes. For example, Giga5 - a commonly used dataset for training deep learning models - contains almost 10 million documents, while systematic reviews usually have only a few thousand records (Parker, Graff, Kong, Chen, & Maeda, 2011).

Shallow neural networks (i.e., networks with less than five layers) can be considered an intermediate step between traditional models and deep learning models in both model complexity and data required for training. In theory, the increased model complexity can lead to improved performance, but more understanding is needed of how the limited dataset sizes in active learning affect neural networks. A commonly used rule of thumb for dataset sample size requirements for neural networks is that the sample size should be at least ten times the number of trainable parameters in the model. Alwosheel et al. demonstrate that this rule of thumb is not conservative enough, they recommend using at least fifty times the number of trainable parameters instead (Alwosheel, van Cranenburgh, & Chorus, 2018). This suggests only very simple networks can be used for active learning in systematic reviewing. However, since their experiments do not consider active learning, their results might not apply here.

Neural network models have the potential to perform even better than traditional models in systematic reviewing. This increased performance would make it easier and less time consuming for researchers to do a systematic review study. Previous research has shown that the optimal classifier model in systematic reviewing differs per dataset (Ferdinands et al., 2020), it would therefore be helpful to know on what kind of datasets neural networks can perform well.

This study aims to contribute to the development of neural networks for systematic reviewing by focusing on dataset size: how does the performance in systematic reviewing of a neural network change when increasing the dataset size? It is expected that the performance of a neural network in systematic reviewing will improve with larger dataset sizes because text classification models in general achieve better results with larger datasets (Wei & Zou, 2019). There is a limit to how much increasing dataset size can improve the model performance, though the work of Alwosheel et al. suggests that this limit is around a dataset size of fifty times the number of trainable parameters in the model (Alwosheel et al., 2018), which means this limit is unlikely to be reached in systematic reviewing.

2 Methodology

2.1 Datasets

Three different datasets were chosen to take samples from, these datasets were described in more detail in Table 1. The first dataset was collected by Nagtegaal et al. during a systematic review study on nudging healthcare professionals to evidence-based medicine (Nagtegaal, Tummers, Noordegraaf, & Bekkers, 2019). The second dataset was collected by Hall et al. (Hall, Beecham, Bowes, Gray, & Counsell, 2011), they performed a systematic review study on fault prediction in software engineering (Hall et al., 2011). The third dataset was collected by Brouwer et al. during a systematic review study on depressive relapse and recurrence (Brouwer et al., 2019). All three authors fully describe the search strategy they used to construct these datasets in their papers (Nagtegaal et al., 2019)(Hall et al., 2011)(Brouwer et al., 2019). All datasets have very low inclusion rates (5.0%, 1.2%, 0.1%) and thus suffer from an imbalanced data problem, which is common in systematic review datasets(Borah, Brown, Capers, & Kaiser, 2017).

Dataset	**Topic**	**Number of records**	**Included records**
Nagtegaal 2019	Nudging	2008	101
Hall 2012	Software fault prediction	8812	104
Brouwer 2019	Depression	48977	62

Table 1: Description of the datasets used in the study. The listed numbers of records are after deduplication.

2.2 Simulation study setup

A simulation study was performed using ASReview, version 0.17 (van de Schoot et al., 2021). ASReview includes a simulation mode that can be used to simulate in what order a particular machine learning model would have suggested the records from a fully labeled dataset. A neural network implemented within ASReview was applied to multiple samples from the datasets described in the previous section. For every combination of dataset and sample size, a single simulation was performed. The scripts needed to reproduce the results in this simulation study are available online (Verberg, 2021).

The machine learning pipeline in ASReview consists of the following steps: balance strategy, feature extraction, classification model, query strategy. The balance strategy changes the weights of relevant and irrelevant records, which is useful for datasets with low inclusion rates. The feature extraction step is used to create a numerical vector representation of the title and abstract for each record. These vectors can then be used as input features for the classification model. The classification model is first trained on the labeled records and then used to predict relevance scores for the unlabeled records. The query strategy then determines which record is screened next, based on these relevance scores. The balance strategy used here was 'double balance', which was the default setting in ASReview (ASReview documentation, 2021b). For feature extraction, the Doc2Vec model included in ASReview was used, which was based on gensim (ASReview documentation, 2021a)(Řehůřek & Sojka, 2010). The classification model used was a dense neural network containing 2 hidden layers of 128 units each, and a sigmoid output unit. It was implemented using Keras (ASReview documentation, 2021d)(Chollet et al., 2015). The total number of trainable parameters in the network was 21889. The query strategy used was 'max certainty', which selects the record with the highest relevance score in the model for screening. 'Max certainty' was the default option in ASReview. (ASReview documentation, 2021c).

2.3 Evaluating performance

The performance of the neural network was evaluated using three different metrics that were also used in earlier simulation studies (Ferdinands et al., 2020). The first of these is Relevant References Found (RRF) after having screened the first 10% of the records (RRF@10). This metric measures how well the model performs in the early stages of screening. The other two metrics are Work Saved over Sampling (WSS) at two different levels of recall: 95% and 100% (WSS@95 and WSS@100). These metrics measure the reduction in the numbers of records that need to be screened as a percentage, compared to random screening. Earlier simulations have shown that finding all relevant records is significantly more difficult than finding 95% of the relevant records when using traditional machine learning models (van de Schoot et al., 2021). Considering also the error rate in manual screening (Wang et al., 2020), finding 95% of relevant records is generally sufficient and more practical than finding all relevant records. However, finding all relevant records would be the preferred result of the screening process, which is why WSS@100 is also included as a metric.

2.4 Prior knowledge and dataset samples

The review process in ASReview requires at least two records (one relevant and one irrelevant) to be chosen as prior knowledge before starting the screening process. These records are used to train the first model. In this simulation study, the same priors were used for each sample to decrease randomness in the results. Samples from each dataset were taken in the following way: Two (one relevant and one irrelevant) records were split off from the full dataset before sampling. Samples were created using stratified sampling, to keep the inclusion rate stable. Sampling was done in such a way that smaller samples are a subset of larger samples, meaning for example that all records included in the software 200 sample are also in the software 400 sample. After creating the samples, the two records that were split off at the start are added to each of the samples. These two records were used as prior knowledge for all runs. For the nudging dataset, samples of the following sizes were taken: 200, 400, 800, 1600. For the software dataset, samples of the following sizes were taken: 200, 400, 800, 1600, 3200, 6400. For the depression dataset, samples of the following sizes were taken: 1600, 3200, 6400, 12800, 25600, 48975.

3 Results

The performance of the neural network during each of the simulation runs can be measured by the metrics described in the methods section. The RRF@10 is shown in Figure 1. Work saved over sampling at 95% and 100% recall are shown in Figure 2 and Figure 3 respectively.

4 Discussion

4.1 Main findings

The goal of the simulation study was to find how dataset size influences neural network performance in systematic reviewing. Three different metrics were used to measure this performance. Because only a single simulation run was performed for every combination of dataset and sample size, there is no margin of error available for the measurements of the metrics.

The RRF@10 is shown in Figure 1. The results follow a different pattern for each dataset. For the nudging dataset the results are stable, though there seems to be some random noise. For the software dataset, the performance in this metric rapidly increases to the maximum score, which occurs when all relevant papers were found within the first 10% of the screening phase. After this point, performance is consistently high. For the depression dataset, the performance starts at the maximum value and then drops to the minimum value when the sample size is increased. A possible explanation for this is that the larger sample contained a lot of new records that were very difficult to find. After this, the performance increases along with the sample size.

The WSS@95 is shown in Figure 2. Again, the results follow a different pattern for each dataset. For the nudging dataset, performance in this metric increases along with the sample size. For the software dataset, performance is high for all but one sample, which could be caused by noise. For the depression dataset, the performance starts high and then decreases first before increasing again.

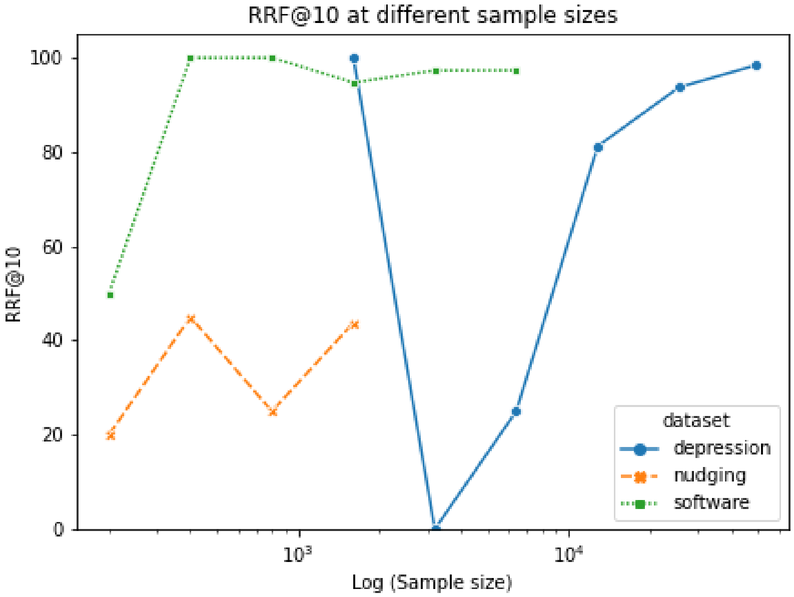

Figure 1: The number of relevant records found after having screened the first 10% of the records, for the different samples and datasets. The line styles and colors indicate the three different datasets from which the samples were taken. A higher score indicates better performance; a score of 100 indicates all relevant records were found within the first 10% of the screening phase.

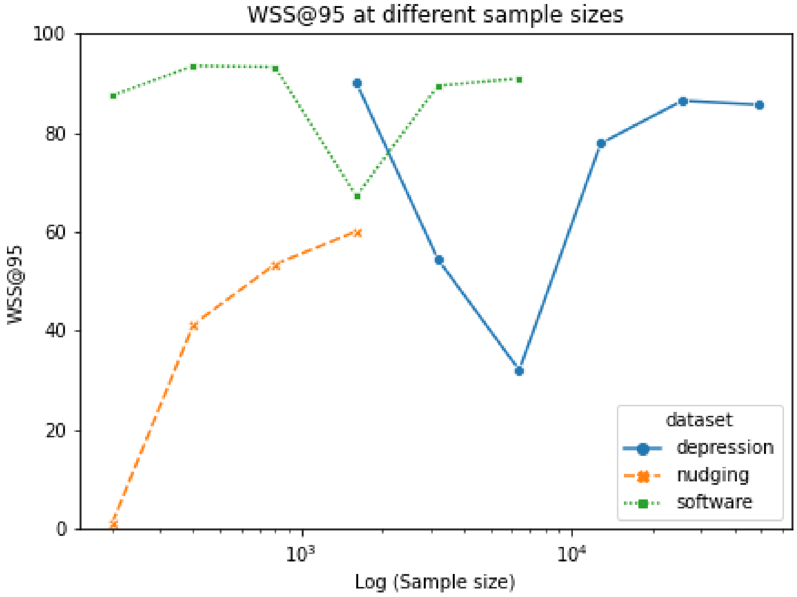

Figure 2: The amount of work that can be saved compared to random sampling, with a goal of finding 95% of the relevant records in the dataset. The line styles and colors indicate the three different datasets from which the samples were taken.

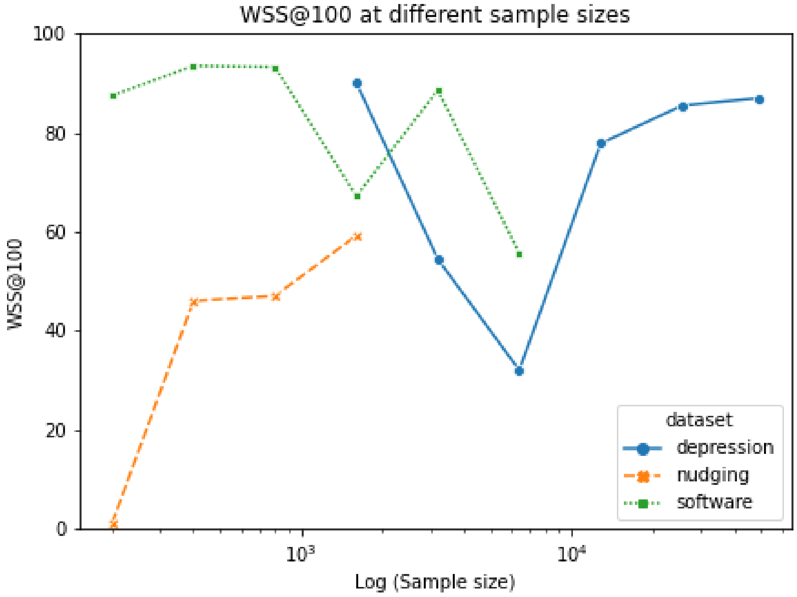

Figure 3: The amount of work that can be saved compared to random sampling, with a goal of finding 100% of the relevant records in the dataset. The line styles and colors indicate the three different datasets from which the samples were taken.

The WSS@100 is shown in Figure 3. results follow a different pattern for each dataset, but a very similar pattern as for the WSS@95. For the nudging dataset, performance increases along with the sample size. For the software dataset, performance is stable for the first few samples but then starts varying. For the depression dataset, the performance starts high and then decreases first before increasing again.

The results show that all three metrics used to evaluate the performance are somewhat correlated. Finding a high percentage of relevant records early in the run will lead to high scores for all of the metrics. The observed changes in performance when increasing the dataset sample size appear to be influenced by random noise. There is currently no likely theoretical explanation for why performance should first decrease and then increase again. This effect has also not been observed in other studies, like the work of Alwosheel et al. (Alwosheel et al., 2018). It is possible that some unknown effect exists in this specific problem setting which could lower the performance of the model when increasing the dataset size. In that case, the observed results could be explained as some kind of tipping point between two different effects that decrease and increase performance with dataset size.

A possible explanation for the differences between datasets is that the relation between the performance and the dataset sample size is dependent on other dataset characteristics, like the systematic review topic or the inclusion rate.

All three metrics used in this study are influenced by results from the start of the screening process. The first model in each simulation run is trained with only the two prior records, and then a single record is added with each step within the simulation process. So the datasets used to train the model are very small in the first phase of the simulation, regardless of the full sample size. It is likely that the initial performance of the model influences the performance of the model during the full simulation run, which could explain why there is not always an increase in performance at higher sample sizes.

4.2 Study limitations and recommendations for future research

The results obtained during this study are statistically weak, which could be improved upon by future research, by increasing the number of simulation runs at each sample size and adding more datasets to sample from. The number of runs at each sample size can be increased by taking different samples of the same size or by using the same samples but switching the priors. The simulation runs are computationally expensive for larger datasets, which should be considered when trying to expand on this simulation study.

Other methods that can be used to increase the performance of neural networks within active learning have not been explored in this study. There are many different neural network architectures that could be considered, while only a single one was used here. In addition, the current neural network implementation in ASReview is training a new model from scratch every time a new record is labeled. This seems unnecessary - the model architecture is the same every time, and the goal of the model is the same too. The only change between consecutive models is the input data. Concepts of transfer learning could be applied: the model weights of a previously trained model can be used to initialize the new model. The amount of epochs trained per model can then be reduced significantly. Reducing computing time this way opens up options that might otherwise be too computationally expensive (like using a deeper network).

4.3 Conclusion

Machine learning can be applied to systematic reviewing, to save researchers a lot of time during the screening process (van de Schoot et al., 2021). Optimizing the machine learning model leads to even more saved time, though the performance of machine learning models in systematic reviewing is dataset dependent (Ferdinands et al., 2020). Neural networks can be used as the classification models within systematic reviewing, but it is not known on what kind of datasets they perform well. In this study, the effect of dataset size on the performance of neural networks within systematic reviewing was simulated.

The results from this study show that increasing the dataset size is not guaranteed to lead to an increase in performance for neural networks within systematic reviewing. Alwosheel et al. show that the performance of neural networks in classification problems increases with dataset sample size (Alwosheel et al., 2018), which is also the accepted theory within the machine learning community. The results from this study are not in conflict with this theory, because systematic reviewing is a specific problem and the metrics used to evaluate model performance are different than the metrics that are generally used in classification problems.

Because there is no proven increase in performance for neural networks on larger datasets, researchers should consider using traditional models even for very large systematic review datasets. Conversely, neural networks might also be viable candidates for small datasets. How neural networks perform in systematic reviewing compared to traditional models has not been tested in this study. Other neural network architectures might exist which do benefit from having large systematic review datasets.

Dataset size is often assumed to improve machine learning model performance. This study shows that for neural networks in systematic reviewing, other paths should be considered when trying to improve model performance. This might also apply to other applications of neural networks in active learning.

References

Alwosheel, A., van Cranenburgh, S., & Chorus, C. G. (2018). Is your dataset big enough? sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of choice modelling*, *28*, 167–182.

ASReview documentation. (2021a). *Api reference doc2vec.* Retrieved from `https://asreview.readthedocs.io/en/latest/API/reference.html#asreview.models.feature_extraction.Doc2Vec`

ASReview documentation. (2021b). *Api reference double balance.* Retrieved from `https://asreview.readthedocs.io/en/latest/API/reference.html#asreview.models.balance.DoubleBalance`

ASReview documentation. (2021c). *Api reference maxquery.* Retrieved from `https://asreview.readthedocs.io/en/latest/API/reference.html#asreview.models.query.MaxQuery`

ASReview documentation. (2021d). *Api reference nn2layer.* Retrieved from `https://asreview.readthedocs.io/en/latest/API/reference.html#asreview.models.classifiers.NN2LayerClassifier`

Boaz, A., Ashby, D., Young, K., et al. (2002). *Systematic reviews: what have they got to offer evidence based policy and practice?* ESRC UK Centre for evidence based policy and practice London.

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.

Brouwer, M. E., Williams, A. D., Kennis, M., Fu, Z., Klein, N. S., Cuijpers, P., & Bockting, C. L. (2019). Psychological theories of depressive relapse and recurrence: a systematic review and meta-analysis of prospective studies. *Clinical psychology review*, *74*, 101773.

Chollet, F., et al. (2015). *Keras.* `https://keras.io`.

Ferdinands, G., Schram, R. D., de Bruin, J., Bagheri, A., Oberski, D. L., Tummers, L., & van de Schoot, R. (2020). Active learning for screening prioritization in systematic reviews-a simulation study.

Hall, T., Beecham, S., Bowes, D., Gray, D., & Counsell, S. (2011). A systematic literature review on fault prediction performance in software engineering. *IEEE Transactions on Software Engineering*, *38*(6), 1276–1304.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, *3*(2), 119–131.

Kremer, J., Steenstrup Pedersen, K., & Igel, C. (2014). Active learning with support vector machines. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(4), 313–326.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., ... Moher, D. (2009). The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology*, *62*(10), e1–e34.

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.

Nagtegaal, R., Tummers, L., Noordegraaf, M., & Bekkers, V. (2019). Nudging healthcare professionals towards evidence-based medicine: A systematic scoping review. *Journal of Behavioral Public Administration*, *2*(2).

Parker, R., Graff, D., Kong, J., Chen, K., & Maeda, K. (2011). English gigaword fifth edition, 2011. *Linguistic Data Consortium, Philadelphia, PA, USA*.

Řehůřek, R., & Sojka, P. (2010, May 22). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA. (`http://is.muni.cz/publication/884893/en`)

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Chen, X., & Wang, X. (2020). A survey of deep active learning. *arXiv preprint arXiv:2009.00236*.

Schröder, C., & Niekler, A. (2020). A survey of active learning for text classification using deep neural networks. *arXiv preprint arXiv:2008.07267*.

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., ... others (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*(2), 125–133.

Verberg, G. (2021, July). *Scripts for a simulation on the effect of dataset size on neural network performance for active learning applied to systematic reviewing.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.5112498` doi: 10.5281/zenodo.5112498

Wang, Z., Nayfeh, T., Tetzlaff, J., O'Blenis, P., & Murad, M. H. (2020). Error rates of human reviewers during abstract screening in systematic reviews. *PloS one*, *15*(1), e0227742.

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.