AI Assisted (Active Learning) Systematic Reviews

Kongkea Ouch

Interim Report for the Master of Science in

Computer Vision, Robotics and Machine Learning

from the

University of Surrey

Department of Electronic Engineering

Faculty of Engineering and Physical Sciences

University of Surrey

Guildford, Surrey, GU2 7XH, UK

May 2021

Supervised by: Professor Mark Plumbley and Andrew Bailey

DECLARATION OF ORIGINALITY

I confirm that the project dissertation I am submitting is entirely my own work and that any material used from other sources has been clearly identified and properly acknowledged and referenced. In submitting this final version of my report to the JISC anti-plagiarism software resource, I confirm that my work does not contravene the university regulations on plagiarism as described in the Student Handbook. In so doing I also acknowledge that I may be held to account for any particular instances of uncited work detected by the JISC anti-plagiarism software, or as may be found by the project examiner or project organiser. I also understand that if an allegation of plagiarism is upheld via an Academic Misconduct Hearing, then I may forfeit any credit for this module or a more severe penalty may be agreed.

AI Assisted (Active Learning) Systematic Review

Kongkea Ouch

Author Signature: Kongkea Ouch Date: 11/05/2021

Supervisor's name: Professor Mark Plumbley

WORD COUNT

Number of Pages: 30

Number of Words: 6511

ABSTRACT

Systematic reviews (SRs) are methodical and formal research literature reviews that are widely performed in several disciplines such as clinical and social science. However, SRs are gruelling and time consuming due to the requirement of manual supervision from researchers. Hence, state-of-the-art artificial intelligence tools based on active learning have been proposed to help researchers save time and effort in performing the screening phase of SRs. Great reduction in screening workload has been reported from using ASReview [20], an active learning screening system, to screen a diverse number of datasets in a reproduced simulation experiment. Future works will revolve around the task of building upon ASReview using latest artificial intelligence techniques that can potentially improve the active learning screening tool.

TABLE OF CONTENTS

LIST OF FIGURES

1 INTRODUCTION

A systematic review (SR) is generally defined as ‘a review of the evidence on a clearly formulated question that uses systematic and explicit methods to identify, select and critically appraise relevant primary research, and to extract and analyse data from the studies that are included in the review’ [1] and the associated methods with systematic review must be repeatable and transparent [2]. The standard process of a systematic review commonly starts with formulating a meaningful research question and generating its respective search query which is then used to retrieve preliminarily related papers across multiple databases. These papers will further be screened by experts in order to be determined as relevant and suitable for including in a summary of findings to the research question. Such an approach requires researchers to dedicate a massive amount of time and effort, especially the step of screening a large number of studies to determine which papers are relevant or not [2]. Hence, it calls for a new approach that incorporate latest efficient technology to assists researchers with expediting the screening stage.

1.1 Background and Context

Over the last 15 years, in order to offset the immense workload in the process of systematic review, several techniques combining machine learning and natural language processing have been developed and have shown positive results in screening documents [2-9]. Among these techniques, active learning algorithm, which makes use of machine learning to classify documents based on regular feedback from reviewers, has been well documented to reduce the labour-intensive effort of having to screen every single title and abstract across gigantic number of papers and make the systematic review process more efficient and effective [8, 9]. More reproductions using active learning on different SR datasets are required to evaluate its robustness and shed new light on its performance.

1.2 Objectives

The aim of this project is to incorporate an optimal active learning algorithm and measure its impact on different SRs particularly multiple existing SRs and a novel SR on ‘to what extent has Artificial Intelligence (AI) been used to diagnose depression?’ conducted by Bailey et al. [42] with the following objectives:

- Research and compare published results on use of active learning in screening SRs
- Implement a baseline AI for screening SRs based on a state-of-the-art paper
- Reproduce a simulation study using the baseline on multiple SRs
- Discuss the simulation results of the baseline to gain a better understanding of the usefulness of the active learning approach using different metrics used in the baseline's paper
- Compare and report on the potential of improving the baseline AI and the extent to which the AI Assisted Systematic Review can be fully automated without requiring human intervention

1.3 Achievements

- Understanding basics of natural language processing (NLP), AI and Pytorch framework
- Implement a baseline AI for systematic review based on a state-of-the-art paper
- Reproduce a simulation study using the baseline on multiple SRs
- Discuss the simulation results of the baseline to gain a better understanding of the usefulness of the active learning approach using different metrics used in the baseline's paper
- Compare and report on the potential of improving the baseline AI

1.4 Overview of Interim Report

Background study and literature theory will be discussed in chapter 2. In this chapter, a brief summary of systematic review and potential application of AI into this field is discussed, followed by an overview of latest tools on this subject and how they work. Important terms will be explained along the way in a high-level manner, so that general audience can grasp the foundation knowledge and relate it to the aforementioned tools. Strengths and weaknesses of these tools will be summarised and one tool will be selected for a simulation study in chapter 3 which is the technical chapter. The choice of technology of the tool and datasets will be described and simulation experiment will be conducted based on this preference. Simulation results will be reported and discussed which highlights the gaps in the current simulation and gives ways to improvement in chapter 4 of future plan. For conclusion of chapter 5, a general view on subject's background, simulation and future plan will be given.

2 BACKGROUND THEORY AND LITERATURE REVIEW

2.1 An Overview on Systematic Reviews

Systematic reviews (SR) are commonly used to gather and summarize research documents from a field of literature in order to deliver critical information necessary to the development of policy and practice on a particular research problem [1, 2]. SRs have been dramatically increasing since the 1970s, and they cost more time and money to perform nowadays than they did fifty years ago, thanks to the exponential rise in publications across multi-disciplinary field of sciences [10]. This is due to the time-consuming requirement of manually screening a large chunk of titles and abstracts to find publications that are applicable to the research question [11]. Because of the significant resources required to develop a comprehensive systematic review, lately there has been a lot of interest in incorporating Artificial Intelligence (AI) to expedite the screening process [12, 15].

2.2 Artificial Intelligence in Systematic Reviews

According to Barr and Feigenbaum [13], AI is a field of computer science that allows us to build intelligent machines that act, think, and decide like humans. AI can solve complex problems in real-world applications such as healthcare, education, agriculture, and so on [13]. One of the applications is assisting researchers in carrying out the screening phase for systematic review using two AI subfields, Machine Learning (ML) and Natural Language Processing (NLP) [2].

NLP [14, 24] is concerned with how computers interpret and understand human language in order to make sense of written or verbal data and perform tasks such as translation, subject modelling, emotion analysis, and other similar tasks.

The aim of machine learning is to learn from data and improve predictive performance over time without having to be programmed [13]. There are three main types of machine learning methods [13]:

- Supervised learning: uses labelled dataset to train the ML model.
- Unsupervised learning: uses unlabelled dataset to train the ML model.
- Semi-supervised learning: incorporates labelled and unlabelled datasets in training the ML model.

Active learning is a semi-supervised machine learning in which the algorithm can iteratively select which data it wants to learn from, reducing the number of papers that need to be manually screened [27, 20, 2]. With this active learning technique, an ML model will ask reviewers to label documents as "relevant" or "irrelevant," and then use those labels to improve its ranking predictions of relevant documents [20, 2]. After spending some time labelling, the reviewers may stop screening to save time once they decide the model has found sufficient relevant studies, which indicates that the ML model has 'semi-automated' the screening process for them [20, 2].

2.3 Feature Extraction

An ML model cannot right away learn from raw textual content of a document. A document's content including title, abstract and keywords must be numerically converted to feature vectors. The process of representing the textual content as feature vectors is called 'feature extraction' which has these key methods:

- The TF-IDF (Term Frequency – Inverse Document Frequency) algorithm is used to assign weight and relevance to a keyword based on its occurrence in the document [21].
- Doc2Vec [22] convert each abstract and title into a vector of scores that can be used to predict relevance.
- BERT (Bidirectional Encoder Representations from Transformers) [23] represents text accurately by considering the surrounding context but it is much slower than the other two approaches.

2.4 Classifier

An ML classifier utilises information (titles, keywords and abstracts) from the dataset of documents to score the relevance of a given document. It is trained on extracted features from the training set to predict accurately on the unlabelled documents. Below are some popular classifiers used in an ML model along with active learning [20]:

- Support vector machines (SVM): find a hyperplane to partition data points into groups [18].
- L2-regularized logistic regression (LR) - model probabilistic outcome of data.
- Naïve Bayes (NB): assumes all features are independent and assign probability to them separately [19, 21].

2.5 Active Learning in Screening Prioritisation

Screening prioritisation [16, 2] is a well-known approach for improving the performance of title and abstract screening. Screening prioritisation has been extensively researched in recent years, and it has been shown to be successful when combined with active learning [16, 2]. Using active learning in screening prioritisation, the screening process typically goes like below [2]:

- Reviewer begins with all unlabelled documents.
- Reviewer labels some documents, e.g., 10, then resulting in a set of labelled documents.
- The cycle of active learning begins with the following steps:
    1. The labelled documents are fed into an ML classifier to learn.
    2. The ML classifier assigns relevant scores for all unlabelled documents.
    3. The ML model picks documents with the highest assigned probability for being relevant.
    4. The ML model asks the reviewer to label this document.
    5. The reviewer screen the document providing a relevant or irrelevant label.
    6. The last labelled document is then added to the training data.
    7. Repeat from step 1 again.

Continue this loop until the point the reviewer wishes to end or all documents are labelled.

2.6 Evaluation Metrics

Various metrics have been used in several active learning papers to evaluate their ML model as studied by O'Mara-Eves et al. [2]. Four important metrics are selected to be briefly explained below:

- Recall [2] is the fraction of relevant references that are retrieved in the screening phase.
- Work Saved Over Sampling (WSS): defined by Cohen et al. [3] as the percentage of papers that do not need to be screened with active learning or assisted screening tool at a certain level of recall or discovery of retrieved relevant papers. The standard level used is WSS@95 which means how much works can be saved by finding only 95% of relevant papers and missing the other 5% papers [2].
- Relevant References Found (RRF): proposed by Schoot et al. [20] as a simple measure that indicates the percentage of relevant references discovered when a researcher has screened a specific percentage of the whole references, typically at 10% which is abbreviately written as RRF@10.

- Average Time to Discovery (ATD): proposed by Schoot et al. [20] as a metric that averages all the numbers of irrelevant papers a researcher needs to go through to reach a relevant paper. Unlike WSS and RRF, which only explain performance at an arbitrary point, ATD is used in this study as it considers performance over the whole screening process.

2.7 Existing SR Active Learning Solutions

The use of active learning models for SRs has been researched comprehensively [7, 32, 5, 7, 17, 24, 25, 26, 28, 29, 2]. Popular deployed tools that use active learning technique in the process of systematic review are Abstrackr [27], Rayyan [31], RobotAnalyst [30] and FASTREAD [11].

However, the existing tools have two main problems [20]. First, their applications do not open-source their codes so that other researcher can replicate their systematic reviews for the purpose of contributing to open science. Second, these tools are not flexible enough to handle a wide range of classifier types that would allow for a powerful benchmark testing.

One tool that fulfils this criterion comes to mind during the study. ASReview, developed by Schoot et al. [20] is an open-source, flexible, and reproducible AI-assisted SR screening platform that integrates active learning as well as a wide range of classifiers and feature extraction techniques. The power of ASReview was recently put to the test in a simulation study using four SRs with labelled data [20]. WSS was used to evaluate performance. WSS@95, for example, saves more time than conventional screening by identifying 95% of relevant documents. For a more rigorous standard, researchers would want to find 100% related documents, so WSS was set to WSS@100 [20]. The results showed that ASReview with WSS@95 saved the screening effort from 67% to 92% of documents, with an average of 83%. As for WSS@100, ASReview still did impressively ranging from 38% to 93% of documents, with an average of 61%. The evidence suggests that ASReview is a powerful and reproducible tool to accelerate the screening process and save significant time and money especially for an open-source tool.

2.8 Gaps and Opportunities

Despite its rich features and reproducibility, ASReview, as other screening tools, suffers from several pitfalls as expressed by Schoot et al. [20].

First, even the active learning considerably significantly cut downs the workload in screening documents, it also has a problem with poor recall estimation in case of unlabelled data. Recall estimation [32] is a method that predicts the fraction of retrieved relevant of documents over all relevant instances up to a certain point and inform reviewers when they can decide to stop screening and save time based on the approximated recall. This problem can be mitigated by experimenting with a novel recall estimation algorithm that uses each space between an irrelevant screening and a relevant screening as a basis for estimating the recall as introduced by Howard et al. [32].

Second, it is vulnerable to hasty generalisation, which is a condition where the classifier is biased towards documents that have existed in the training set the longest and would result in a biased training set which overlooks newer documents that might be of importance. To alleviate this bias effect, Singh et al. [7] has proposed a novel active learning algorithm based on Latent Dirichlet Allocation (LDA) topic modelling that dynamically explores various topics of a dataset in the early stage and switches to focusing on relevance of documents in the later stage.

3 TECHNICAL CHAPTER

In this chapter, a simulation study of active learning is conducted using ASReview on seven SR labelled datasets to partially reproduce and evaluate screening results as presented by Schoot et al. [20]. These datasets will be trained in simulation mode provided by ASReview on its default choice of model that has been proven as consistently outperforming other models in term of computational efficiency and classification performance over multiple experiments with number of diverse datasets from different fields (Table 1). This active learning simulation progressively learns labelled papers in each dataset the same way as how a researcher using ASReview screens each paper in a semi-automated active learning process [20]. The purpose of the simulation mode is to reproduce the whole screening process in order to assess performance and time saved using ASReview on existing labelled datasets. Quantitative results will be reported and discussed after performing the simulation.

3.1 Active Learning Model

The default model [20, 2] used for this active learning is consisted of:

- TF-IDF (Term Frequency – Inverse Document Frequency) feature extraction: as previously explained [21], assigns weight to a word based on its occurrences in a document. Even though it does not take into account the order and context of words, it has been widely evaluated as performing the same or better than other feature extraction techniques.
- Naïve Bayes classifier: a probabilistic ML model based on Bayes' Theorem that assumes independence among all features and assign separate weight to them accordingly [19, 21].
- Dynamic resampling: a novel undersampling technique that dynamically adjusts the ratio between irrelevant and relevant papers so that the irrelevant ones will not dominate the relevant ones during the training.
- Certainty-based sampling: queries only the papers with the highest chance of being relevant and presents to the researcher during the active learning cycle.

3.2 Datasets

Six SR datasets compiled by Schoot et al. [20] are integrated in the simulation based on the ground that they are truly open-source, labelled, reproducible and comprised of a diverse range of fields including social science, medical science and computer science. One more dataset has recently been manually labelled and obtained from Bailey et al. [42] who is undertaking research on the application of AI in diagnosing depression. The vast mixture of academic fields is highly preferred so as to

capture how the active learning system generalises in different research environments. From table 1, it is shown from the seven datasets that the number of relevant papers is only a sfraction of the total papers, ranging from under 1% to 6%. Hence, the task of manually screening these datasets is equivalent to finding a needle in a haystack. This laborious mission can be expedited and assisted by active learning and demonstrated thorough this simulation.

Dataset	ID	Total	Relevant	Relevance rate	Topic	Field	License
Nudging	Nagtegaal_2019 [41]	2019	101	5.00%	Nudging healthcare	Social science	CC0
Wilson	Appenzeller-Herzog_2020 [40]	3453	29	0.84%	Wilson disease	Medical science	CC-BY Attribution 4.0 International
Software	Hall_2012 [39]	8911	104	1.17%	Software Fault Prediction	Computer science	CC-BY Attribution 4.0 International
Ace	Cohen_2006_ACEInhibitors [3]	2544	41	1.61%	ACEInhibitors	Medical science	Custom open source
PTSD	van_de_Schoot_2017 [37]	6189	43	0.69%	PTSD Trajectories	Social science	CC-BY Attribution 4.0 International
Virus	Kwok_2020 [38]	2481	120	4.84%	Virus Metagenomics	Medical science	CC-BY Attribution 4.0 International
Depression	Bailey_2021 [42]	10319	604	5.85%	Depression detection	Computer science Medical science	Custom open source

Table 1: SR Datasets

3.3 Utilised Metrics

Three metrics are used in ASReview's active learning system to measure the performance of the seven datasets:

- Work Saved Over Sampling (WSS): in this simulation, WSS is evaluated at both 95% and 100% of recall and plotted on a recall curve. WSS@100 is incorporated because it is often the ultimate goal of some researchers to identify all relevant documents which is equivalent to a recall of 100% [20].
- Relevant References Found (RRF): this simulation incorporates both RRF@5 and RRF@10 to offer an early view into the number of found documents during the screening.
- Average Time to Discovery (ATD)

3.4 Simulation Results

Having performed the simulation of active learning screening of all seven datasets, the models are evaluated using several metrics and the findings are presented below.

Dataset	WSS@95	WSS@100	RRF@5	RRF@10	ATD
Nudging	48.50%	47.55%	35.00%	65.00%	10.90%
Wilson	51.05%	48.48%	71.43%	78.57%	7.62%
Software	91.90%	82.48%	99.03%	99.03%	1.33%
Ace	79.34%	25.65%	77.50%	85.00%	6.89%
PTSD	92.47%	92.00%	97.62%	100.00%	1.40%
Virus	67.80%	48.81%	42.86%	62.18%	9.18%
Depression	72.46%	18.45%	55.22%	77.78%	7.12%

Table 2: Simulation results

Based on table 2 and figure 1, after approximately screening between 8% to 50% of all the documents across all datasets, 95% relevant documents will be found. Notably, more than 90% of relevant documents can be identified at just 5% of the screening for Software and PTSD datasets which might have benefited from minimal number of relevant papers and lack of complex relevance condition. On the other hand, possibly more complex datasets like Nudging, Wilson, Ace, Virus and Depression can take considerable number of screenings before meaningful fraction of relevant papers is revealed.

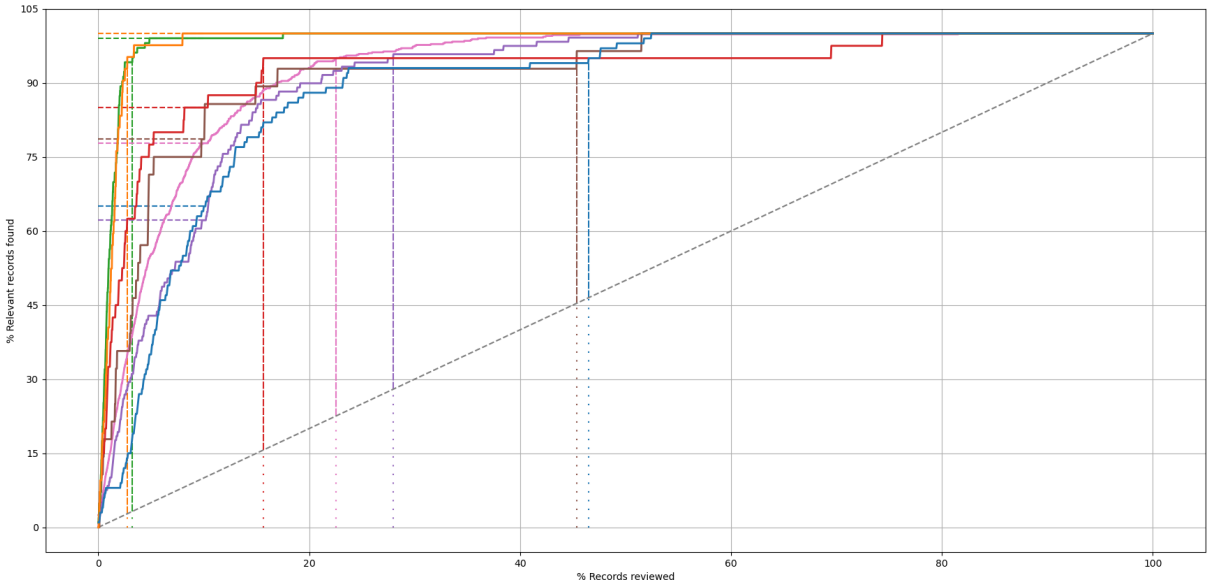

Figure 1: Recall curve

Ace and Depression datasets appear to take the longest time to screen in order to yield 100% of the relevant documents with a save in workload between 18% to 25%. Nevertheless, the majority of relevant papers can be identified from 60% up to 100% at the cost of only 10% of screening.

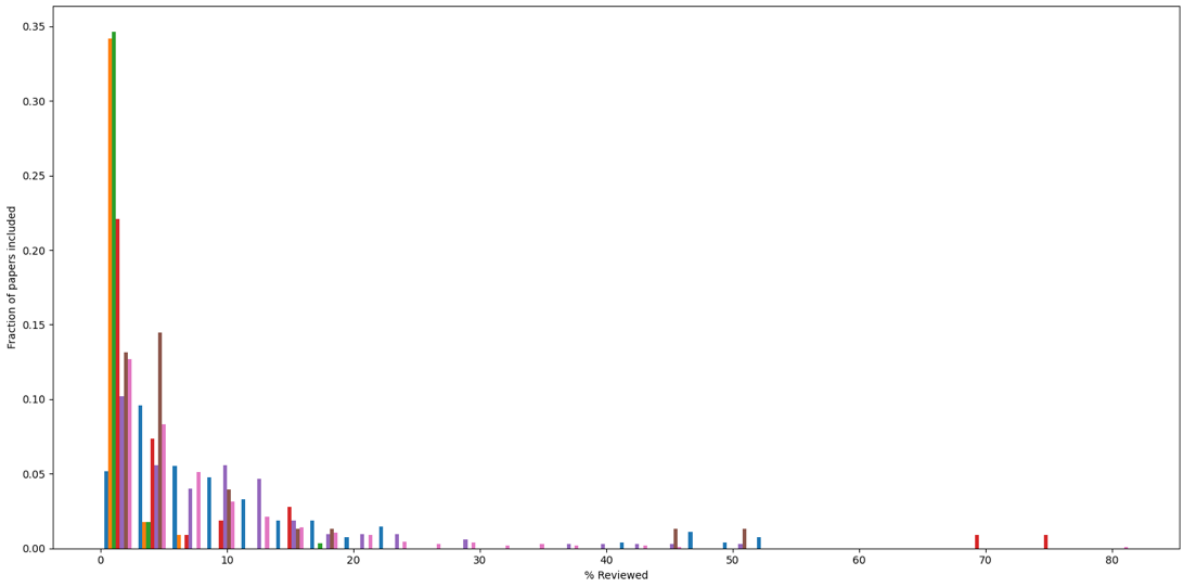

Figure 2: Discovery chart

As can be seen from figure 2, most of the relevant papers are already identified at the early stage of reviewing, which is in line to the finding from figure 1. Some remaining relevant papers belonging to datasets of Ace, Virus, Nudging and Depression can be quite challenging to detect as they take until the later stage of screening to reveal themselves. Moreover, most of the reproduced simulated results compare well with the original simulation results from [20] but some of them suffer from variations due to lack of multiple runs to get an overall average, which was employed by the original paper.

Overall, the active learning system has been proven by the simulation to deliver great reduction in screening effort. More than 48% to 92% of workload can be cut and yet 95% relevant papers can still be obtained which is extremely significant. There is no consistent pattern that indicates that certain academic fields might offer a faster and more simple screening as all three fields in the simulations vary in performance. Though, it can be inferred that size, complex context and relevance criteria of relevant papers can affect the process of active learning. The introduction of deep neural network into classifier and feature extraction might mitigate these issues and improve the active learning model in the future.

4 FUTURE PLAN

Based on the gaps identified and background study and simulation, there are several avenues for improvement that can be undertaken in the next stage of final report. Given that there is limited time and resource, the target implementations for enhancing the active learning model are listed in order of priority below:

- Feature extractions:
    - Bidirectional Encoder Representations from Transformers (BERT) [23]
    - Embedding LSTM (Long Short-Term Memory) [35]
    - Paragraph Vectors (Doc2Vec) [22, 24]
- Classifiers:
    - Dense neural network [33]
    - LSTM base classifier [34, 36]
    - LSTM pool classifier [34, 36]
    - Support vector machine (SVM) [18]
- Additional cross-field datasets
- Ensemble or average of classifiers
- Comparing performance with other screening tools
- Topic modelling prior to relevance classification [7]
- Enhanced recall estimation [32]

5 CONCLUSION

Systematic review is essential for summarising information for the purpose of guiding policies and practices. As the screening phase of systematic review requires a massive amount of time and resource, artificial intelligence has been introduced in order to speed up this process and save effort. Active learning system is a recent tool that incorporates artificial intelligence and assists researchers in finding relevant references without having to screen everything. Significant saving in time and effort has been proven by using ASReview, an active learning screening system, to screen multiple datasets from various research fields in a simulation study. Even though early results are promising in alleviating the screening effort, more experiments need to be conducted in the future by incorporating modern artificial intelligence techniques that can potentially enhance the screening system.

REFERENCES

[1]. K. S. Khan, G. ter Riet, J. Glanville, A. J. Sowden, and J. Kleijnen, *Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews*. York, England: Centre for Reviews and Dissemination, University of York, 2001.

[2]. A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou, "Using text mining for study identification in systematic reviews: a systematic review of current approaches," *Systematic Reviews*, vol. 4, no. 1, 2015.

[3]. A. M. Cohen, W. R. Hersh, K. Peterson, and P.-Y. Yen, "Reducing Workload in Systematic Review Preparation Using Automated Citation Classification," *Journal of the American Medical Informatics Association*, vol. 13, no. 2, pp. 206–219, 2006.

[4]. S. Ananiadou, B. Rea, N. Okazaki, R. Procter, and J. Thomas, "Supporting Systematic Reviews Using Text Mining," *Social Science Computer Review*, vol. 27, no. 4, pp. 509–523, 2009.

[5]. B. C. Wallace, T. A. Trikalinos, J. Lau, C. Brodley, and C. H. Schmid, "Semi-automated screening of biomedical citations for systematic reviews," *BMC Bioinformatics*, vol. 11, no. 1, 2010.

[6]. J. J. García Adeva, J. M. Pikatza Atxa, M. Ubeda Carrillo, and E. Ansuategi Zengotitabengoa, "Automatic text classification to support systematic reviews in medicine," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1498–1508, 2014.

[7]. G. Singh, J. Thomas, and J. Shawe-Taylor, "Improving Active Learning in Systematic Reviews," *arXiv.org*, 29-Jan-2018. [Online]. Available: https://arxiv.org/abs/1801.09496. [Accessed: 10-May-2021].

[8]. Y. Mo, G. Kontonatsios, and S. Ananiadou, "Supporting systematic reviews using LDA-based document representations," *Systematic Reviews*, vol. 4, no. 1, 2015.

[9]. S. Gaurav, J. Thomas, and J. Shawe-Taylor, "Improving active learning in systematic reviews," *arXiv preprint arXiv:1801.09496*, 2018.

[10]. H. Bastian, P. Glasziou, and I. Chalmers, "Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up?," *PLoS Medicine*, vol. 7, no. 9, 2010.

[11]. J. Lau, "Editorial: Systematic review automation thematic series," *Systematic Reviews*, vol. 8, no. 1, 2019.

[12]. A. M. O'Connor, G. Tsafnat, S. B. Gilbert, K. A. Thayer, and M. S. Wolfe, "Moving toward the automation of the systematic review process: a summary of discussions at the second

meeting of International Collaboration for the Automation of Systematic Reviews (ICASR)," *Systematic Reviews*, vol. 7, no. 1, 2018.

[13]. A. Barr, E. Feigenbaum, and C. Roads, "The Handbook of Artificial Intelligence, Volume 1," *Computer Music Journal*, vol. 6, no. 3, p. 78, 1982.

[14]. C. Dreisbach, T. A. Koleck, P. E. Bourne, and S. Bakken, "A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data," *International Journal of Medical Informatics*, vol. 125, pp. 37–46, 2019.

[15]. A. Bannach-Brown, P. Przybyła, J. Thomas, A. S. Rice, S. Ananiadou, J. Liao, and M. R. Macleod, "Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error," *Systematic Reviews*, vol. 8, no. 1, 2019.

[16]. I. Shemilt, A. Simon, G. J. Hollands, T. M. Marteau, D. Ogilvie, A. O'Mara-Eves, M. P. Kelly, and J. Thomas, "Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews," *Research Synthesis Methods*, vol. 5, no. 1, pp. 31–49, 2013.

[17]. M. Miwa, J. Thomas, A. O'Mara-Eves, and S. Ananiadou, "Reducing systematic review workload through certainty-based screening," *Journal of Biomedical Informatics*, vol. 51, pp. 242–253, 2014.

[18]. R. Wang, S. Kwong, and Q. He, "Active learning based on support vector machines," *2010 IEEE International Conference on Systems, Man and Cybernetics*, 2010.

[19]. H. D. Berrar, "Bayes' Theorem and Naïve Bayes Classifier," *Encyclopedia of Bioinformatics and Computational Biology*, pp. 403–412, 2019.

[20]. R. van de Schoot, J. de Bruin, R. Schram, P. Zahedi, J. de Boer, F. Weijdema, B. Kramer, M. Huijts, M. Hoogerwerf, G. Ferdinands, A. Harkema, J. Willemsen, Y. Ma, Q. Fang, S. Hindriks, L. Tummers, and D. L. Oberski, "An open source machine learning framework for efficient and transparent systematic reviews," *Nature Machine Intelligence*, vol. 3, no. 2, pp. 125–133, 2021.

[21]. J.-Y. Yoo and D. Yang, "Classification Scheme of Unstructured Text Document using TF-IDF and Naïve Bayes Classifier," 2015.

[22]. D. Kim, "Text Genre Detection Using Doc2Vec Word-embedding Language Model," *Language and Information*, vol. 23, no. 2, pp. 23–43, 2019.

[23]. K. Irie, A. Zeyer, R. Schlüter, and H. Ney, "Language Modeling with Deep Transformers," *Interspeech 2019*, 2019.

[24]. X. Qin, J. Liu, Y. Wang, Y. Liu, K. Deng, Y. Ma, K. Zou, L. Li, and X. Sun, "Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews," *Journal of Clinical Epidemiology*, vol. 133, pp. 121–129, 2021.

[25]. Z. Yu and T. Menzies, "FAST2: An intelligent assistant for finding relevant papers," *Expert Systems with Applications*, vol. 120, pp. 57–71, 2019.

[26]. Z. Yu, N. A. Kraft, and T. Menzies, "Finding better active learners for faster literature reviews," *Empirical Software Engineering*, vol. 23, no. 6, pp. 3161–3186, 2018.

[27]. A. Gates, C. Johnson, and L. Hartling, "Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool," *Systematic Reviews*, vol. 7, no. 1, 2018.

[28]. A. Carvallo, D. Parra, H. Lobel, and A. Soto, "Automatic document screening of medical literature using word and text embeddings in an active learning setting," *Scientometrics*, vol. 125, no. 3, pp. 3047–3084, 2020.

[29]. S. H. Cheng, C. Augustin, A. Bethel, D. Gill, S. Anzaroot, J. Brun, B. DeWilde, R. C. Minnich, R. Garside, Y. J. Masuda, D. C. Miller, D. Wilkie, S. Wongbusarakum, and M. C. McKinnon, "Using machine learning to advance synthesis and use of conservation and environmental evidence," *Conservation Biology*, vol. 32, no. 4, pp. 762–764, 2018.

[30]. P. Przybyła, A. J. Brockmeier, G. Kontonatsios, M. A. Le Pogam, J. McNaught, E. Elm, K. Nolan, and S. Ananiadou, "Prioritising references for systematic reviews with RobotAnalyst: A user study," *Research Synthesis Methods*, vol. 9, no. 3, pp. 470–488, 2018.

[31]. M. Ouzzani, H. Hammady, Z. Fedorowicz, and A. Elmagarmid, "Rayyan—a web and mobile app for systematic reviews," *Systematic Reviews*, vol. 5, no. 1, 2016.

[32]. B. E. Howard, J. Phillips, A. Tandon, A. Maharana, R. Elmore, D. Mav, A. Sedykh, K. Thayer, B. A. Merrick, V. Walker, A. Rooney, and R. R. Shah, "SWIFT-Active Screener: Accelerated document screening through active learning and integrated recall estimation," *Environment International*, vol. 138, p. 105623, 2020.

[33]. L. SiChen, "A Neural Network Based Text Classification with Attention Mechanism," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), 2019.

[34]. X. Bai, "Text classification based on LSTM and attention," 2018 Thirteenth International Conference on Digital Information Management (ICDIM), 2018.

[35]. Y. Li and M. Ye, "A Text Classification Model Base On Region Embedding AND LSTM," Proceedings of the 2020 6th International Conference on Computing and Artificial

Intelligence, 2020.

[36]. X. Shi and R. Lu, “Attention-Based Bidirectional Hierarchical LSTM Networks for Text Semantic Classification,” 2019 10th International Conference on Information Technology in Medicine and Education (ITME), 2019.

[37]. R. van de Schoot, M. Sijbrandij, S. D. Winter, S. Depaoli, and J. K. Vermunt, “The GRoLTS-Checklist: Guidelines for Reporting on Latent Trajectory Studies,” *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 24, no. 3, pp. 451–467, 2016.

[38]. K. T. Kwok, D. F. Nieuwenhuijse, M. V. Phan, and M. P. Koopmans, “Virus Metagenomics in Farm Animals: A Systematic Review,” *Viruses*, vol. 12, no. 1, p. 107, 2020.

[39]. T. Hall, S. Beecham, D. Bowes, D. Gray, and S. Counsell, “A Systematic Literature Review on Fault Prediction Performance in Software Engineering,” *IEEE Transactions on Software Engineering*, vol. 38, no. 6, pp. 1276–1304, 2012.

[40]. C. Appenzeller-Herzog, T. Mathes, M. L. S. Heeres, K. H. Weiss, R. H. J. Houwen, and H. Ewald, “Comparative effectiveness of common therapies for Wilson disease: A systematic review and meta-analysis of controlled studies,” *Liver International*, vol. 39, no. 11, pp. 2136–2152, 2019.

[41]. R. Nagtegaal, L. Tummers, M. Noordegraaf, and V. Bekkers, “Nudging healthcare professionals towards evidence-based medicine: A systematic scoping review,” *Journal of Behavioral Public Administration*, vol. 2, no. 2, 2019.

[42]. A. Bailey, Y. Abbassi, M. D. Bailey, and R. Nilforooshan, “Why is AI not utilised for systematic reviews?,” *The Lancet: Digital Health (Under Review)*, 2021.

APPENDIX 1 - WORK PLAN

Task	March 4th - April 4th	April 5th - April 21st	April 22nd - May 10th	May 11th - May 17th	May 11th - June 7th	June 8th - July 8th	July 8th - August 8th	August 8th - September 3rd	September 4 - september 11th
Learn basics of NLP, AI and Pytorch									
Research and compare published results on use of active learning in screening SRs									
Manual screening papers for AI Use for Depression Diagnosis									
Solving courseworks									
Write iterim report									
Implement a baseline AI for screening SRs based on a state-of-the-art paper									
Reproduce a simulation study using the baseline on multiple SRs									
Discuss the simulation results of the baseline using different metrics used in the baseline's paper									
Compare and report on the potential of improving the baseline AI for future work									
Interim report and viva									
Examination									
Test the baseline on SVM and Doc2Vec									
Implement deep learning (neural nework, BERT and LSTM) for feature extraction and classifier									
Use ensemble/average of classifiers to reduce variation									
Write final report									
Target 20 labelled datasets for diversity									
Explore the potential of integrating of a novel recall estimation method that may help push the screening towards complete automation									
Explore the potential of integrating an LDA-based topic modelling to migitate hasty generalization of the baseline									
Attempt replicating experiments on other screening tools and compare performance									
Final report and viva									

Figure 3: Work plan

APPENDIX 2 – TRAINING SUMMARY

In your first report you will have completed a training needs analysis table and an updated version of that template is in this appendix which you should complete into an updated version of that table, with guidance given in the template below. This is very important that you complete correctly as this will form part of your interim assessment as to how you are analysing your training needs and planning out your strategy to complete a successful dissertation.

Soft IT Skills

General IT Applications	Specific skills	Training needed? Y/N	Training completed if applicable? (Tick)
Word Processing (with Microsoft Word)	How to insert Powerpoint files into a Word document: https://www.youtube.com/watch?v=QVBoY2csyyE	N	
	How to insert pictures into a Word document: https://www.youtube.com/watch?v=sOrmu0RBMrU	N	
	How to insert tables into a Word document: https://www.youtube.com/watch?v=7mXRd3ZPW0k	N	
	How to insert equations into a Word document: https://www.youtube.com/watch?v=kXtxyx7hq7Y	N	
	How to do equation numbering in a Word document: https://www.youtube.com/watch?v=wM57WvO20KA	N	
	How to use heading styles in a Word document and create an automatic contents page: https://www.youtube.com/watch?v=OkyisWIE3kQ	N	
	How to insert captions into a Word document and create a list of figures or list of tables: https://www.youtube.com/watch?v=Xj9dgX0Bbew	N	
	How to make italic text for variables as well as subscripts and superscripts in Word: https://www.youtube.com/watch?v=L_u9sXm5IP8 https://www.youtube.com/watch?v=8SjbzLzSYhY	N	

	How to do cross referencing in Word: https://www.youtube.com/watch?v=vPsKMMDk5eo	N	
	How to insert a footnote in Word: https://www.youtube.com/watch?v=Hn1W7HjivFU	N	
	How to insert Greek symbols in Word: https://www.youtube.com/watch?v=xNQgUUi1mZo	N	
	How to insert a page break into Word: https://www.youtube.com/watch?v=kZpstbMAtNs	N	
	How to change line spacing in Word: https://www.youtube.com/watch?v=I5yhL5hKBx8	N	
Spreadsheets (with Microsoft Excel)	How to do summation of cells in Excel: https://www.youtube.com/watch?v=v3Ec5GupNjA	N	
	How to multiply two cells in Excel: https://www.youtube.com/watch?v=NXJOFJ9YESo	N	
	How to format cells in Excel: https://www.youtube.com/watch?v=L0cQ_lvL8LQ	N	
	How to do trigonometric functions in Excel: https://www.youtube.com/watch?v=Nlm2cl6OzaE	N	
	How to use the Log10 function in Excel: https://www.youtube.com/watch?v=tQl8xK4KTUY	N	
	How to make a graph or chart in Excel: https://www.youtube.com/watch?v=8B8kFVNzlQ8	N	
	Insert Excel table into Word https://www.youtube.com/watch?v=Jtk7lKGuhTA	N	
	Insert Excel Graph into Word https://www.youtube.com/watch?v=4tB-VSRJnWA	N	
Linux operating systems	How to log into Ubuntu in the FEPS computing labs. Self learning tutorial at: http://www.ee.surrey.ac.uk/Teaching/Unix/	N	

Professional Skills

Professional Skills	**Training needed Y/N?**	**Status of training. If not complete, what is the plan to complete it. If training was not needed, state how it had been completed previously.**
Report writing	N	Training provided during undergraduate years.
Revision techniques	N	Training provided during undergraduate years.
Research methods, literature reviewing (training provided by the library later in the semester – see timetable)	N	Training provided during undergraduate years.
Oral presentation skills	N	Training provided during undergraduate years.
CV writing (note free CV writing course from IET Lifeskills on October 24th!)	N	Training provided during undergraduate years.
Time planning/Project planning	N	Training provided during undergraduate years.
Plagiarism training (Go to SurreyLearn and find it in The Student Common Room)	N	Training provided during undergraduate years.
Applying for a job Note that “applying” to do a	N	Training provided during undergraduate years.

project is like applying for a job. You need to convince an academic that you are a student worth supervising, just like you would have to convince an employer you are worth employing.		

Specialist Skills

Note you should adapt this table specifically to your project, this template is just given as a guidance.

Specialist skill	**Training required Y/N?**
Refresher of common AI and NLP techniques required for AI Assisted SR project.	Y
How will I undertake training in above skills if required? Online tutorials.	
Mathematics – Refreshing or developing knowledge in maths. e.g. Maths Drop In Centre available: http://personal.ee.surrey.ac.uk/Personal/W.Wang/MathsDropInCentre.html The "Casual Drop-In Use" is available to MSc students.	N
How will I undertake training in mathematics if required?	
Hardware programming, e.g. Arduino, Raspberry PI, FPGA Note that the Electronics and Amateur Radio Society is a great extra-curricular way of broadening your skills in firmware programming with devices like Arduino. Why not go and join them and attend the courses that students deliver to students? More info at: https://www.ussu.co.uk/ClubsSocieties/societies/ears/Pages/home.aspx	N
How will I undertake training in hardware programming if required?	
Practical skills, e.g. soldering, test and measurement, PCB design You will need to discuss more specifically the details with your supervisor if you undertake a practical project of this nature and also you should ensure you have completed a risk assessment for specialist health and safety.	N
How will I undertake training in practical skills if required?	
Specialist simulation packages. Is there a specialist simulation package that you need to learn use?	N

If you do need to learn a specific simulation package you will certainly need to discuss this with your supervisor and find out how you will learn it.	
How will I undertake training in specialist simulation packages if required?	
Specialist test and measurement equipment. Some projects involve undertaking a number of practical measurements and may require learning specifically how to use the equipment required for the task.	N
How will I undertake training in specialist test and measurement equipment if required?	

Specific knowledge

Note you should adapt this table specifically to your project, this template is just given as a guidance.

Background state of the art
What are the key words that you need to search for in the literature regarding the state of the art in relation to your project? To discuss with your supervisor? There may be more than five key words so you can add more if you need to. 1. Artificial Intelligence 2. Systematic Review 3. Depression 4. Depressive Disorder 5. Active Learning
Who are the key scholars in the field where research papers they produce are of relevance to read? You can note any useful names and rough titles of papers that you should get hold of. Ensure you check you have found the right sources with your supervisor. Gaurav Singh A.M. Cohen Byron C. Wallace Yuanhan Mo

Theoretical knowledge for your project
What are the key items of theory you need to understand in order to undertake the work relevant to your project? Note that there may be more than five. If so, you need to insert them. 1. Active learning 2. Systematic review

3. Topic modelling 4. Keyword detection 5. Transfer learning
How will you obtain this knowledge, is it essential for you to study particular module(s) to ensure you have the knowledge you require to do your project? What textbooks or other resources are there available to help deepen your learning of the background theory? How is the best way to learn what you need to know? These you should also discuss carefully with your supervisor. Relevant online tutorials will be utilized to solidify my understanding of these topics.