A Critical Analysis of Studies that Address the Use of Text Mining for Citation Screening in Systematic Reviews

Babatunde K. Olorisade*
b.k.olorisade@keele.ac.uk

Ed de Quincey*
e.de.quincey@keele.ac.uk

Pearl Brereton*
o.p.brereton@keele.ac.uk

Peter Andras*
p.andras@keele.ac.uk

*School of Computing and Mathematics. Keele University

ABSTRACT

Background: Since the introduction of the systematic review process to Software Engineering in 2004, researchers have investigated a number of ways to mitigate the amount of effort and time taken to filter through large volumes of literature.

Aim: This study aims to provide a critical analysis of text mining techniques used to support the citation screening stage of the systematic review process.

Method: We critically re-reviewed papers included in a previous systematic review which addressed the use of text mining methods to support the screening of papers for inclusion in a review. The previous review did not provide a detailed analysis of the text mining methods used. We focus on the availability in the papers of information about the text mining methods employed, including the description and explanation of the methods, parameter settings, assessment of the appropriateness of their application given the size and dimensionality of the data used, performance on training, testing and validation data sets, and further information that may support the reproducibility of the included studies.

Results: Support Vector Machines (SVM), Naïve Bayes (NB) and Committee of classifiers (Ensemble) are the most used classification algorithms. In all of the studies, features were represented with Bag-of-Words (BOW) using both binary features (28%) and term frequency (66%). Five studies experimented with n-grams with n between 2 and 4, but mostly the unigram was used. χ^2, information gain and tf-idf were the most commonly used feature selection techniques. Feature extraction was rarely used although LDA and topic modelling were used. Recall, precision, F and AUC were the most used metrics and cross validation was also well used. More than half of the studies used a corpus size of below 1,000 documents for their experiments while corpus size for around 80% of the studies was 3,000 or fewer documents. The major common ground we found for comparing performance assessment based on independent replication of studies was the use of the same dataset but a sound performance comparison could not be established because the studies had little else in common. In most of the studies, insufficient information was reported to enable independent replication. The studies analysed generally did not include any discussion of the statistical appropriateness of the text mining method that they applied. In the case of applications of SVM, none of the studies report the number of support vectors that they found to indicate the complexity of the prediction engine that they use, making it impossible to judge the extent to which over-fitting might account for the good performance results.

Conclusions: There is yet to be concrete evidence about the effectiveness of text mining algorithms regarding their use in the automation of citation screening in systematic reviews. The studies indicate that options are still being explored, but there is a need for better reporting as well as more explicit process details and access to datasets to facilitate study replication for evidence strengthening. In general, the reader often gets the impression that text mining algorithms were applied as magic tools in the reviewed papers, relying on default settings or default optimization of available machine learning toolboxes without an in-depth understanding of the statistical validity and appropriateness of such tools for text mining purposes.

EASE '16, June 01-03, 2016, Limmerick, Ireland
 ISBN 978-1-4503-3691-8/16/06$15.00
DOI: http://dx.doi.org/10.1145/2915970.2915982

CCS Concepts

•*Computing methodologies*→*Artificial intelligence*→*Machine learning* •*General and reference*→*Cross-computing tools and techniques*→*Experimentation* •*General and reference*→*Document types*→*Surveys and overviews*

Keywords

Study Selection; Citation Screening; Systematic Review; Text Mining; Automation; Study reproduction.

1. INTRODUCTION

In recent years, there has been a marked increase in the use of Systematic Reviews (SRs) in software engineering (SE). Whilst this has generated feedback identifying possible implementation problems and proposed solutions, it has also aided the review of the initial guidelines proposed by Kitchenham [8, 19, 37, 44, 45]. One of the major problems faced by SR users is the amount of time and effort required to conduct a thorough SR [36, 42].

Therefore, there are ongoing efforts to automate part, or all of the stages of the SR process. One such approach is the application of Machine Learning (ML) techniques using text mining (TM) to automate the citation screening (CS) stage (also called study selection). However, there are currently no studies that focus on analysing the methods being used and the reproducibility of the reported results. The underlying question is how appropriate and transparent is the application of the ML techniques being used? This covers finding out if the parameters for the techniques are set in an informed way, are the methods applied in a statistically valid way – considering data size and method complexity and are the methods applied in a transparent way to enable independent verification?

This paper takes the form of a critical analysis of studies previously included in a SR of TM techniques for automatic citation screening [62]. The previous review focused only on non-technical aspects of the TM techniques used.

Here we look at the availability of information about the TM methods being used, including the description and explanation of the methods, parameter settings, assessment of the appropriateness of their application given the size and dimensionality of the data used, performance on training, testing and validation data sets, and further information that may aid the reproducibility of the included studies.

In the rest of the paper, section 2 discusses a range of methods proposed for automation of different Phases/stages of SR. The process of the study is presented in section 3, the results of the process in section 4, while section 5 focused on assessing the difficulty or otherwise of reproducing the studies. The paper is closed by a discussion and conclusions section.

2. BACKGROUND

There have been several attempts at reducing the amount of human time and effort required for SRs. While some attempts are being made at automating the entire process, others are focused on specific stages such as citation screening, or data extraction. The following sections discuss attempts at automation of SR process and some specific stages of the process.

2.1 Entire Process Automation

SLuRp, is a web based tool designed for the management of all types of data involved in the SR process [7]. It was developed to 'semi-automatically' search and retrieve studies from limited databases, capture data relating to the review being carried out, inclusion/exclusion criteria, reasons for acceptance/rejection, disagreement reconciliation and storage of full copies of included papers. Another tool with similar functionality is SLR-Tool [26]. The tool uses TM techniques to enhance decision making. SLR-Tool can store papers in pdf, communicate with bibliography management software and can also collect and import data to Excel among other functions [26]. StArt, is a tool reported in the literature for managing all phases of the SR except literature search however, it can read citations in BibTex format. It can rank papers and record information and decisions regarding each paper at different phases of the review process [38]. A recent addition to these tools is SESRA, a web based SR management tool [61].

Based on information provided in the papers, the major tasks of the SR supported - limited or fully - by each of the tools are presented in Table 1. The '√' sign indicates supported feature while '-', indicates otherwise. A more detailed comparative analysis of the features offered by these tools can be found in [53].

2.2 Automation of Specific Stages

A number of studies in recent reviews on methods for SR automation have indicated that there are more studies published on the automation of specific stages of the SR, most especially, citation screening and data extraction, than on the entire process [42, 62]. Work in this area is now focused beyond basic – software support development – of the SR processes and instead aims to create intelligent system (using Artificial Intelligent methods) that can make independent decisions and therefore reduce the human effort required in SR [42, 62].

Study identification, citation screening and data extraction are the three stages that are currently being focused upon based on available publications. These three are discussed below:

Table 1. SR phase managed by the tools

SR Stage	SLuRp	StArt	SLR-Tool	SESRA
Protocol development	-	√	-	√
Study identification	√	√	-	√
Study selection	√	√	√	√
Study evaluation	√	√	√	√
Data extraction	√	√	√	√
Data synthesis	√	√	√	√
Reporting	√	√	-	√

Study identification: A federated search tool has been developed to automate searching and retrieval of literature across multiple databases [33]. Tool developers reported promising result from its use across more than 10 databases. However, this tool is not publicly available nor has it been independently evaluated.

Citation screening: This stage has attracted the most attention in terms of an individual SR stage automation [52]. The majority of the efforts to automate the citation screening stage are centered on text mining techniques; these are explored in the context of easing the task of selecting the relevant studies from the results of the study search. Forty-four of these studies were reviewed and reported in [62]. The studies focused on a range of interests, from reducing screening workload to prioritisation of documents for screening. There is no overarching or widely accepted tool/method yet, but results are promising.

Data extraction: A recent review by Jonnalagadda et al. identified 26 studies focused on automation of the data extraction stage in SR [42]. The majority of the studies reviewed also used ML techniques for automation.

2.3 General Purpose Tools

There are also other software applications that support the SR even though they might not have been specifically developed for SR; mostly in this category are reference management applications such as - EndNote, Mendely, Refworks, Zotero, Excel etc. [38].

3. METHOD

This section presents the details of the process followed to conduct this study. We conducted a mapping study based on the articles reviewed in another SR publication that was the most recent on the subject when this study was conducted. We followed the SR guidelines in [45] for our study.

3.1 Research Questions

The research questions for this study focus on information regarding the techniques and how they were used. They were defined as follows:

RQ1.: What information is available on the use and distribution of specific TM algorithms being proposed to automate citation screening in SR - How well are the algorithms used described and/or justified in the context of use, what information is provided about the data size and to what extent is the effect of data size on the TM algorithm used taken into account?

RQ2.: What is the proportion of the included (positive example)/excluded (negative example) documents and how did the classifiers perform during training, validation and testing given the metrics used?

RQ3.: How comparable are the results of the different studies reviewed?

3.2 Search Strategy

We only retrieved and worked with the papers O'Mara-Eves et al. included in [62]. The O'Mara-Eves et al.'s study [62], selected papers on TM methods or metrics that were applied to the screening stage of a SR (or similar evidence review), however, the study did not look at the methods in any depth since their intended audience were users of the technologies rather than computer scientists. We chose O'Mara-Eves et al.'s article because it is a recent review on the subject and the most widespread we are aware of.

3.3 Selection Criteria

The initial inclusion criteria for this study is that the paper be one of the 44 papers reported in [62]; in addition to their criteria a set of secondary criteria particular to this study were also defined. They are:

- The publication must be reporting the outcome of a research exercise/experiment/case study/development.
- The topic of discussion or field of application must relate to ML-based TM classification model.
- The context of use must be citation screening in SR

To avoid duplication, studies reported across multiple publications are considered together and where papers report multiple studies, the studies are considered separately.

3.4 Data Extraction

The review team consists of four reviewers – the authors. The first author, the lead reviewer, reviewed all the papers. The papers were randomly divided amongst the other three reviewers. The data extraction form was designed using Excel. A pilot study was initially conducted to assess the form and reviewers' understanding of its fields. The extraction form was modified after the exercise to correct the inconsistencies identified. After the full data extraction, differences in the extracted data were resolved through two meetings involving all the reviewers; any outstanding differences were resolved through meetings between the lead reviewer and the other review team member concerned. No situation warranted inviting a third reviewer to mediate in any of the latter resolution meetings.

4. RESULTS

In this section, the research questions are answered based on the studies reviewed.

Eight of the 44 papers were excluded from this review because they did not fully meet one or more of the inclusion criteria for this study. Three were excluded because they are communication between different research teams as a follow up discussion on their previous studies' results [13, 57, 58]; there was no text mining experiment conducted in [73], though, systematic review was discussed in [72], the technique used is not ML-based. Two of the studies were excluded because the techniques used were neither ML-based nor applied within the SR context [27, 28]. The last paper was excluded because the focus of the study was on the performance of different feature selection techniques and not the classification model [67]. Additionally, we were unable to retrieve an unpublished article they included. The total number of papers included was 35. The number of studies reviewed in these 35 papers was 45.

4.1 Research Question 1

RQ1: *What information is available on the use and distribution of specific TM algorithm being proposed to automate citation screening in SR - How well are the algorithms used described and/or justified in the context of use, what information is provided about the data size and to what extent is the effect of data size on the TM algorithm used taken into account?*

Support Vector Machine (SVM) was the most used algorithm. It was used in 31% of the studies, excluding its usage in Ensemble of classifiers, and has been used in at least one experiment annually since 2006 (see Table 3). Ensemble of classifiers was used in 22% (see Table 3 and Figure 1) while Naïve Bayes (NB) was used in 14% of the studies. About 50% of the studies tried and reported more than one classifier. Their usage in the papers reviewed including other algorithms used is presented in Table 2.

Table 2. Classification algorithms used

S/N	Classifier	Paper
1	Support Vector Machine (SVM)	[9, 11, 12, 14–16, 31, 43, 51, 54, 76, 78–81]
2	EvoSVM	[4, 5]
3	Naïve Bayes (NB)	[4, 6, 9, 31, 51, 56, 74]
4	K-Nearest Neighbour	[4, 23, 31]
5	K-Means	[20, 23]
6	Complement Naïve Bayes (cNB)	[4, 29, 30]
7	Decision Tree (DT)	[5, 51]
8	WAODE	[5]
9	Neural Network (NN}	[17, 18]
10	Regression	[18]
11	Ensemble	[17, 29, 30, 48, 49, 60, 67, 69, 78, 81]
12	Rocchio	[31],
13	Distributional semantics with relevance feedback	[40]

Apart from the individual techniques, different variant options have been tried as shown in Table 4.

Less than 20% of the studies explained the algorithms they used in the context of their studies and provide some justification as to why the particular algorithm was chosen over other classification algorithms. None of the studies that used variants of an SVM classification algorithm or optimisation settings, e.g., kernels, C or gamma values justified or provided insights into why they chose one option over others.

In 70% of the cases, the studies reported using open access ML implementation frameworks like WEKA [35] with different settings mostly the default without discussing why they are suitable within the context of their own experiment(s).

The summary of the corpus sizes used in the studies is presented in Figure 2. None of the papers considered the impact of the corpus size on the statistical appropriateness of the application of the ML methods that they used.

In particular, the papers describing the application of SVM did not report the number of support vectors in the final classifier, which is critical information to confirm that overfitting by the classifier was avoided.

Table 3: Classifiers usage by year

Algorithm	2006	2007	2008	2009	2010	2011	2012	2013	2014	Total	Percentage
SVM	1	1	3	1	4	1	4		1	16	31%
EvoSVM					1		1			2	4%
NB		1			1	1	2		2	7	14%
cNB					1	1	1			3	6%
KNN						1	1		1	3	6%
k-Means						1	1			2	4%
Decision Tree		1			1					2	4%
WAODE					1					1	2%
NN	1									1	2%
Ensemble	1			2	3	1	2	1	1	11	22%
Regression							1			1	2%
Rocchio									1	1	2%
D. Semantics								1		1	2%

Except where explicit information was not provided, all the studies used the vector space model – 'Bag of Words' - for feature representation [46, 50].

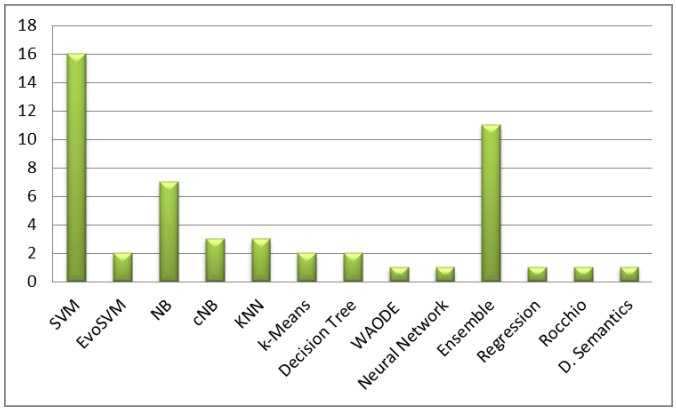

Figure 1. Number of classifiers used in the studies

Frequency based representation is the most used while a few have used binary feature representation (see Table 5).

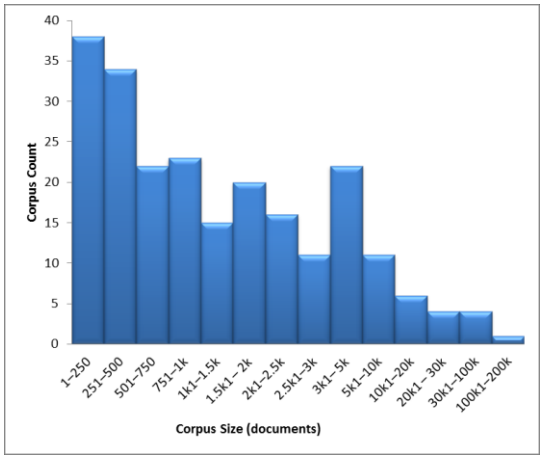

Figure 2. Corpus size range used across all studies

Some studies also experimented with multiple n-grams [4, 6, 12, 14, 16].

Table 4. Different kernels and classifier variants used

Classifier	Variant	Studies	Year
SVM	Linear Kernel	[9, 76, 80]	2012, 2011, 2010
	Radial Basis Function kernel	[5, 9, 69, 82]	2010, 2012, 2013, 2008
	Polynomial Kernel	[9]	2012
	Sigmoid	[9]	2012
	Epanechnikov (degree 3, 4)	[5]	2010
	Active Learning	[60, 76, 78, 79]	2014, 2011, 2012, 2010
KNN	K = 1	[4]	2012
Naïve Bayes	Multinomial	[4, 9]	2012, 2012
	Complimentary	[4, 6]	2012, 2014
Neural Network	Voting Perceptron	[17]	2006
	Generalized Linear Model	[18]	2012
Regression	Gradient Boosting Machine	[18]	2012
Ensemble	Voting	[11, 29, 30]	2011, 2006, 2010
	Bagging	[69, 77]	2013, 2012
	Unspecified	[49, 60, 81]	2009, 2009, 2014, 2010
	Query by Committee	[48]	2010

Feature selection/extraction (FS) techniques used across the studies are: term frequency (TF), term frequency – inverse document frequency (tf-idf), information gain (IG), Okapi BM25 (BM25), bi-normal separation (BNS), odds ratio (OR), signed

margin distance (SMD), normalized compression distance (NCD), cosine distance (CD), covariate shift (CS), aggressive under sampling + weighting (AU + W), linked document enrichment (LDE) and random indexing (RI).

Table 5. Feature representation usage in the studies

S/N	Feature representation	Count
1	Term frequency	25
2	Binary vector	7
3	SOSCO[1]	2
4	No Explicit Information[2]	4

Feature selection/extraction (FS) techniques used across the studies are: term frequency (TF), term frequency – inverse document frequency (tf-idf), information gain (IG), Okapi BM25 (BM25), bi-normal separation (BNS), odds ratio (OR), signed margin distance (SMD), normalized compression distance (NCD), cosine distance (CD), covariate shift (CS), aggressive under sampling + weighting (AU + W), linked document enrichment (LDE) and random indexing (RI).

Figure 3 shows the techniques and the number of times each was used across all the studies. There are situations where studies did not provide information concerning how FS was handled, 'INP' is used to signify such in Figure 3, whilst 'NA' means 'Not Applicable', for situations with no information. About 50% of the studies used multiple techniques to compare performance. Feature extraction approach was rarely used, LDA was used in [60] and topic modelling in [6].

Some of the studies have proposed novel tools, approaches or algorithms. An SVM based tool called GAPScreener was proposed in [82], ABSTACKR, an Active Learning based system was proposed in [78, 79]. A ranking algorithm was proposed in [48, 49], while a 'ranked-retrieval-re-rank' approach was proposed in [54]; a factorized form to cNB was proposed in [56]. Tomasseti proposed an enriched approach for feature selection based on linked data [74]. In [76], Wallace et al proposed a 'metacognitive Multiple Experts Active Learning (MEAL)' algorithm.

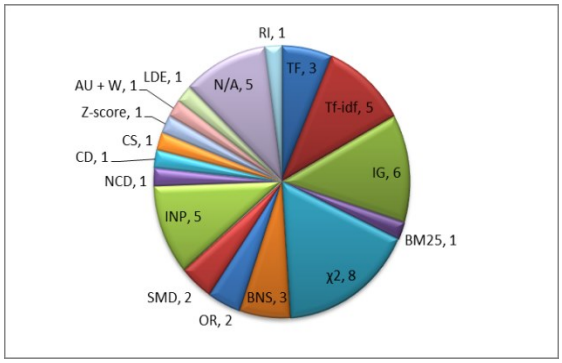

Figure 3. Feature selection/extraction distribution

ML toolboxes were used to carry out the experiments reported in 28 of the papers. The main toolboxes used are: WEKA [35], Projclus [65], Revis, PEx tool[3], Pimiento [1], RapidMiner[4], LibSVM[5] and SVMLight[6].

4.2 Research Question 2

RQ2: *What is the proportion of the included (positive example)/excluded (negative example) documents and how did the classifiers perform during training, validation and testing given the metrics used?*

The average percentage ratio of the positive to negative examples in the corpus used for 90% or more of the studies is 10%:90%. The studies tried to maintain this ratio in the training and test data. This issue of class imbalance was handled in different ways across the studies, see [62] for a summary of the different approaches used.

The majority of the studies used cross validation (CV) – 5x2 cross validation was used in [6, 10, 12, 16, 17, 49, 56] and 10-fold cross validation in [4, 5, 9, 31, 48, 74, 76], 5-fold cross validation was used in [18]; [48, 49] used both 5x2 and 10-fold CV with stratified random sampling, multiple n-way cross validation with *n* ranging between 2 to 256 increasing by power of 2 was used in [15] while cost rejection sampling was used in [11].

In terms of performance metrics, (mean) recall, (mean) precision, (mean) F and the area under the receiver operating characteristics curve (AUC) were mostly used. High recall implies few false negatives in the result while high precision implies few false positives. The F-measure is a weighted harmonic mean assessing the precision-recall trade-off and AUC is the probability that a model will rank a randomly chosen positive sample higher than a randomly chosen negative sample.

Mean recall was 95% and above in [4, 5, 17, 30, 48, 56, 74, 77, 82] while it was below 95% in [9, 49]. Precision on the other hand was over 10% in [4, 5, 17, 30, 48, 49, 82]. AUC was used in [12, 18, 48, 55, 60, 82] and the result was over 0.5 in all the studies. Cohen et al proposed a metric based on the amount of manual work saved – Work Saved over Sampling (WSS) which they defined in [17]. This measure was also used in [29, 41, 54, 56] to determine how much manual screening effort was saved given the classification result. Training performance was mostly sustained during testing or cross validation.

4.3 Research Question 3

RQ3: *How comparable are the results of the different studies reviewed?*

Comparing the performance of classifiers from different experimental settings is not trivial in ML. The performance of classifiers is usually specific to the context of use, thus, it is not easy to compare classifiers trained and used on different datasets [68] or from different experiments [2, 47]. It may be possible to compare, when the same dataset is used for different classifiers in different experiments, but if, for example, the dataset was not split in exactly the same way the comparison is still questionable. This is the case with most of the studies reviewed above, though, two of them used the same data set in their experiments (Table 6), and there was no record of whether a replica of the training set and test set in an experiment was repeated in another. Some of the researchers have attempted to establish some comparison between

[1] Second Order Soft Co-Occurrence

[2] The studies did not provide explicit information and we avoid drawing inference.

[3] http://infoserver.lcad.icmc.usp.br/infovis2/PEx

[4] https://rapidminer.com/

[5] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[6] http://svmlight.joachims.org/

the techniques based on same dataset used [13, 17, 40, 58]. None of the studies explicitly explained which portion of their dataset was used as training and test portions.

The datasets used in more than one paper are presented in Table 6 along with the classifiers and metrics used. Where classifiers are compared in a study '>' is used to denote 'better than' in respect of reported performance, otherwise, the classifier used is just listed under comment. The table is not presented for the purposes of comparison purpose but to gain insight into study variability based on dataset, metrics and classification model. It can be inferred from the extent of variability in metrics and techniques (comment) in Table 6 that datasets are being reused without any actual relation to the results (and/or process) of previous experiments that had used the same data.

Table 6. Studies with same dataset

S/N	Dataset	Paper	Metrics	Comment
1	Drug Evaluation Review Project (DERP)[7]	[9]	AUC	SVM
		[56]	WSS@95%	FCNB
		[43]	accuracy	SVM
		[12]	AUC	SVM
		[5]	Recall, precision, F_1	EvoSVM > WAODE > NB
		[16]	Recall, precision, F	SVM
		[55]	WSS, AUC	SVR
		[11]	U_n	SVM
		[40]	WSS	Relevance feedback
		[17]	Recall, precision, F_1, WSS	Perceptron
		[12, 14, 15]	AUC	SVM
2	TrialStat SR	[29]	Recall, precision, F, WSS	cNB
		[30]	Recall, precision, F	SVM ≈ NB
		[67]		Ensemble
		[49]	Recall, precision, workload save	Ensemble
		[48]	False negatives	Ensemble
3	Chronic Obstructive Pulmonary Disease (COPD)	[76]	U_{19}	MEAL (SVM) > PAL (SVM)
		[80]	U_{19}	SVM(coFeature) > (Simple) > (Random) > (Features Simple)
		[81]	Yield, burden	SVM (AL)
		[60]	Utility, coverage, AUC	Ensemble SVM
4	Proton beam	[79]		SVM (AL)
		[80]	U_{19}	SVM(coFeature) > (Simple) > (Random) > (Features Simple)
		[81]	Yield, burden	SVM Ensembles
		[60]	Utility, coverage	Ensemble SVM
5	Micro nutrients	[80]	U_{19}	SVM(coFeature) = (lp) > (Random) > (Simple) > (Features Simple)
		[81]	Yield, burden	SVM (AL)
		[60]	Utility, coverage	Ensemble SVM

[7] Some of the studies used fewer review set than others but they mostly share common 15 studies. The dataset is the same as Text Retrieval Conference (TREC) genomics data.

5. STUDY REPLICATION

Replication of experiments is an established practice in science and engineering to underpin theories and techniques, especially in a growing field [3, 63, 64]. This principle has also been recognized and encouraged in software engineering demonstrated by the existence of research groups with 'empirical' or 'evidence based' attached to their names [34]. Study reproductions with the same, similar or different dataset are useful to verify, extend or complement existing results [34, 71, 75].

Although, study replication was not part of our research questions, it's relevance manifested during the study and we felt strongly to report our experience briefly because of its importance; it can aid the building of a sustainable knowledge to advance any discipline [71, 75]. We will prepare a separate publication to fully address it within this context.

Considering the nascent stage of systematic review in software engineering and the application of text mining to the automation of some of its stages, it is thus important for independent research teams to reproduce published studies in whole or part [71] as a means to establish efficiency, maturity and applicability of proposed methods and techniques [59].

Adapting the guidelines suggested in [34], essential artefacts that can influence reproduction in the context of using text mining to automate the citation screening phase of systematic review as we identify are: data source, dataset, pre-processing, dimensionality reduction technique, classification method, model assessment approach, machine learning (third party) implementation tool used and any other local software built by the researchers used.

Based on this, we evaluate how difficult or easy it will be to replicate the studies reviewed. The individual results are not reported since the focus of this paper is not about issues regarding reproducible research. Extensive details will be addressed in a separate publication. Our perspective is to reflect on broader view of consciousness or otherwise of reproducibility and replication during studies and not target any particular paper.

González-Barahona and Robles proposed assessment of reproducibility/replication under five criteria [34]:

i. Identification (ID) – source of the (original) data
ii. Description (Desc.) – assessment of details of published description (internal organization and structure) of the dataset
iii. Availability (Av.) – Accessibility to the dataset by other researcher
iv. Persistence (Per.) – Chances of being available at source for a long period
v. Flexibility (Flex.) – Adaptability of the data to new environments

These criteria are judged as: N – not usable for reproduction, D – usable for reproduction with some difficulty, U – usable for reproduction, '-' – irrelevant or non-existent, '+' – availability foreseeable in future and '*' – flexibility.

Guided by these options and using the information provided in the studies we constructed a simplified chart of possible TM experiment activities (MS) in Table 7 and explain below:

i. Data source (DS): Out of the 35 papers reviewed, only [21–25] did not mention the source of their raw data but they all provided information on the structure (usually title, abstract and optional keywords or metadata) of their data; none of them provided the link to the data except [5, 11, 30, 48]. The datasets are no more available at the links provided in [30, 48]. In addition, the link provided for the DERP datasets in [5] is broken. All of the datasets are for private use where membership of a certain professional group may be required to access them except the 'TREC' dataset.
The data source can always be accessed for the same dataset but may be with difficulty in some situations like the need to contact previous researchers for link update or seek approval from the custodians before use.

ii. Dataset (DT): Dataset in this context refers to the set specifically used in different studies. The majority of the datasets are reusable and have actually been used in several studies. The datasets have the potential of being available in the future for use and may be adaptable to different formats. Some studies used smaller private datasets that cannot be accessed independently and no description was provided.
There are cases where the whole dataset from a source was not completely utilized. Access to this set is not offered in any of the studies.
None of the studies gave details about data partitioning – training and test set. It is unclear which part of the dataset was used as the training set and which part as the test set. This information may be essential for partial reproduction using the same dataset but different classifier for example. Even if data was split randomly information about any seed value applied may help in obtaining a similar split in future.
Overall the datasets can be reused in reproduction with some difficulty.

iii. Dimensionality reduction (DR): The pre-processing and other dimensionality reduction techniques used are the general ones available in regular Machine Learning literatures. However, less than 30% of the studies provided information about their final feature set – size or access to the set.
The pre-processing (PP) and feature extraction activities are fully accessible and re-usable but the feature set are not reusable except where they have been stored, in which case there is a chance that they can be made available.

iv. Classification (CL): Classical machine learning algorithms were mostly used. They are available and accessible.
Classification methods are reusable, accessible and are likely to be around for some time.

v. Third party ML framework (3rd): All third party machine learning implementations used are open source and publicly available. These are tools developed for machine learning and made available for public use utilized in the studies.
The tools are reusable and are likely to be accessible in the future.

vi. Local tool (LT): This refers to any other software or algorithm developed/proposed by the research team for the purpose of the study or as extension to the public third party tool. When new methods or algorithms are proposed, they are described in sufficient details only in about 30% of the studies. Proposed algorithm implementations were not provided; some tools mentioned in the studies are neither described nor accessible publicly.
The tools are not reusable since limited information is available about them. This status may change in the future if they are made accessible to the public with sufficient documentation.

Table 7. Compressed reproducibility assessment chart

MS	ID	Desc.	Av.	Per.	Flex.
DS	Partial	No	Public/ Private	Likely	Complete
DT	Partial	No	Public/ Private	Likely	Complete
PP	Complete	Detailed	Public	Likely	N/A
DR	Complete	Detailed	Public	Likely	N/A
CL	Partial	Fairly	Public	Likely	N/A
ASS	Complete	Detailed	Public	Likely	N/A
3rd	Complete	No	Public	Likely	-
LT	No	No	No	Likely	Possible

6. VALIDITY THREAT

The articles used in this study are limited to those previously included in an existing SR. our results therefore, are affected by the completeness of the published SR. Furthermore, we have not extended the search to include relevant studies published since February 2014. It is likely, however, that the studies we included are representative of the field. The results about dataset accessibility are based strictly on the information provided in the reviewed papers.

7. DISCUSSION AND CONCLUSIONS

This section presents a general discussion on the results presented in section 4 and section 0.

The SVM has the advantage of coping with high dimensional data without significant impact from class imbalance. It is less affected by the size of its input and requires moderate samples for training [2, 47, 68]. It is also suited to high feature to low training instance situation [39]. These facts might have accounted for the performance recorded and substantial use of SVM in the studies. Attempts to ensure more reliable classification performance

results might have accounted for the high use of ensemble methods as well. NB on the other hand did not perform well. Replicated studies by independent research teams are required to verify and extend existing results. Replications in the studies reviewed were often conducted by the same research groups. One such in-team replication led to the creation of ABSTACKR [78, 79], a tool developed by Wallace et al., that has been evaluated by another group in [66].

The corpus sizes used across the studies as shown in Figure 2 suggest that the majority of the experiments used corpus sizes that calls into question the statistical reliability of the classification model built through such corpus. There was rarely any justification across all the studies for the different decisions about the choice of a certain technique or approach within the context of use.

Insufficient information makes it hard to assess the process and statistical validity of the majority of the studies, for example, none of the studies that used SVM reported the number of support vectors they found. Similarly, in the case of the application of neural networks, there was no information on the number of neurons or hidden layers used. Thus, it is hard to judge how over-fitting was controlled and to what extent the complexity of the classifier was considered. There was no mention of the bias/variance trade-off characteristics of the classification algorithms and the impact of the data size in this context. The role data size plays in learning, generalisation ability and classification performance of a model was not emphasized in any of the studies. Notably, the positive to negative example ratio with the number of effective parameters (complexity) is quite important to determine the size of data necessary for the statistical validity of a model; the higher the complexity and the lower the positive to negative ratio, the more data is required to train an appropriate model.

Replication by independent research teams is possible but with different levels of difficulty specific to each study. Studies in this field need to be reported with more information than is currently the practice to aid independent reproduction of the studies. One suggestion would be to create a common repository where research results can be stored along with associated datasets, partition information and process details [32]. This ensures persistence and availability of datasets, as well as information not included in publications. Also, communication may improve between researchers due to the need for further explanation or elicitation of undocumented tacit knowledge or ideas used in the original experiment. This type of communication has been established to help better replication [70, 75].

We note that some of the studies have been replicated; in fact six datasets were found to be used by more than one study (see Table 6). In addition, Cohen et al. and Matwin et al.'s teams are already comparing model results based on use of the same dataset [12, 13, 56, 58], Felizardo et al. have also replicated their study [23, 25]. However, more work needs to be done given the fact that SR is now cutting across disciplines from medicine to social science, software engineering and computer science. In order to build useful tools, research teams may require access to data used in studies from other disciplines which may not be as readily available compared to data from within the discipline.

Although, a lot of studies are already published in this area, there is yet to be any concrete headway commensurate with the amount of research effort so far. Obviously, there is a need for more collaborative effort among research teams. Public availability of data and process description need to be considered for convenient study reproduction. More efforts need to be channelled into tools packaged for cross-domain use.

8. REFERENCES

[1] Adeva, J.J.G. and Calvo, R. 2006. Mining text with pimiento. *IEEE Internet Computing*. 10, 4 (2006), 27–35.

[2] Baharudin, B., Lee, L.H. and Khan, K. 2010. A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*. 1, 1 (2010), 4–20.

[3] Basili, V.R., Shull, F. and Lanubile, F. 1999. Building Knowledge through Families of Software Studies : *Software Engineering, IEEE Transactions on*. 25, 4 (1999), 456–473.

[4] Bekhuis, T. and Demner-Fushman, D. 2012. Screening nonrandomized studies for medical systematic reviews: A comparative study of classifiers. *Artificial Intelligence in Medicine*. 55, 3 (2012), 197–207.

[5] Bekhuis, T. and Demner-Fushman, D. 2010. Towards automating the initial screening phase of a systematic review. *Studies in Health Technology and Informatics*. 160, PART 1 (2010), 146–150.

[6] Bekhuis, T., Tseytlin, E., Mitchell, K.J. and Demner-Fushman, D. 2014. Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. *PLoS ONE*. 9, 1 (2014), 1–10.

[7] Bowes, D., Hall, T. and Beecham, S. 2012. SLuRp : A Tool to Help Large Complex Systematic Literature Reviews Deliver Valid and Rigorous Results. *Proceedings of the 2nd international workshop on Evidential assessment of software technologies - EAST '12*. (2012), 33–36.

[8] Carver, J.C., Hassler, E., Hernandes, E. and Kraft, N.A. 2013. Identifying Barriers to the Systematic Literature Review Process. *2013 ACM / IEEE International Symposium on Empirical Software Engineering and Measurement* (2013), 203–212.

[9] Choi, S., Ryu, B., Yoo, S. and Choi, J. 2012. Combining relevancy and methodological quality into a single ranking for evidence-based medicine. *Information Sciences*. 214, (2012), 76–90.

[10] Cohen, Aaron M. Ambert, K. and McDonagh, M. 2010. A Prospective Evaluation of an Automated Classification System to Support Evidence-based Medicine and Systematic Review. *AMIA Annual Symposium Proceedings*. 2010, (2010), 121–125.

[11] Cohen, A.M. 2006. An effective general purpose approach for automated biomedical document classification. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* (2006), 161–165.

[12] Cohen, A.M. 2008. Optimizing feature representation for automated systematic review work prioritization. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* (2008), 121–125.

[13] Cohen, A.M. 2011. Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@95 measure. *Journal of the American Medical Informatics Association : JAMIA*. 18, 1 (2011), 104; author reply 104–105.

[14] Cohen, A.M., Ambert, K. and McDonagh, M. 2010. A Prospective Evaluation of an Automated Classification

System to Support Evidence-based Medicine and Systematic Review. *AMIA Annual Symposium Proceedings* (2010), 121–125.

[15] Cohen, A.M., Ambert, K. and McDonagh, M. 2009. Cross-Topic Learning for Work Prioritization in Systematic Review Creation and Update. *Journal of the American Medical Informatics Association*. 16, 5 (2009), 690–704.

[16] Cohen, A.M., Ambert, K. and McDonagh, M. 2012. Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC Medical Informatics and Decision Making*. 12, 1 (2012), 33.

[17] Cohen, A.M., Hersh, W.R., Peterson, K. and Yen, P.Y. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*. 13, 2 (2006), 206–219.

[18] Dalal, S.R., Shekelle, P.G., Hempel, S., Newberry, S.J., Motala, A. and Shetty, K.D. 2012. A Pilot Study Using Machine Learning and Domain Knowledge to Facilitate Comparative Effectiveness Review Updating. *Medical Decision Making*. (2012), 1–13.

[19] Dybå, T., Dingsøyr, T. and Hanssen, G.K. 2007. Applying systematic reviews to diverse study types: An experience report. *Proceedings - 1st International Symposium on Empirical Software Engineering and Measurement, ESEM 2007* (2007), 225–234.

[20] Felizardo, K.R., Andery, G.F., Paulovich, F. V., Minghim, R. and Maldonado, J.C. 2012. A visual analysis approach to validate the selection review of primary studies in systematic reviews. *Information and Software Technology*. 54, 10 (2012), 1079–1091.

[21] Felizardo, K.R., Nakagawa, E.Y., Feitosa, D., Minghim, R., Mapping, S., Mining, V.T. and Maldonado, J.C. 2010. An approach based on visual text mining to support categorization and classification in the systematic mapping. *Proc. of EASE* (2010), 1–10.

[22] Felizardo, K.R., Riaz, M., Sulayman, M., Mendes, E., MacDonell, S.G. and Maldonado, J.C. 2011. Analysing the use of graphs to represent the results of systematic reviews in software engineering. *Proceedings - 25th Brazilian Symposium on Software Engineering, SBES 2011* (2011), 174–183.

[23] Felizardo, K.R., Salleh, N., Martins, R.M., Mendes, E., MacDonell, S.G. and Maldonado, J.C. 2011. Using Visual Text Mining to Support the Study Selection Activity in Systematic Literature Reviews. *2011 International Symposium on Empirical Software Engineering and Measurement*. (2011), 77–86.

[24] Felizardo, K.R., Salleh, N., Martins, R.M., Mendes, E., MacDonell, S.G. and Maldonado, J.C. 2011. Using Visual Text Mining to Support the Study Selection Activity in Systematic Literature Reviews. *2011 International Symposium on Empirical Software Engineering and Measurement* (2011), 77–86.

[25] Felizardo, K.R., Souza, S.R.S. and Maldonado, J.C. 2013. The use of visual text mining to support the study selection activity in systematic literature reviews: A replication study. *Proceedings - 2013 3rd International Workshop on Replication in Empirical Software Engineering Research, RESER 2013*. (2013), 91–100.

[26] Fernández-Sáez, A.M.., Bocco, M.G.. and Romero, F.P.. 2010. SLR-Tool a tool for performing systematic literature reviews. *ICSOFT 2010 - Proceedings of the 5th International Conference on Software and Data Technologies*. 2, September 2015 (2010), 157–166.

[27] Fiszman, M., Bray, B.E., Shin, D., Kilicoglu, H., Bennett, G.C., Bodenreider, O. and Rindflesch, T.C. 2010. Combining relevance assignment with quality of the evidence to support guideline development. *Studies in Health Technology and Informatics*. 160, PART 1 (2010), 709–713.

[28] Fiszman, M., Ortiz, E., Bray, B.E. and Rindflesch, T.C. 2008. Semantic processing to support clinical guideline development. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* (2008), 187–191.

[29] Frunza, O., Inkpen, D. and Matwin, S. 2010. Building systematic reviews using automatic text classification techniques. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (2010), 301–311.

[30] Frunza, O., Inkpen, D., Matwin, S., Klement, W. and O'Blenis, P. 2011. Exploiting the systematic review protocol for classification of medical abstracts. *Artificial Intelligence in Medicine*. 51, 1 (2011), 17–25.

[31] García Adeva, J.J., Pikatza Atxa, J.M., Ubeda Carrillo, M. and Ansuategi Zengotitabengoa, E. 2014. Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*. 41, 4 PART 1 (2014), 1498–1508.

[32] Gentleman, R. and Lang, D.T. 2004. Statistical Analyses and Reproducible Research. *Journal of Computational and Graphical Statistics*. 5, 1 (2004), 38.

[33] Ghafari, M., Saleh, M. and Ebrahimi, T. 2012. A federated search approach to facilitate systematic literature review in software engineering. *International Journal of Software Engineering Applications*. 3, 2 (2012), 13–24.

[34] González-Barahona, J.M. and Robles, G. 2011. On the reproducibility of empirical software engineering studies based on data retrieved from development repositories. *Empirical Software Engineering*. 17, 1-2 (2011), 75–89.

[35] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*. 11, 1 (2009), 10–18.

[36] Handbook for Systematic Reviews of Interventions: 2015. *http://handbook.cochrane.org/*. Accessed: 2015-09-10.

[37] Hassler, E., Carver, J.C., Kraft, N.A. and Hale, D. 2014. Outcomes of a community workshop to identify and rank barriers to the systematic literature review process. *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14* (2014), 1–10.

[38] Hernandes, E., Zamboni, A., Fabbri, S., Carlos, S., Thommazo, A. Di and Carlos, S. 2012. Using GQM and TAM to evaluate StArt-a tool that supports Systematic Review. *CLEI Electronic Journal*. 15, 1 (2012), 3.

[39] Ikonomakis, M., Kotsiantis, S. and Tampakas, V. 2005.

Text classification using machine learning techniques. *WSEAS Transactions on Computers*. 4, 8 (2005), 966–974.

[40] Jonnalagadda, S. and Petitti, D. 2013. A New Iterative Method to Reduce Workload in the Systematic Review Process. *International journal of computational biology and drug design*. 6, 0 (Feb. 2013), 5–17.

[41] Jonnalagadda, S., Petitti, D. and Manuscript, A. 2013. A New Iterative Method to Reduce Workload in the Systematic Review Process. *International journal of computational biology and drug design*. 6, 0 (Feb. 2013), 5–17.

[42] Jonnalagadda, S.R., Goyal, P. and Huffman, M.D. 2015. Automating data extraction in systematic reviews: a systematic review. *Systematic Reviews*. 4, 1 (2015), 78.

[43] Kim, S. and Choi, J. 2012. Improving the performance of text categorization models used for the selection of high quality articles. *Healthcare Informatics Research*. 18, 1 (2012), 18–28.

[44] Kitchenham, B. and Brereton, P. 2013. A systematic review of systematic review process research in software engineering. *Information and Software Technology*. 55, 12 (2013), 2049–2075.

[45] Kitchenham, B. and Charters, S. 2007. Guidelines for performing Systematic Literature Reviews in Software Engineering. *Technical report, Keele University and Durham University Joint Report. EBSE Technical Report. EBSE*. Ver. 2.3.

[46] Korde, V. and Mahender, C.N. 2012. Text Classification and Classifiers: A Survey. *International Journal of Artificial Intelligence & Applications*. 3, 2 (2012), 85–99.

[47] Kotsiantis, S., Zaharakis, I. and Pintelas, P. 2007. Supervised machine learning: A review of classification techniques. *Informatica*. 31, (2007), 249–268.

[48] Kouznetsov, A. and Japkowicz, N. 2010. Using classifier performance visualization to improve collective ranking techniques for biomedical abstracts classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 6085 LNAI, (2010), 299–303.

[49] Kouznetsov, A., Matwin, S., Inkpen, D., Razavi, A.H., Frunza, O., Sehatkar, M., Seaward, L. and O'Blenis, P. 2009. Classifying biomedical abstracts using committees of classifiers and collective ranking techniques. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 5549 LNAI, (2009), 224–228.

[50] Kumar, A.A. 2012. Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering. *International Journal of Engineering Research & Technology (IJERT)*. 1, 5 (2012), 1–6.

[51] Ma, Y. 2007. *Text classification on imbalanced data: Application to Systematic Reviews Automation*. University of Ottawa.

[52] Marshall, C. and Brereton, P. 2013. Tools to Support Systematic Literature Reviews in Software Engineering: A Mapping Study. *2013 ACM / IEEE International Symposium on Empirical Software Engineering and Measurement* (2013), 296–299.

[53] Marshall, C., Brereton, P. and Kitchenham, B. 2014. Tools to Support Systematic Reviews in Software Engineering : A Feature Analysis. *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (2014), 139–148.

[54] Martinez, D., Karimi, S., Cavedon, L. and Baldwin, T. 2008. Facilitating Biomedical Systematic Reviews Using Ranked Text Retrieval and Classification. *Australasian Document Computing Symposium ADCS*. December (2008), 53–60.

[55] Martinez, D., Karimi, S., Cavedon, L. and Baldwin, T. 2008. Facilitating biomedical systematic reviews using ranked text retrieval and classification. *Australasian Document Computing Symposium (ADCS)* (2008), 53–60.

[56] Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O. and O'Blenis, P. 2010. A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association : JAMIA*. 17, 4 (2010), 446–453.

[57] Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O. and O'Blenis, P. 2011. Performance of SVM and Bayesian classifiers on the systematic review classification task. *Journal of the American Medical Informatics Association*. 18, 1 (2011), 104–105.

[58] Matwin, S. and Sazonova, V. 2012. Direct comparison between support vector machine and multinomial naive Bayes algorithms for medical abstract classification. *Journal of the American Medical Informatics Association*. 19, 5 (2012), 917–917.

[59] Miller, J. 2005. Replicating software engineering experiments: A poisoned chalice or the Holy Grail. *Information and Software Technology*. 47, 4 (2005), 233–244.

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

[61] Molléri, J.S. and Benitti, F.B.V. 2015. SESRA: a web-based automated tool to support the systematic literature review process. *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering* (2015), 24.

[62] O Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M. and Ananiadou, S. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*. 4, 1 (2015), 1–22.

[63] Olorisade, B.K. 2012. Informal Aggregation Technique for Software Engineering Experiments. *IJCSI International Journal of Computer Science Issues*. 9, (2012), 199–204.

[64] Olorisade, B.K., Vegas, S. and Juristo, N. 2013. Determining the effectiveness of three software evaluation techniques through informal aggregation. *Information and Software Technology*. 55, 9 (2013), 1590–1601.

[65] Paulovich, F.V. and Minghim, R. 2006. Text Map Explorer: A tool to create and explore document maps. *Proceedings of the International Conference on Information Visualisation*. (2006), 245–251.

[66] Rathbone, J., Hoffmann, T. and Glasziou, P. 2015. Faster title and abstract screening? Evaluating Abstrackr, a semi-

automated online screening program for systematic reviewers. *Systematic Reviews*. 4, 1 (2015), 80.

[67] Razavi, A.H., Matwin, S., Inkpen, D. and Kouznetsov, A. 2009. Parameterized contrast in second order soft co-occurrences: A novel text representation technique in text mining and knowledge extraction. *ICDM Workshops 2009 - IEEE International Conference on Data Mining* (2009), 471–476.

[68] Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. *ACM computing surveys (CSUR)*. 34, 1 (2002), 1–47.

[69] Shemilt, I., Simon, A., Hollands, G.J., Marteau, T.M., Ogilvie, D., O'Mara-Eves, A., Kelly, M.P. and Thomas, J. 2013. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods*. January (2013), n/a–n/a.

[70] Shull, F., Mendoncça, M.G., Basili, V., Carver, J., Maldonado, J.C., Fabbri, S., Travassos, G.H. and Ferreira, M.C. 2004. Knowledge-Sharing Issues in Experimental Software Engineering. *Empirical Software Engineering*. 9, (2004), 111–137.

[71] Shull, F.J., Carver, J.C., Vegas, S. and Juristo, N. 2008. The role of replications in empirical software engineering. *Empirical Software Engineering*. 13, 2 (2008), 211–218.

[72] Sun, Y., Yang, Y., Zhang, H., Zhang, W. and Wang, Q. 2012. Towards evidence-based ontology for supporting Systematic Literature Review. *16th International Conference on Evaluation & Assessment in Software Engineering (EASE 2012),* (2012), 171–175.

[73] Thomas, J. and O'Mara-Eves, A. 2011. How can we find relevant research more quickly ? *NCRM Newsletter: MethodsNews*. (2011).

[74] Tomassetti, F., Rizzo, G., Vetro, A., Ardito, L., Torchiano, M. and Morisio, M. 2011. Linked data approach for selection process automation in systematic reviews. *15th Annual Conference on Evaluation and Assessment in Software Engineering (EASE '11),* (2011), 31–35.

[75] Vegas, S., Juristo, N., Moreno, A., Solari, M. and Letelier, P. 2006. Analysis of the influence of communication between researchers on experiment replication. *International Symposium on Empirical Software Engineering* (2006), 28.

[76] Wallace, B. and Small, K. 2011. Who should label what? Instance allocation in multiple expert active learning. *Proceedings of the SDM* (2011), 176–187.

[77] Wallace, B.C., Small, K., Brodley, C.E., Lau, J., Schmid, C.H., Bertram, L., Lill, C.M., Cohen, J.T. and Trikalinos, T.A. 2012. Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining. *Genetics in Medicine*. 14, 7 (2012), 663–669.

[78] Wallace, B.C., Small, K., Brodley, C.E., Lau, J. and Trikalinos, T.A. 2012. Deploying an interactive machine learning system in an evidence-based practice center. *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI '12* (2012), 819.

[79] Wallace, B.C., Small, K., Brodley, C.E., Lau, J. and Trikalinos, T.A. 2010. Modeling Annotation Time to Reduce Workload in Comparative Effectiveness Reviews Categories and Subject Descriptors Active Learning to Mitigate Workload. *Proceedings of the 1st ACM International Health Informatics Symposium. ACM,* (2010), 28–35.

[80] Wallace, B.C., Small, K., Brodley, C.E. and Trikalinos, T.A. 2010. Active Learning for Biomedical Citation Screening. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (2010), 173–182.

[81] Wallace, B.C., Trikalinos, T.A., Lau, J., Brodley, C. and Schmid, C.H. 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*. 11, (2010), 55.

[82] Yu, W., Clyne, M., Dolan, S.M., Yesupriya, A., Wulf, A., Liu, T., Khoury, M.J. and Gwinn, M. 2008. GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. *BMC bioinformatics*. 9, (2008), 205.