Reproducibility of Studies on Text Mining for Citation Screening in Systematic Reviews: Evaluation and Checklist

Babatunde Kazeem Olorisade[a,*], Pearl Brereton[a], Peter Andras[a]

[a]*School of Computing and Mathematics*
Keele University, Staffs. ST5 5BG, UK

Abstract

Context: Independent validation of published scientific results through study replication is a pre-condition for accepting the validity of such results. In computation research, full replication is often unrealistic for independent results validation, therefore, study reproduction has been justified as the minimum acceptable standard to evaluate the validity of scientific claims. The application of text mining techniques to citation screening in the context of systematic literature reviews is a relatively young and growing computational field with high relevance for software engineering, medical research and other fields. However, there is little work so far on reproduction studies in the field.

Objective: In this paper, we investigate the reproducibility of studies in this area based on information contained in published articles and we propose reporting guidelines that could improve reproducibility.

Methods: The study was approached in two ways. Initially we attempted to reproduce results from six studies, which were based on the same raw dataset. Then, based on this experience, we identified steps considered essential to successful reproduction of text mining experiments and characterized them to measure how reproducible is a study given the information provided on these steps. 33 articles were systematically assessed for reproducibility using this approach.

*Corresponding Author
Email addresses: b.k.olorisade@keele.ac.uk (Babatunde Kazeem Olorisade), o.p.brereton@keele.ac.uk (Pearl Brereton), p.andras@keele.ac.uk (Peter Andras)

Results: Our work revealed that it is currently difficult if not impossible to independently reproduce the results published in any of the studies investigated. The lack of information about the datasets used limits reproducibility of about 80% of the studies assessed. Also, information about the machine learning algorithms is inadequate in about 27% of the papers. On the plus side, the third party software tools used are mostly free and available.
Conclusions: The reproducibility potential of most of the studies can be significantly improved if more attention is paid to information provided on the datasets used, how they were partitioned and utilized, and how any randomization was controlled. We introduce a checklist of information that needs to be provided in order to ensure that a published study can be reproduced.

Keywords: Citation screening, systematic review, reproducibility, text mining, reproducible research

1. Introduction

A scientific claim cannot be considered credible until it can be independently verified [1–3]. Despite the few arguments against the reproduction of studies, the most notable being that it generates no new knowledge, the practice has been justified as the minimum acceptable standard to assess the validity of scientific claims in computational research [4] where full replications are often impractical [5].

Reproducibility, for the purpose of verification, understanding and consequently knowledge extension, is an essential requirement of all scientific studies — theoretical or experimental [6]. A theoretical study requires only mental understanding, pen and paper to reproduce and verify; while an experimental study requires similar laboratory settings and equipment to reproduce and verify [7]. However, the emergence of computational studies in the last few decades has put additional challenges on study reproduction. The independent researcher requires access to actual data, software and hardware specifications

for effective reproduction of computational studies [7, 8]. Two examples of these new challenges are the fact that:

(i) software modules are in continual development with possible alterations to internal implementation algorithms

(ii) datasets may be updated or moved without notice

These unpredictable circumstances necessitate the reporting of additional details that may facilitate future access to similar experimental materials for reproduction purposes. Consequently, published articles need to maintain a persistent link to all the digital materials of their experiment.

Study reproducibility or reproduction is the extent to which the results of a specific study can be independently reproduced based strictly on the published text, data, as well as experimental and analysis procedures [9]. Reproduction of experiments is essential in computational research for two main reasons. On the one hand, it supports the validation, verification and/or extension of computational results published in papers [10]. On the other hand, it is a precondition for accepting published claims as part of a body of knowledge [1]. However, articles are often published without the details — codes, datasets, experiment design parameters etc. — essential to the reproduction of experiments [5]. Study reproduction particularly requires access to all essential experimental elements in order that teams independent of the original research group can use them to verify the published results.

Text mining (TM), is generally the process of using the computer to automatically explore and analyze (unstructured) multiple textual sources to discover fresh information for further use [11, 12]. TheTM process can be viewed as comprising three major steps or activities: document (corpus) collection, text transformation and knowledge extraction.Each of these steps is informed by a major research area, namely, and respectively, Information Retrieval (IR), Information Extraction (IE) and Data Mining [13, 14].

TM activities start with the collection of documents relevant to the purpose of the TM. The document collection step involves the process of searching, lo-

cating, identifying and retrieving documents suspected of being relevant to the intended purpose. This step relies on technologies and techniques from the IR domain. IR research is primarily concerned with the development, optimization and delivery of techniques for searching, assessing, ranking and presenting information resources with respect to the users' information needs [15]. At present, the complexities associated with TM (such as lack of structure in the text and the dynamic nature of the databases) mean that research activities rely mainly on the use of standard corpuses and simple retrieval of the documents from their known location without the need to use further more sophisticated IR techniques. This is the case for the studies investigated in this work.

The text transformation step entails cleaning the text and converting it into a more structured format by means of some natural-language information extraction techniques. IE filters structure data from unstructured text by identifying references to named entities as well as relationships between such entities [12]. Typically, current works in TM favour the vector space model using bag-of-words for representation of the text without requiring the use of more sophisticated IE techniques [12].

The knowledge extraction step utilizes data mining algorithms and techniques to build models that can learn the data pattern and predict new knowledge from new similar dataset through regression, classification or clustering. Data mining is the process of discovering non-trivial knowledge or patterns from databases using machine learning algorithms [14].

Systematic Review (SR) is a literature review approach used in software engineering and other disciplines (particularly medicine and education) [16, 17]. It provides a rigorous, dependable and auditable review methodology with the main goal of building an impartial and complete synthesis of available evidence on a specific topic, based upon which decisions can be made and conclusions drawn. The SR process is divided into three major phases: planning, execution and documentation. These phases are further divided into stages [16, 18, 19]. 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 TM to automate the citation screening ((CS), also called study selection) stage [20].

The application of TM techniques to support the citation screening stage of SRs (e.g. in evidence-based software engineering or medical research) is an emerging research field with the first reported publication on the subject in 2005 [21] and a systematic review of 44 papers published in 2015 [22]. This field can, therefore, benefit from tools and techniques for improved experiment reproduction to verify published results, establish efficiency, maturity, applicability of proposed methods and techniques and advance findings [23]. This has even become imperative given the fact that funding agencies and publishers of data driven studies have now begun to stipulate that researchers make digital components of their research available.

In this study, we address reproduction issues in the field by assessing how easy it is to reproduce the results published in 33 papers. These 33 papers, which report 33 studies, were reviewed by Olorisade et al. [20] and are a subset of the 44 papers reviewed by O'Mara et al. [22]. As far as we know, there is no published work yet addressing experiment reproduction issues in the field.

The assessment involves three steps: initially, we tried to reproduce six of the experiments that used the Drug Evaluation Review Program (DERP) dataset. Then, we used the experience from this to identify elements of the TM process critical to reproduction and finally, we undertook a systematic assessment of 33 published studies using a proposed set of essential elements in TM experiments. As a result of this work, we suggest a checklist that authors could use to ascertain whether their articles contain enough relevant details to enable reproduction of the research and reviewers could also use it to assess computational studies for compliance to reproduction requirements.

The rest of the paper is structured as follows: Section 2 — Background, presents a brief overview of related work on reproducibility of computational studies, the studies that apply TM techniques to citation screening using the same dataset and existing work on assessing the reproducibility of data driven studies in software engineering. The details of the reproduction analysis and

the systematic assessment are presented in Section 3 — Methodology. Section 4 — Results, presents the outcomes of this work, while Section 5 addresses the threats to validity. Finally, Section 6 proposes a reproducibility checklist and summarizes the conclusions of this research.

2. Background

2.1. Reproduction of Computational Studies

The issue surrounding the ability of independent researchers to reproduce computational studies has been identified in the past few decades and researchers have made several proposals about how to make computational studies reproducible. [2, 24] advised cultivating reproducibility into a habit and everyday research culture before its effect can be successfully noticed in publications.

Explicit and unambiguous description of process and results is the first step towards ensuring independent researchers can clearly understand a study to the level that it can be reproduced by them [2]. Undocumented implicit knowledge is often the main impediment to the implementation of proposed algorithms and models [25].

Technology can support reproducibility [9]. For example, it has been suggested that researchers should utilize whenever they can, available libraries and packages that are easily accessible to the public, are robust and are continually maintained [2, 24]. Cross platform software should be chosen where possible for experiment purposes [24, 25]. However, it is practically impossible to capture all the decisions and situations during a computational study, so employing an automatic means of storing the details of every decision, process and result is encouraged [4, 5, 24]. GitHub and other similar version control applications can aid capturing of the different stages and changes in experiments as well as providing long term storage and access to the digital artefacts [4, 5, 24].

Public repositories and publishers are helping to ensure digital components of publications are available to readers [2, 4]; however, this does not guarantee that a study will be reproducible. Understanding the provided files is key to making

independent (active) use of them but data files are still formatted haphazardly; partially or insufficiently annotated [26, 27]; codes are poorly commented while graphs and charts are sparsely annotated amongst other issues [28]. Though, the digital components storage provision facilities is a step in the right direction.

In order to ensure reproducibility, comparability and generalizability of studies, the IR community have dedicated considerable efforts (notably) to the standardization of data formats to facilitate uniform storage, access and exchange of data, as well as the creation of common evaluation methods for techniques [29, 30]. Notable initiatives that have pushed research achievements in IR are TREC[1], CLEF[2] and NTCIR[3] [31–33]. These efforts are inherently beneficial to and directly utilized in TM research. Some of the experimental collections used in TM are part of the experimental collections from real domains of interest like medicine, made available through the efforts of IR research at ensuring reproducibility and comparability in the field [30]. An example is the TREC collection, one of which is the corpus used in this work and in studies reviewed in this work. The evaluation metrics proposed and used in IR research are also beneficial to and utilized by TM studies [34].

The Big Data to Knowledge (BD2K), trans-National Institute of Health (NIH) initiative has been established to facilitate the standardization, discovery and reuse of digital assets in biomedical research through innovative approaches and tools so that machines without human intervention can automatically access and (re)use study data. This initiative led to the agreement on the Findable, Accessible, Interoperable and Reusable (FAIR) principles that should guide such big data driven research. The guidelines for these principles are described in [28] and a sample tool implementation is provided in [35].

These principles along with other aims of the BD2K initiative[4] support reproducibility of experiments by facilitating digital assets discovery (open knowl-

[1] http://trec.nist.gov/
[2] http://www.clef-initiative.eu/
[3] http://research.nii.ac.jp/ntcir
[4] https://commonfund.nih.gov/bd2k

edge) for verification, knowledge advancement and community wide research engagement. The realization of the BD2K objectives will not only be useful to biomedical research but also for the general science communities' effort on reproducibility of scientific research.

Data format is also key to access and reuse. Researchers should attempt to store their data in common formats [25] like the comma separated values (csv) or similar formats. This way, other researchers will find it easier to retrieve and manipulate the data.

Prior to publication, it has been suggested that researchers should conduct a reproducibility check by asking colleagues not involved in the research to attempt to reproduce their studies based strictly on the information contained in their manuscript. This way, it will be possible to anticipate areas of ambiguities and insufficient information [2, 5, 26].

Though reproducibility is not a license to a study's correctness, validity or quality, it is however, a precursor to these qualities as utilizing these principles will not only aid the reproducibility of studies but also further the development of the means to ensure it.

2.2. Replication/Reproduction in CS Automation Studies

The earliest work we found on applying TM to CS automation is the work of Aphinyanagons et al. [21] published in 2005. Several works have since been and are still being published in the field. Most of the published studies share common experimental components in terms of datasets and machine learning algorithms. Olorisade et al. [20] conducted a critical analysis of 44 studies and found Support Vector Machine (SVM) and the ensemble method to be the most used among the studies, 31% and 22% respectively.

In terms of datasets, we found that the DERP review topics data have been used in 13 studies [20]. Table 1 shows the usage pattern of datasets common in some studies. Despite the extensive use of the DERP dataset in the publications (most especially the 15 review topics first used in [36]), there have been very few attempts at independent replication of existing studies. Thus, there seems

to be little ground for comparability of results as the study settings vary a lot — except in cases where the same research team replicate their own work.

The researchers in the field have demonstrated an awareness of the need for reproducible research. This is evidenced in [36], where the authors published the intermediate output from each step of the study and provided a link to supplementary materials including the datasets[5]. This level of detail has not been found in subsequent studies. Matwin et al. conducted a replication of [36] and both groups were able to compare the performance of their SVM and Multinomial Nave Bayes models [59–61]. Khabsa et al. [47] also compared their results to those reported in [36, 37, 45]. Several other studies that used the same 15 review topics as [36] have found the details useful by following at least the same preprocessing steps. [48, 62–64] are among the studies that have provided access to the supplementary materials of their studies.

Despite these efforts, the field has not witnessed any significant replication efforts. Overall, the field has a few clusters of studies sharing common datasets, machine learning algorithm and performance assessment metrics but within the clusters, the same research teams have mostly conducted the studies. Thus, there is a need for more independent replication of some of these studies to further validate the published results and consequently extend the findings.

2.3. Reproducibility Assessment

Gonzalez-Barahoma and Robles identified a set of information elements required to support the reproducibility of software engineering studies based on data [65]. The elements are: data source, retrieval methodology, raw dataset, study parameters, extraction methodology, processed data, analysis methodology and results dataset [65]. Their proposal built on the Knowledge Discovering in Databases (KDD) schema proposed by Fayyad et al.'s in [66] where data, selection, target data, preprocessing, preprocessed data, transformation, trans-

[5] `http://skynet.ohsu.edu/~cohenaa/systematic-drug-class-review-data.html` (The original link provided in their publication is now broken)

Table 1: Studies using common datasets and the algorithms

S/N	Dataset	Comment	Paper
1	DERP*	SVM	[37–44]
		FCNB	[45]
		EvoSVM, NB	[46]
		Random Forest	[47]
		Latent Dirichlet Allocation	[48]
		Perceptron	[36]
2	TrialStat SR	cNB	[49]
		SVM, NB	[50]
		Ensemble	[51–53]
3	Chronic Obstructive Pulmonary Disease (COPD)	cNB	[54–57]
		Ensemble	[57]
4	Proton beam	SVM	[55–58]
		Ensemble	[56, 57]
5	Micro nutrients	SVM	[55–57]
		Ensemble	[57]

* Some of the studies used fewer review set than others but they mostly share common 15 studies. The dataset is also referred to in some studies as the Text Retrieval Conference (TREC) data

Table 2: Values for the assessment of the information elements attributes

S/N	Attribute	Values
1	Identification	Complete (Classical), Partial, No, N/A
2	Description	Complete (Textual), Partial, No, N/A
3	Availability	Private, Public (Free), No, N/A
4	Persistence	Likely, Unknown, N/A
5	Flexibility	Complete, Partial, No, N/A

formed data, data mining, patterns, interpretation/evaluation and knowledge were identified as the elements composing the KDD process.

In their study, Gonzalez-Barahoma and Robles [65] defined five attributes and some values that can be used to describe the information elements associated with computational studies like TM. The five attributes are:

(i) Identification (location): where the information element can be accessed e.g. web-link.

(ii) Description: level of published details provided about the information element including it's internal organization and structure, and its semantics.

(iii) Availability: a measure of the difficulty involved to currently access or acquire the information element.

(iv) Persistence: the possibility of the information element being available for future use.

(v) Flexibility: how adaptable is the information element to different formats and/or environments.

These attributes are assessed independently of each other based on the available information in a publication. The values that can be assigned to each attribute are listed in Table 2.

The interpretation of the values as used within this study is described in

section 3 (Methodology).

Robles et al. [65] also defined a set of six (summary) assessment tags (Table 3) that may be combined, as applicable, to summarize the strength or otherwise of the contribution of an element to the reproducibility of a study.

Table 3: Summary assessment tags for defined reproducibility elements

S/N	Tag	Meaning
1	U	Usable for reproduction
2	D	Usable for reproduction with some difficulty
3	N	Not usable for reproduction
4	+	Future availability is foreseeable
5	*	Flexible
6	-	Irrelevant

3. Methodology

In this study, we assess the reproducibility of studies that investigate the application of TM techniques to the automation of the CS stage in SR. Our aim is to assess how reproducible are existing studies about the use of TM techniques to automating CS in SR.

We approached the reproduction assessment of the studies as follows:

i Reproduction Analysis: to try to reproduce six studies that used the DERP dataset

ii Assessment framework definition: to formulate an assessment framework using experience from `(i)` and the literature

iii Reproducibility assessment: applying the assessment framework to measure the reproducibility of 33 studies

These steps are further discussed in the following subsections.

3.1. Reproduction Analysis

For the reproduction analysis, we selected six studies [36, 38, 39, 41, 44, 45] that were based on various topics in the DERP dataset, particularly the topics contained in the TREC 2004 Genomics Track corpus[6].

After searching the Internet, we were able to locate the repository for the raw dataset - TREC 2004. The raw dataset contains 4,367,228 articles, separated into a few files in eXtended Markup Language (XML) or Standard Generalized Markup Language (SGML) formats, of which the studies we were trying to reproduce used less than 19,000. In order to select the subset that was of interest to us, we had to study and harmonize information contained in over ten different files and then write a parser to retrieve the articles and portions of each article of interest to us. We used the studies' PubMed Identification (PMID) information from a file provided by Cohen et al.[5] to cross-reference the raw dataset and extract the required dataset. The original supplementary materials link provided in [10] did not work.

It was easier to replicate the text pre-processing steps reported in the studies. The pre-processing in this context involves:

- removing commonly used words (e.g. articles and prepositions) referred to as stopwords
- breaking the sentences into words or phrases known as features
- storing all tokens in a feature vector using the Bag-of-Words (BoW) approach
- representing or encoding the features in a numeric usually binary or frequency based - format
- appending special tags to features from the MeSH and publication type before the above preprocessing steps as was done in [36, 45].

[6] `http://skynet.ohsu.edu/trec-gen/data/2004/`

We followed the protocol provided in [36]. Accordingly, we distinguished three type of features from the corpus — title and abstract, the MeSH terms and publication type. The MeSH terms were appended with 'mh' while the publication type were appended with 'pt' to distinguish them from similar title and abstract terms. We appended these terms before removing stopwords. We used binary representation for the features. In binary representation, if a feature is present in a document, it is represented by 1 as the corresponding element of the feature-document matrix and by 0 otherwise.

We conducted feature selection, the process of selecting the most discriminative subset of all the features for use, according to the process implemented in [36] by selecting statistically significant features using χ^2. We used the RapidMiner data science platform[7] and the FSelector (version 0.21) package in R for feature selection. Such feature selection techniques are used here to reduce the dimensionality of the data representation vectors.

The authors of [36, 41, 45] did not provide the codes for the algorithms they proposed, therefore, we used the base algorithms in each case to see how close the results were.

We conducted experiments using the simple Perceptron and SVM algorithms in Python's 'sklearn' package [67] and the implementation of the 'votedperceptron' algorithm provided in Weka (with no weighting) [68], which is the algorithm that was modified in [36].

We stored the data of the different studies in order of the PMID in the file provided by Cohen et al.[5]. Supporting materials — codes and data files — to aid the reproduction of this experiment is hosted on github[8]. In our implementation of the algorithms the classifiers parameters were set as follows:

- SVM: C = 1.0 and class_weight = 'balanced', others are left at the default.
- Perceptron: penalty = 'l1', class_weight = 'balanced', shuffle = True,

[7] `https://rapidminer.com/`

[8] `https://github.com/raylite/reproducibility-data`

random_state = 0; other parameters are left at default.

In both cases, the sample weight for the negative to positive class was set at 1:4 during fitting. We chose this sample weighting following [36], which used the same weight for some of the studies reported there. Although the best performance for each of the fifteen studies is recorded at different weights in [36], the weighting of 1:4 showed a consistent acceptable performance comparable to the best cases for all studies (in some cases providing the best results) [36].

Model validation was fairly well reported. Cross validation was mostly used. This might be due to the small size of the datasets. We used the 'Stratified-KFold' method also from python's sklearn package to divide the datasets into training and test data for the 5x2 cross validation. The method ensures negative:positive class ratio in the training and test data comparable to the original dataset. The random_state parameter was set to '67' in both cases; random_state is the seed of the pseudo random number generator to use when shuffling the data. The shuffling ensures that each run of the algorithm produces different results. However, if the random_state is set to a value, this value can be used to repeat a previous result provided other factors are kept constant.

The average precision, recall and F1 scores were calculated using the precision_score, recal_score and f1_score methods in sklearn. The average parameter in these methods was set to binary since this is a binary classification. A brief definition of precision, recall and the f1 score follows below:

- Recall is the fraction of the total number of positive examples in the whole corpus that is correctly classified [69].

$$recall = \frac{tp}{tp + fn}$$

- Precision is the ratio of actual positive examples and the total number of the predicted positives [69].

$$recall = \frac{tp}{tp + fp}$$

- F1 score is the weighted harmonic mean of the recall and the precision [69].

$$F1 = 2.\frac{precision.recall}{precision + recall}$$

where,

tp →true positive fp →false positive

tn →true negative fn →false negative

3.2. Assessment framework

In order to systematically assess the studies, we follow the approach proposed in [65] by identifying information elements (see sub-section 4.2) that supports reproducibility within the context of TM. The relationship between the identified elements in the TM process is depicted in fig. 1. We adopted the KDD process proposed by Fayyad et al. and its adaptation proposed by Robles et al. [65, 66], but further adapted to the TM context. We added the data source element as suggested by Robles to capture data retrieval (see fig. 1). This is essential as all the studies use existing data from some organizations and rarely make their particular experiment data available for reuse. The interpretation/evaluation step is replaced with model assessment.

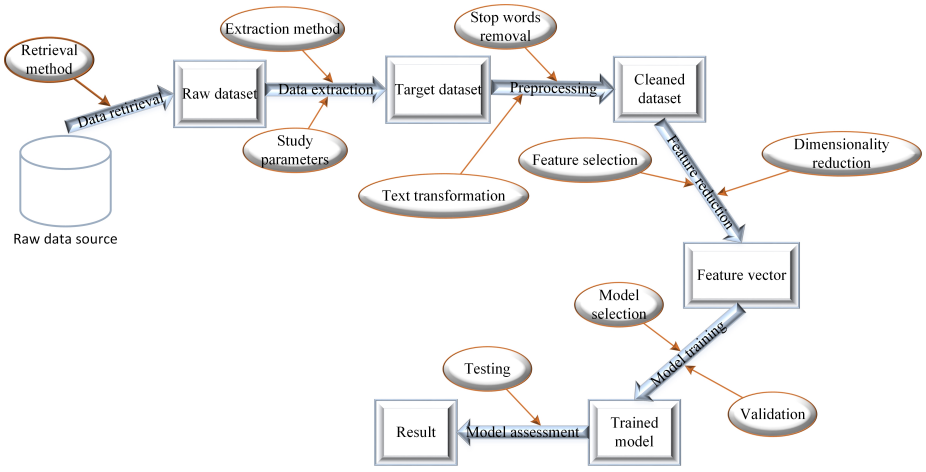

Figure 1: Basic text mining process

These information elements are assessed under the five attributes presented in sub-section 2.3 and described with a defined set of values (Table 2).The interpretation of the values (Table 2) depends on the attribute-element context. The meanings as used in this study are provided below:

- Complete: this generally implies that basic information needed to locate or identify the element in question is provided. For example, in the case of raw datasets, this may imply the general name of a particular dataset with the associated link from where it can be retrieved. Notable variations are:
  - Classical: the term classical is sometimes used (instead of complete) under identification, if one of the traditional machine learning algorithms is used out of the box with no (significant) alteration. This term is preferred to indicate that insufficient description may be tolerated in such cases.
  - Textual: Textual is used to indicate a new method, tool or algorithm proposed by the researchers and described only with text in the publication with neither source code nor executable file provided.
- Partial: This value is used to indicate situations where the information provided about an element is too general or insufficient. For example, a dataset (source) named with no link information to its exact webpage but rather to the index page of the provider where the researcher will be left to try and navigate to the desired resource.
- No: No implies complete absence of the attribute.
- N/A (Not Applicable): This implies the attribute is not applicable to the element in question. For example, for a study that did not make use of any of the information elements described above, the corresponding entries will be N/A.
- Likely: This value applies to the persistence attribute if there is a possibility that a relevant element is likely to be available for future access.

- Private/Public/Free: The term private is used to indicate elements, in this case data or tools, located but inaccessible due to extra constraints like membership, application, subscription etc. imposed before access may be granted. Public on the other hand means that the dataset (raw or processed) is provided for public use. Free is used in the case of a tool used that is available for free download.

- Unknown: We use this term when it was not easy to determine whether or not a relevant element will be available for future access.

See Appendix A for further illustration.

Not all attributes are defined for every element. Table 4 shows an example of the set of attributes applicable to each element.

Table 4: Example of attributes defined for each element type

	Data sources	Datasets	Technique	Parameters	Tools/Algorithms
Identification	✓	✓	✓	✓	✓
Description	✓	✓	✓	✓	✓
Availability	✓	✓			✓
Persistence	✓	✓			✓
Flexibility		✓			✓

3.3. Reproducibility Assessment

After the attempt to reproduce the experiments reported in the six papers, we were in a better position to evaluate how easy it might be to reproduce the rest of the studies and to identify what factors determine the extent of reproducibility. In each study, we identify the different information elements (depicted in Fig. 1 and explained in sub-section 4.2) and use the assessment

attributes and their associated metrics defined by Gonazalez and Robles highlighted in section 2.3 to indicate the presence or otherwise of each of them. If present, the appropriate metrics value is selected to indicate the extent of usefulness of the provided information.

4. Results

In this section we present the results of the three basic activities of this study described in Section 3 – Methodology.

4.1. Reproduction analysis

Here we report the outcomes from the reproduction analysis of the six studies (described in sub-section 3.1). The difficulties encountered are very similar across all of the studies. Nevertheless, when there is need to show concrete example(s), we refer to [36], which provides the most detailed step by step measurable outputs.

Generally, it was difficult to acquire the raw/cleaned dataset used in the considered studies. Often the referenced web links were either broken or pointed to the index page of the hosting institution. In most cases, however, there was no link even to the location of the raw dataset. The papers contain sufficient information that identifies the classification algorithm used but the provided information was insufficient to reproduce the classification results. Beyond the standard algorithms, all the studies attempted something new to try to optimize the performance of the traditional algorithms and mitigate the effect of any known TM problems like class imbalance. However, they provided only textual descriptions of the changes or at best an algorithm of the changes but not the code that was used.

Starting from the dataset, analysis of the details available in each of the studies are as below:

- The link to supplementary materials provided in [36] is broken. We were able to locate the new link, but the cleaned extracted dataset is not provided. The site contains a web link to the TREC 2004 Genomics Track

webpage but not directly to where the raw data was supposed to be located; we have provided the direct link[6]. They also provided a file with the PMIDs for the dataset they used.

- [45] referenced the information provided in [36].
- Dataset source or location was not provided in [38].
- Data source or body providing the data was named in [39, 41, 44] but neither a link nor retrieval information was provided for the dataset used.

Though we were eventually able to locate the raw DERP data source, we were unable to extract the exact full dataset. We used the PMID file provided by Cohen et al. [36] and retrieved 18,431 from the directories: "2004_TREC_ASCII MEDLINE_1" and "2004_TREC_ASCII_MEDLINE_2". Out of the 18, 733 data items of the 15 review topics used in [36, 38, 45], however, we could not locate the 302 missing items (see Table 5. for the number of studies retrieved for each topic). Thus, our reproduction analysis relied on an incomplete dataset, which was a significant setback from the perspective of reproducibility. In order to circumvent this problem, we emailed the corresponding author of [36] requesting the extracted dataset used in their experiment and stated our mission but got no response. [39, 41, 44] used part or all of this dataset and also additional data.

The information provided about pre-processing — data cleaning, feature representation and selection — was mostly useful for reproduction across the papers. Only the feature representation used was reported in [39]. There was no explicit explanation of the representation.

In [36], the paper described how they selected statistically significant features using χ^2 with 0.05 α level, thus it was easy to compare results. The two applications we used agree on more than the top 50% of the results and above 80% in total for selected features. Despite this, we were not able to produce the exact number of features for a 0.05 confidence interval using the χ^2 method. This might have been because we did not have the complete dataset in the first

Table 5: Retrieved corpus size(s) and number of features significant for each study (Cohen et al.'s appears in italics)

Review topics	Corpus	χ^2 top features	MeSH features	PubType features
ACEInhibitors	2498	242	54	7
	2544	*210*	*40*	*5*
ADHD	845	115	39	0
	851	*80*	*24*	*0*
Antihistamines	308	31	10	1
	310	*29*	*9*	*0*
AtypicalAntipsychotics	1115	173	44	7
	1120	*302*	*71*	*8*
BetaBlockers	2043	129	26	3
	2072	*194*	*42*	*5*
CalciumChannelBlockers	1190	166	43	4
	1218	*329*	*77*	*5*
Estrogens	362	102	26	4
	368	*233*	*44*	*5*
NSAIDs	389	146	39	5
	393	*242*	*51*	*5*
Opioids	1883	78	25	0
	1915	*55*	*14*	*0*
OralHypoglycemics	493	97	22	3
	503	*234*	*55*	*4*
ProtonPumpInhibitors	1314	165	40	4
	1333	*206*	*54*	*6*
SkeletalMuscleRelaxants	1610	67	14	4
	1643	*11*	*2*	*2*
Statins	3402	173	39	5
	3465	*467*	*87*	*6*
Triptans	657	226	42	6
	675	*121*	*22*	*3*
Triptans	322	21137	37	6
	327	*215*	*45*	*5*

place. Another possibility is that there may be some fine–tuning not reported in the paper because the discrepancy in our number of features and theirs is too wide in some cases. The results of our data retrieval and feature selection compared to [36] (in italics) are presented in Table 5.

The 5x2 cross validation average results for precision, recall and harmonic mean (F1) are presented in Table 6, alongside an extract from Cohen et al.'s results [36] in italics. Our 'votedperceptron' precision values are better than Cohen et. al.'s but the recall and F1 score are worse. The lower recall in this case accounts for the higher precision values, since there is always a trade-off between recall and precision. But the simple perceptron and SVM show comparable and sometimes lower recall with higher precision performance compared to Cohen et al.'s. This shows that the results of the studies could be reproduced only if the authors were to provide sufficient information on experimental procedure and data. If we have access to the full dataset, it might still be impossible for us to get the exact classification outcomes given that randomization is usually involved in the procedures of text classification algorithms and none of the papers provide access to the data partition or indices they used for the training and test/validation sets. They only provide proportion information about training and test sets (i.e. what percentage of the data was used for these purposes). The seed value used (if any), would have been sufficient to reproduce any randomised step but that was not provided either. Overall, based on our reproduction analysis experience, we conclude that it is difficult to reproduce the studies. This difficulty could have been significantly reduced if the studies had made available the datasets they used, the seed value for each randomisation steps or the data partition or indices for the training and test/validation sets, and the implementation details of any algorithm or method used.

4.2. Assessment framework definition

Following our attempt to reproduce the results of six studies in this paper, we identified the following information elements required to support the reproducibility of TM application in the context of citation screening:

Table 6: 5X2 folds cross validation results based on top features as used in [28]

Review topics	Method	Precision	Recall	F1
ACEInhibitors	Cohen	*0.0387*	*0.9561*	*0.0745*
	Votedperceptron	0.414	0.101	0.16
	Simple perceptron	0.11	0.86	0.19
	SVM	0.15	0.75	0.25
ADHD	Cohen	*0.0945*	*0.9200*	*0.1713*
	Votedperceptron	0.53	0.514	0.521
	Simple perceptron	0.35	0.95	0.50
	SVM	0.46	0.94	0.62
Antihistamines	Cohen	*0.0502*	*0.8500*	*0.0948*
	Votedperceptron	0.571	0.467	0.517
	Simple perceptron	0.40	0.83	0.53
	SVM	0.40	0.98	0.57
AtypicalAntipsychotics	Cohen	*0.1534*	*0.9493*	*0.2642*
	Votedperceptron	0.582	0.533	0.556
	Simple perceptron	0.42	0.80	0.53
	SVM	0.33	1.00	0.49
BetaBlockers	Cohen	*0.0334*	*0.9286*	*0.0644*
	Votedperceptron	0.459	0.201	0.279
	Simple perceptron	0.19	0.85	0.31
	SVM	0.18	0.97	0.30
CalciumChannelBlockers	Cohen	*0.0952*	*0.9460*	*0.1730*
	Votedperceptron	0.581	0.447	0.503
	Simple perceptron	0.38	0.78	0.49
	SVM	0.41	0.97	0.26
Estrogens	Cohen	*0.2252*	*0.9725*	*0.4044*
	Votedperceptron	0.645	0.440	0.519
	Simple perceptron	0.32	0.83	0.44
	SVM	0.38	0.96	0.54
NSAIDs	Cohen	*0.2631*	*0.9317*	*0.4103*
	Votedperceptron	0.651	0.568	0.603
	Simple perceptron	0.36	0.95	0.51
	SVM	0.44	0.92	0.59
Opioids	Cohen	*0.0092*	*0.9467*	*0.0182*
	Votedperceptron	0.359	0.068	0.114
	Simple perceptron	0.04	0.84	0.07
	SVM	0.08	0.56	0.14
OralHypoglycemics	Cohen	*0.4004*	*0.9471*	*0.4561*
	Votedperceptron	0.35	0.86	0.49
	Simple perceptron	0.67	0.75	0.68
	SVM	0.28	1.00	0.44
ProtonPumpInhibitors	Cohen	*0.0602*	*0.9373*	*0.1132*
	Votedperceptron	0.519	0.301	0.380
	Simple perceptron	0.26	0.80	0.38
	SVM	0.24	0.93	0.38
SkeletalMuscleRelaxants	Cohen	*0.0055*	*1.0000*	*0.0109*
	Votedperceptron	0.428	0.067	0.120
	Simple perceptron	0.03	0.94	0.05
	SVM	0.04	0.67	0.08
Statins	Cohen	*0.0311*	*0.9647*	*0.0603*
	Votedperceptron	0.272	0.039	0.070
	Simple perceptron	0.07	0.87	0.12
	SVM	0.11	0.69	0.19
Triptans	Cohen	*0.0365*	*0.9583*	*0.0703*
	Votedperceptron	0.647	0.634	0.641
	Simple perceptron	0.45	0.92	0.82
	SVM	0.48	0.93	0.63
UrinaryIncontinence	Cohen	*0.1559*	*0.9850*	*0.2691*
	Votedperceptron	0.473	0.465	0.465
	Simple perceptron	0.33	0.84	0.46
	SVM	0.26	0.97	0.41

(i) Data source: The actual location of the raw dataset — direct webpage.

(ii) Raw data: The whole of the dataset retrievable from (i). Necessary information may include the description of the internal structure of the dataset, the retrieval method, the file format(s) etc.

(iii) Dataset: The focused dataset used in a particular TM experiment which may be the whole of (ii) or a subset. Information required may involve any new location of the extracted dataset, the extraction technique and the parts extracted.

(iv) Pre-processing: This involves preprocessing steps of tokenization and noise removal from the resulting dataset.

(v) Feature representation: The method used for numerical encoding of the text tokens.

(vi) Feature Selection: The feature selection/reduction approach used with sufficient details.

(vii) Dimensionality reduction: Any other method used to further reduce the dimensionality of the feature vector beside feature selection.

(viii) Data partitions: Partitions (or indices) of the data used for the different classification operations — training, testing and or validation or seed value used to control randomised partitioning.

(ix) Modelling: Details of the machine learning algorithm used for mining the text, seed values for randomisation control, algorithm parameters and code or executable file for newly proposed algorithms.

(x) Model assessment: The testing or validation approach used.

(xi) Third party framework: Available machine learning software or packages used during the experiments.

(xii) Custom method: This refers to algorithms or techniques proposed by the authors in a study.

4.3. Reproducibility Assessment

Based on the information elements, attributes, metrics and tags defined in sub-sections 4.2 and 2.3, we assessed the reproducibility of 33 studies on the

application of TM to citation screening in systematic reviews. A typical detailed assessment of a study is shown in Table 7 while the overall assessment is presented in Table 8. The issues relating to data sources and datasets pose a key challenge to reproduction as information found in 28 (89%) of the papers (in both elements) are only useful with some difficulty while four (12%) were found to not have useful information about the data source and six (18%) about the dataset. 13 (39%), 16 (48%) and 11 (33%) of the papers respectively provided pre-processing, feature selection and dimensionality reduction information that is fully useful to reproduction; an additional six (18%), eight (24%) and four (12%) respectively with some difficulty. This leaves an average of five (15%) with either irrelevant or not useful information. Pre-processing and feature representation recorded values higher than 30% on no useful information mainly because the authors might assume implicit understanding thereby failing to mention what steps were specifically taken in data cleaning e.g. were stopwords removed? This information is necessary because there have been situations where experiments were conducted with stopwords. In the case of data split, we found only five (15%) papers providing information that may be useful for reproduction. The information about the machine learning algorithms can be used for reproduction in nine papers and with difficulty in another 19 (57%). However, information provided on custom (proposed) methods in three papers were found to be useful, 16 with difficulty while 13 (39%) have no provision for this element.

Validation and testing information were found useful in 13 (39%) of the papers and in 12 (36%) were useful for reproduction with some level of difficulty. Finally, all third party tools or frameworks used in the studies were found to be free and accessible. The information provided on them was sufficient to locate the tools.

Table 7: A typical assessment output of a study (see footnote for abbreviations in column 1)

	Identification	Description	Availability	Persistence	Flexibility	Assessment
DS	Partial	No	Private	Likely	N/A	D+
Dataset	No	No	Unknown	Unknown	No	N
PP	Classical	Complete	N/A	N/A	N/A	U
FS	Classical	Complete	N/A	N/A	N/A	U
DR	Classical	Complete	N/A	N/A	N/A	U
Split	No	Partial	N/A	N/A	N/A	D
Technique	Classical	Partial	N/A	N/A	N/A	D
Testing	Complete	Partial	N/A	N/A	N/A	D
TPF	Complete	Complete	Free	Likely	No	U+
CM	N/A	N/A	N/A	N/A	N/A	—

Note: DS – Data source; PP – Pre-processing; FS – Feature selection; DR – Dimensionality reduction; Split – dataset partition; Technique – Modelling technique/algorithm used; Testing – testing or cross validation technique; TPF – Third party framework and CM – Custom method.

5. Validity Threats

The assessment presented in this study is based mainly on our subjective interpretation of the content of the papers. The number of studies chosen for the reproduction analysis is quite small and thus the results might not be representative of all of the studies in the field. The studies involved in the assessment are also quite limited and thus may not represent the whole research area though they were chosen from a systematic review published in 2015.

6. Discussion and Conclusions

For the reproduction analysis, we were unable to reproduce any of the results of the original studies because we could not retrieve the complete datasets and, for all six studies, critical data usage information was missing. In particular, more information was needed about how the dataset was partitioned and about the seed values used for randomization.

Table 8: Summary assessment output on the reproducibility assessment of the 33 studies

Paper	Data Source	Dataset	Pre-processing	Feature Representation	Dimensionality reduction	Dataset Split	ML Technique	Validation or Testing	Third Party Framework	Customized Tool/Method	Ref
P1	D+	N	U	U	U	D	D	D	U+	-	[46]
P2	D	D	U	U	U	N	U	D	U	D	[70]
P3	D+	D	U	U	U	N	U	U	U+	D	[71]
P4	D+	D+	U	U	U	N	U	U	U+	D	[42]
P5	D	D	U	U	D	N	D	D	-	D	[36]
P6	D+	D+*	U	U	U	N	D	D	-	D	[40]
P8	D	D+	D	U	U	N	U	U	U+	-	[37]
P9	D	D+	D	U	U	N	D	U	-	D	[41]
P10	D+	D+*	N	N	N	N	U	U	-	-	[39]
P11	D+	D+	N	N	D	N	U	U	-	-	[72]
P12	D	D+	N	U	N	N	U	U	-	-	[44]
P13	N	N	-	-	-	-	-	-	U	U	[73]
P14	N	N	-	-	-	-	-	-	U	-	[74]
P15	N	N	-	-	-	-	-	-	U	-	[75]
P18	D+	N	D	U	U	D	U	D	U+	D	[49]
P19	D	D	U	U	U	N	D	D	U+	D	[50]
P20	N	D	U	U	U	N	U	D	U+*	-	[76]
P21	D	D+	N	N	N	N	-	-	-	D	[77]
P22	D+	D+	U	D	N	N	D	D	-	-	[38]
P23	D	D	N	U	N	N	D	U	U+	N	[53]
P24	D	D	N	D	D	D	D	U	-	D	[52]
P26	D	N	-	-	-	-	-	-	U	-	[78]
P27	U	D+*	N	D	N	-	D	-	U	-	[43]
P29	D+	D	U	U	N	N	D	U	-	D	[45]
P30	D	D	D	D	N	D	D	D	U+*	D	[57]
P32	D+	D	N	N	N	N	D	D	-	-	[79]
P34	D+	D+	D	D	N	N	D	N	U	D	[80]
P37	D+	D+	N	D	N	N	D	D	-	D	[81]
P38	D+	D	U	U	N	N	D	D	-	-	[62]
P40	D	D	N	N	-	-	D	U	-	U	[58]
P41	D+	D	D	D	D	N	D	-	-	D	[55]
P42	D	D	U	D	-	-	D	U	-	U	[56]
P43	D+	D	U	U	U	N	D	U	-	D	[82]

Note: U = Usable for reproduction; D = Usable for reproduction with some difficulty; N = Not usable for reproduction; + = Future availability is foreseeable; * = Flexible; - = Irrelevant

Some of the papers assessed for reproducibility (e.g. P1 — P9, as shown in Table 8.) did exhibit some potential for reproducibility providing good accessibility to raw datasets and useful explanations of their preprocessing, feature representation and dimensionality reduction process. However, for many of the papers, information about dataset partitioning was inadequate.

In addition, access to and information about the dataset used and details about the algorithms used in the studies were insufficient for reproduction. In particular, information about parameters and new (proposed) algorithms was lacking.

Generally, the accessibility of third party tools was good although, of course, we cannot be sure about their persistence and flexibility.

As a result of our research, we propose a checklist (Table 9) which is based on the information elements we have identified. This can be used by authors reporting TM experiments for citation screening in systematic reviews or any text classification experiment to help improve reproducibility.

Reviewers may also use the checklist to assess the level of reproducibility of TM studies in the context of citation screening for systematic reviews. We expect that the checklist will continue to be evaluated and upgraded until its usefulness and completeness is confirmed by many researchers. The checklist is in partial compliance with the FAIR principle as described in [35]. The data source and storage details will ensure the data is Findable, while being hosted on the internet at a published address will ensure it is Accessible. Interoperability is still a challenge, given that the data is being stored in popular formats on general-purpose repositories making it usable by humans, but not automatically usable by machines. The information about data format and partitioning will facilitate the Reusability of the data.

The reproduction analysis and reproducibility assessment in this study reveal that the studies are hard to reproduce due to missing information regarding access to and availability of raw, target or processed datasets. Reproduction by independent research teams is possible but with different levels of difficulty specific to each study.

Table 9: Reproducibility enabling information checklist for text mining studies

Item No.	Information elements	Yes	No	N/A
1	Original location of the raw dataset			
2	Provided link to local copy of: a. Raw dataset b. Target dataset c. Cleaned dataset			
3	Described the internal structure of: a. Raw dataset b. Target dataset c. Cleaned dataset			
4	Data retrieval method details			
5	Data extraction method described			
6	Pre-processing details			
7	Feature representation technique			
8	Feature selection technique			
9	Dimensionality reduction technique			
10	Final feature vector download link			
11	Training algorithm			
12	Custom algorithm a. Text b. Code c. Algorithm d. Executable file			
14	Model assessment method			
15	Detailed model assessment result			
16	Necessary seed values provided			
17	Training/test data partition available or indices provided a. Link to data partitions provided b. (link to) Indices provided c. Seed value provided			
18	Provide name and version number of third party or external software package used			

Studies in this field need to be reported with more information than is currently the practice, to aid independent reproduction of the studies. One possibility would be to create a common repository where research results can be stored along with associated datasets, partition information and process details [83]. This would ensure persistence and availability of datasets, as well as providing additional experiment information not included in publications. In fact, we advise making available the full code used during experiments. 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. Such communication has been established to help better replication [84, 85].

Data and process descriptions need to be made publicly available in order to support study reproduction and consequently enhance external validation and maturity chances of claims and discoveries.It will also help improve the availability of evidence about the effectiveness of the methods that have been proposed for the application of TM techniques to citation screening in SRs.

Reference

[1] D. L. Donoho, A. Maleki, I. U. Rahman, M. Shahram, V. Stodden, Reproducible research in computational harmonic analysis, Computing in Science & Engineering 11 (1) (2009) 8–18. doi:10.1126/science.1213847.

[2] J. P. Mesirov, Accessible reproducible research, Science 327 (5964) (2010) 415–416.

[3] G. King, Replication, replication, PS: Political Science & Politics 28 (03) (1995) 444–452.

[4] R. D. Peng, Reproducible research in computational science, Science 334 (6060) (2011) 1226–1227. doi:10.1126/science.1213847.Reproducible.

[5] G. K. Sandve, A. Nekrutenko, J. Taylor, E. Hovig, Ten simple rules for reproducible computational research, PLoS Comput Biol 9 (10) (2013) e1003285.

[6] D. Waltemath, R. Adams, F. T. Bergmann, M. Hucka, F. Kolpakov, A. K. Miller, I. I. Moraru, D. Nickerson, S. Sahle, J. L. Snoep, et al., Reproducible computational biology experiments with sed-ml-the simulation experiment description markup language, BMC systems biology 5 (1) (2011) 198. doi:10.1186/1752--0509--5--198.

[7] S. Fomel, G. Hennenfent, Reproducible computational experiments using scons, in: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, Vol. 4, IEEE, 2007, pp. IV–1257.

[8] S. Fomel, J. F. Claerbout, Reproducible research, Computing in Science & Engineering 11 (1) (2009) 5–7.

[9] T. Hothorn, L. Held, T. Friede, Biometrical journal and reproducible research, Biometrical Journal 51 (4) (2009) 553–555.

[10] L. Madeyski, B. Kitchenham, Would wider adoption of reproducible research be beneficial for empirical software engineering research?, Journal of Intelligent & Fuzzy Systems 32 (2) (2017) 1509–1521.

[11] V. Gupta, G. S. Lehal, et al., A survey of text mining techniques and applications, Journal of emerging technologies in web intelligence 1 (1) (2009) 60–76.

[12] R. J. Mooney, R. Bunescu, Mining knowledge from text using information extraction, ACM SIGKDD explorations newsletter 7 (1) (2005) 3–10.

[13] A. Hotho, A. Nürnberger, G. Paaß, A brief survey of text mining., in: Ldv Forum, Vol. 20, 2005, pp. 19–62.

[14] A.-H. Tan, et al., Text mining: The state of the art and the challenges, in: Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases, Vol. 8, sn, 1999, pp. 65–70.

[15] W. B. Croft, D. Metzler, T. Strohman, Search engines: Information retrieval in practice, Vol. 283, Addison-Wesley Reading, 2010.

[16] B. A. Kitchenham, T. Dyba, M. Jorgensen, Evidence-based software engineering, in: Proceedings of the 26th international conference on software engineering, IEEE Computer Society, 2004, pp. 273–281.

[17] J. P. Higgins, S. Green, Cochrane handbook for systematic reviews of interventions, Vol. 4, John Wiley & Sons, 2011.

[18] B. A. Kitchenham, D. Budgen, P. Brereton, Evidence-Based Software engineering and systematic reviews, Vol. 4, CRC Press, 2015.

[19] E. Hassler, J. C. Carver, N. A. Kraft, D. Hale, Outcomes of a community workshop to identify and rank barriers to the systematic literature review process, in: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, ACM, 2014, p. 31.

[20] B. K. Olorisade, E. de Quincey, P. Brereton, P. Andras, A critical analysis of studies that address the use of text mining for citation screening in systematic reviews, in: Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering, ACM, 2016, p. 14.

[21] Y. Aphinyanaphongs, I. Tsamardinos, A. Statnikov, D. Hardin, C. F. Aliferis, Text categorization models for high-quality article retrieval in internal medicine, J. Am. Med. Informatics Assoc. 12 (2) (2005) 207–216.

[22] A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, S. Ananiadou, Using text mining for study identification in systematic reviews: a systematic review of current approaches, Systematic reviews 4 (1) (2015) 5.

[23] J. Miller, Replicating software engineering experiments: a poisoned chalice or the holy grail, Information and Software Technology 47 (4) (2005) 233–244.

[24] A. Davison, Automated capture of experiment context for easier reproducibility in computational research, Computing in Science & Engineering 14 (4) (2012) 48–56.

[25] T. Crick, B. A. Hall, S. Ishtiaq, "Can I Implement Your Algorithm?": A Model for Reproducible Research Software, ArXiv e-prints`arXiv:1407.5981`.

[26] J. P. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, et al., Repeatability of published microarray gene expression analyses, Nature genetics 41 (2) (2009) 149–155.

[27] J. Rung, A. Brazma, Reuse of public genome-wide gene expression data, Nature Reviews Genetics 14 (2) (2013) 89–99.

[28] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3.

[29] D. C. Comeau, R. Islamaj Doğan, P. Ciccarese, K. B. Cohen, M. Krallinger, F. Leitner, Z. Lu, Y. Peng, F. Rinaldi, M. Torii, et al., Bioc: a minimalist approach to interoperability for biomedical text processing, Database 2013 (2013) bat064.

[30] J. Zobel, W. Webber, M. Sanderson, A. Moffat, Principles for robust evaluation infrastructure, in: Proceedings of the 2011 workshop on Data infrastructurEs for supporting information retrieval evaluation, ACM, 2011, pp. 3–6.

[31] M. Agosti, E. Di Buccio, N. Ferro, I. Masiero, S. Peruzzo, G. Silvello, Directions: Design and specification of an ir evaluation infrastructure., in: CLEF, Springer, 2012, pp. 88–99.

[32] J. Freire, N. Fuhr, A. Rauber, Reproducibility of data-oriented experiments in e-science (dagstuhl seminar 16041), in: Dagstuhl Reports, Vol. 6, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[33] N. Ferro, Reproducibility challenges in information retrieval evaluation, Journal of Data and Information Quality (JDIQ) 8 (2) (2017) 8.

[34] W. Hersh, Evaluation of biomedical text-mining systems: lessons learned from information retrieval, Briefings in bioinformatics 6 (4) (2005) 344–356.

[35] M. D. Wilkinson, R. Verborgh, L. O. B. da Silva Santos, T. Clark, M. A. Swertz, F. D. Kelpin, A. J. Gray, E. A. Schultes, E. M. van Mulligen, P. Ciccarese, et al., Interoperability and fairness through a novel combination of web technologies, PeerJ Computer Science 3 (2017) e110.

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

[37] A. M. Cohen, Optimizing feature representation for automated systematic review work prioritization, in: AMIA annual symposium proceedings, Vol. 2008, American Medical Informatics Association, 2008, p. 121.

[38] S. Kim, J. Choi, Improving the performance of text categorization models used for the selection of high quality articles, Healthcare informatics research 18 (1) (2012) 18–28.

[39] A. M. Cohen, K. Ambert, M. McDonagh, Studying the potential impact of automated document classification on scheduling a systematic review update, BMC medical informatics and decision making 12 (1) (2012) 33.

[40] A. M. Cohen, An effective general purpose approach for automated biomedical document classification, in: AMIA Annual Symposium Proceedings, Vol. 2006, American Medical Informatics Association, 2006, p. 161.

[41] A. M. Cohen, K. Ambert, M. McDonagh, Cross-topic learning for work prioritization in systematic review creation and update, Journal of the American Medical Informatics Association 16 (5) (2009) 690–704.

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

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

[44] A. M. Cohen, K. Ambert, M. McDonagh, A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review, in: AMIA Annual Symposium Proceedings, Vol. 2010, American Medical Informatics Association, 2010, p. 121.

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

[46] T. Bekhuis, D. Demner-Fushman, Towards automating the initial screening phase of a systematic review, Stud. Health Technol. Inform. 160 (PART 1) (2010) 146–150. doi:10.3233/978-1-60750-588-4-146.

[47] M. Khabsa, A. Elmagarmid, I. Ilyas, H. Hammady, M. Ouzzani, Learning to identify relevant studies for systematic reviews using random forest and external information, Machine Learning 102 (3) (2016) 465–482.

[48] B. E. Howard, J. Phillips, K. Miller, A. Tandon, D. Mav, M. R. Shah, S. Holmgren, K. E. Pelch, V. Walker, A. A. Rooney, et al., Swift-review: a text-mining workbench for systematic review, Systematic reviews 5 (1) (2016) 87.

[49] O. Frunza, D. Inkpen, S. Matwin, Building systematic reviews using automatic text classification techniques, in: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 2010, pp. 303–311.

[50] O. Frunza, D. Inkpen, S. Matwin, W. Klement, P. Oblenis, Exploiting the systematic review protocol for classification of medical abstracts, Artificial intelligence in medicine 51 (1) (2011) 17–25.

[51] A. H. Razavi, S. Matwin, D. Inkpen, A. Kouznetsov, Parameterized contrast in second order soft co-occurrences: A novel text representation technique in text mining and knowledge extraction, in: Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on, IEEE, 2009, pp. 471–476.

[52] A. Kouznetsov, S. Matwin, D. Inkpen, A. H. Razavi, O. Frunza, M. Sehatkar, L. Seaward, P. OBlenis, Classifying biomedical abstracts using committees of classifiers and collective ranking techniques, in: Canadian Conference on Artificial Intelligence, Springer, 2009, pp. 224–228.

[53] A. Kouznetsov, N. Japkowicz, Using classifier performance visualization to improve collective ranking techniques for biomedical abstracts classification, in: Canadian Conference on Artificial Intelligence, Springer, 2010, pp. 299–303.

[54] B. C. Wallace, K. Small, C. E. Brodley, T. A. Trikalinos, Who should label what? instance allocation in multiple expert active learning, in: Proceedings of the 2011 SIAM International Conference on Data Mining, SIAM, 2011, pp. 176–187.

[55] B. C. Wallace, K. Small, C. E. Brodley, T. A. Trikalinos, Active learning for biomedical citation screening, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2010, pp. 173–182.

[56] B. C. Wallace, T. A. Trikalinos, J. Lau, C. Brodley, C. H. Schmid, Semi-automated screening of biomedical citations for systematic reviews, BMC bioinformatics 11 (1) (2010) 55.

[57] M. Miwa, J. Thomas, A. OMara-Eves, S. Ananiadou, Reducing systematic review workload through certainty-based screening, Journal of biomedical informatics 51 (2014) 242–253.

[58] B. C. Wallace, K. Small, C. E. Brodley, J. Lau, T. A. Trikalinos, Modeling annotation time to reduce workload in comparative effectiveness reviews, in: Proceedings of the 1st ACM International Health Informatics Symposium, ACM, 2010, pp. 28–35.

[59] A. M. Cohen, 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 18 (1) (2011) 104–104.

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

[61] S. Matwin, V. Sazonova, 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.

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

[63] Z. Yu, N. A. Kraft, T. Menzies, How to read less: Better machine assisted reading methods for systematic literature reviews, arXiv preprint arXiv:1612.03224.

[64] Y. Mo, G. Kontonatsios, S. Ananiadou, Supporting systematic reviews using lda-based document representations, Systematic reviews 4 (1) (2015) 172.

[65] J. M. González-Barahona, G. Robles, On the reproducibility of empirical software engineering studies based on data retrieved from development repositories, Empirical Software Engineering 17 (1–2) (2012) 75–89. doi:10.1007/s10664--011--9181--9.

[66] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI magazine 17 (3) (1996) 37.

[67] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (Oct) (2011) 2825–2830.

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

[69] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management 45 (4) (2009) 427–437.

[70] T. Bekhuis, D. Demner-Fushman, Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers, Artificial intelligence in medicine 55 (3) (2012) 197–207.

[71] T. Bekhuis, E. Tseytlin, K. J. Mitchell, D. Demner-Fushman, Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence, PloS one 9 (1) (2014) e86277.

[72] S. R. Dalal, P. G. Shekelle, S. Hempel, S. J. Newberry, A. Motala, K. D. Shetty, A pilot study using machine learning and domain knowledge to facil-

itate comparative effectiveness review updating, Medical Decision Making 33 (3) (2013) 343–355.

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

[74] K. R. Felizardo, N. Salleh, R. M. Martins, E. Mendes, S. G. MacDonell, J. C. Maldonado, Using visual text mining to support the study selection activity in systematic literature reviews, in: Empirical Software Engineering and Measurement (ESEM), 2011 International Symposium on, IEEE, 2011, pp. 77–86.

[75] K. R. Felizardo, S. R. Souza, J. C. Maldonado, The use of visual text mining to support the study selection activity in systematic literature reviews: a replication study, in: Replication in empirical software engineering research (RESER), 2013 3rd International Workshop On, IEEE, 2013, pp. 91–100.

[76] J. G. Adeva, J. P. Atxa, M. U. Carrillo, E. A. Zengotitabengoa, Automatic text classification to support systematic reviews in medicine, Expert Systems with Applications 41 (4) (2014) 1498–1508.

[77] S. Jonnalagadda, D. Petitti, A new iterative method to reduce workload in systematic review process, International journal of computational biology and drug design 6 (1-2) (2013) 5–17.

[78] V. Malheiros, E. Hohn, R. Pinho, M. Mendonca, A visual text mining approach for systematic reviews, in: Empirical Software Engineering and Measurement, 2007. ESEM 2007. First International Symposium on, IEEE, 2007, pp. 245–254.

[79] I. Shemilt, A. Simon, G. J. Hollands, T. M. Marteau, D. Ogilvie, A. O'Mara-Eves, M. P. Kelly, J. Thomas, Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in

extremely large scoping reviews, Research Synthesis Methods 5 (1) (2014) 31–49.

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

[81] K. Small, B. Wallace, T. Trikalinos, C. E. Brodley, The constrained weight space svm: learning with ranked features, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 865–872.

[82] W. Yu, M. Clyne, S. M. Dolan, A. Yesupriya, A. Wulf, T. Liu, M. J. Khoury, M. Gwinn, Gapscreener: an automatic tool for screening human genetic association literature in pubmed using the support vector machine technique, BMC bioinformatics 9 (1) (2008) 205.

[83] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al., Bioconductor: open software development for computational biology and bioinformatics, Genome biology 5 (10) (2004) R80.

[84] F. Shull, M. G. Mendoncça, V. Basili, J. Carver, J. C. Maldonado, S. Fabbri, G. H. Travassos, M. C. Ferreira, Knowledge-sharing issues in experimental software engineering, Empirical Software Engineering 9 (1) (2004) 111–137.

[85] S. Vegas, N. Juristo, A. Moreno, M. Solari, P. Letelier, Analysis of the influence of communication between researchers on experiment replication, in: Proceedings of the 2006 ACM/IEEE international symposium on Empirical software engineering, ACM, 2006, pp. 28–37.

Appendix A. Further explanation of tags in Table 8

U(Usable for reproduction): This option is used if the information provided for a certain element are precise and was useful to repeat the study action. This is normally associated with a combination of 'complete' tag in 'identification' and 'description'; and 'public' in 'availability' attributes.

D (Usable for reproduction with some difficulty): Any variation in the identification, description and public attributes from the description above will likely result in a 'D' measure if the information is still found useful. For example, if a data source is precisely described but it is stored on a private repository requiring certain membership or the reader has to take some personal initiative to achieve the expected task.

N (Not usable for reproduction): This is the case when the information provided is does not help the reader in any way to repeat the author's action(s).

+ (Future availability is foreseeable): This sign is used to indicate that a concrete artefact e.g. tool or dataset will still be available in foreseeable future. May be because it's open source, well maintained, funded, managed or because it's been around for some time with an active team and technical support etc.

***** (Flexible): The asterisk sign is used to indicate perceived level of flexibility of:

- Data: In terms of storage or format. The ease of the possibility to transform it from one format or storage technology to another.
- Tools, algorithms or techniques: Is the method or tool written in a popular language with codes made available to the public and easy to modify and/or extend?

- (Irrelevant): Used when an attribute is irrelevant to a given element.

The tags are an overall decision on how useful to reproducibility was the information provided in the study being assessed regarding each information element and its attribute rating. Table 7 provides an example of the attributes judgement per information element for a sample study. In the table, data source has

an assessment of 'D+', the 'D' simply implies that the information regarding the data source given in the study being assessed is found useful (i.e. a reader can use it to find the data) but with some level of difficulty (e.g. the link given was to a general page and the reader have to figure out how to navigate to the specific data webpage). The '+' implies that the data is likely to be persistent may be because its hosted in a public well maintained website or provided by a reputable body that is interested to continue to make it available.

Appendix B. Explanation of some terms/phrases in Table 9

Following are the definitions of some of the phrases used in Table 9:

Raw dataset: This refers to the whole body of the dataset in its original form, in situations where the study under review utilized only a subset of a larger data body. For example, the TREC 2004 dataset consists of 50 DERP review topics where some of the studies reviewed in this study used only 15 or at most 24. The raw dataset in this case is the complete 50 review topics because they were bundled together. Any user will first have to download the whole set before extracting the part required. This may sometimes be the same as the target dataset when the whole set is being used.

Target dataset: The target dataset is the subset (data) of interest in its original form, for a particular study in cases where the data used for the study is part of a larger set. An example is the 15 review topics used in [36] which is a subset of the 50 review topics of the TREC 2004 dataset. This may sometimes be the same as the raw dataset.

Cleaned dataset: This is the processed (through preprocessing or any other data cleaning approach) version of the target dataset.

Internal structure: This entry requires the researcher to describe the different headings under which each data record was categorized and which part is of interest to the study. For example, the TREC 2004 used 50 or more categorical heading to describe each document, part of which are: Title, Abstract, MeSH tag, PMID, publication type, publication year etc. The storage format and or-

der of heading arrangement might also be useful.

Data retrieval method: Information about how the dataset is packaged or stored and what method was used or will be required to gain access to the data e.g direct download from a universal resource locator (URL) or automated retrieval (e.g. web scraping) because the dataset are not bundled together or are from different sources.

Data extraction: Most of the data files are sometimes too large to be opened directly or loaded into memory at once, so, after gaining access to the raw dataset, how were the records of interest for each datum extracted. This is more useful in cases where only partial record of each datum is desired. Again, using the TREC 2004 dataset as an example, most of the studies we reviewed are interested only in four information - title, abstract, MeSH and the publication type out of about 50 information available for each document.

Custom algorithm: In situations where a researcher proposed a new or an improvement to an existing algorithm, the type of description provided for this proposal will determine how well or not it can be reused.