Reproducibility in Machine Learning-Based Studies: An Example of Text Mining

Babatunde K. Olorisade
School of Computing and Mathematics
Keele University
Keele, ST5 5BG
b.k.olorisade@keele.ac.uk

Pearl Brereton
School of Computing and Mathematics
Keele University
Keele, ST5 5BG
o.p.brereton@keele.ac.uk

Peter Andras
School of Computing and Mathematics
Keele University
Keele, ST5 5BG
p.andras@keele.ac.uk

Abstract

Reproducibility is an essential requirement for computational studies including those based on machine learning techniques. However, many machine learning studies are either not reproducible or are difficult to reproduce. In this paper, we consider what information about text mining studies is crucial to successful reproduction of such studies. We identify a set of factors that affect reproducibility based on our experience of attempting to reproduce six studies proposing text mining techniques for the automation of the citation screening stage in the systematic review process. Subsequently, the reproducibility of 30 studies was evaluated based on the presence or otherwise of information relating to the factors. While the studies provide useful reports of their results, they lack information on access to the dataset in the form and order as used in the original study (as against raw data), the software environment used, randomization control and the implementation of proposed techniques. In order to increase the chances of being reproduced, researchers should ensure that details about and/or access to information about these factors are provided in their reports.

1 Introduction

Independent verification of published claims for the purpose of credibility confirmation, extension and building a 'body of knowledge' is a standard scientific practice [13]. Machine learning methods based research are not excluded from this strict scientific research requirement. However, it may sometimes be hard or even impossible to replicate computational studies of this nature [12]. This is why the minimum standard expected of any computational study is for it to be reproducible [11].

In order for a study to be reproduced, an independent researcher will need at least full information and artefacts of the experiment - datasets, experiment parameters, similar software and hardware environment etc., as used in the original study. However, the experience in studies today shows a lack of sufficient information that can enable an independent researcher reproduce majority of the studies successfully.

Submitted to Reproducibility in ML Workshop at the 34th International Conference on Machine Learning (ICML 2017), Sydney, Australia.

Our focus in this work is to explore the state of reproducibility in a discipline adopting machine learning techniques and identify the necessary improvements required. Particularly, we focus on studies adopting text mining techniques for the automation of the citation screening stage in the systematic reviews process.

Systematic review is a structured review approach popular in evidence based research in software engineering and other disciplines like medicine and education. It is used to investigate and draw evidence on the current state of knowledge on any particular topic of research interest in the disciplines, through exhaustive collection and consideration of available publications on the topic [8, 6]. Citation screening is a stage in the systematic reviews process where all the publications retrieved from the initial search are screened for relevance to the review need.

We used our experience from attempting to reproduce six studies to identify high level reproducibility-relevant aspects common to all studies. Then, we assessed 24 more studies regarding the availability or otherwise of information about the identified aspects.

In the rest of the paper, a list information necessary for successful reproduction of a text mining study is presented in Section 2. The methodology of this study is presented in Section 3, while Section 4 highlights the results. The results were further discussed in Section 5, while Section 6 presents the conclusions from this study.

2 Aspects of TM studies critical to reproduction

Contrary to some views, reproducing a study is useful in the sense that it will at least give independent researchers the opportunity to gain a better insight into the situations surrounding the outcome of a certain study. This in turn may facilitate the extension or advancement of such results by independent researchers. An attempt to reproduce six published studies on the automation of citation screening in systematic reviews found that there was insufficient information for successful reproduction [10]. The authors, however, were able to identify key aspects of the text mining experiments where information was needed to facilitate reproduction. These aspects are as listed below:

- Dataset: Information about the location and the retrieval process of the dataset is needed to ensure access to the dataset as used in the study.
- Data preprocessing: The process of ridding the input data of noise and encoding it into a format acceptable to the learning algorithm. Explicit preprocessing information is the first step towards a successful reproduction exercise. An independent researcher should be able to follow and repeat how the data was preprocessed in the study. Also, it will be useful to find preprocessing output information to compare to e.g. final feature vector dimension.
- Dimensionality reduction: In text mining, the feature vector from the preprocessing exercise is usually large and sparse. Therefore, an optional dimensionality reduction technique is employed to further reduce the vector dimension and keep as much as possible only the features that are the most discriminatory. If the dimension of the resulting feature vector from the initial preprocessing activity was reduced, the details of the dimensionality reduction technique(s) should be provided alongside output details to allow for comparison.
- Dataset Partitions: Details of how the dataset was divided for use as training and test data.
- Model training: The process of fitting the model to the data. Making available, as much information as possible regarding every decision made during this process is particularly crucial to reproduction. Necessary information include but not limited to:
  1. Study parameters
  2. Proposed technique details – codes, algorithms etc. (if applicable)
- Model assessment: Measuring the performance of the model trained in 2. Similar information as in 2 applies here as well.
- Randomization control: Most operations of machine learning algorithms involves randomization. Therefore, it is essential to set seed values to control the randomization process in order to be able to repeat the same process again.
- Software environment: Due to the fact that software packages/modules are in continual development with possible alterations to internal implementation algorithms, it is important

that the details of the software environment used (modules, packages and version numbers) be made available.

- Hardware environment (for large data volume): Some data intensive studies are only reproducible on the same machine capacity as was used to produce the original result. So, the hardware information are sometimes essential.

The experience has shown that if the information regarding these aspects are explicitly provided or externally linked to in a study the chances of the study being reproduced will be greatly increased.

3 Assessing the reproducibility of text mining studies

In this study, we assess studies that focus on the application of text mining techniques to the automation of the citation screening stage in systematic reviews for information that might support their reproduction. Based on a reproduction exercise of six studies in this field [4, 2, 7, 9, 5, 3], we identified the common aspects in the text mining studies as discussed in section 2, whose absence of information will influence the successful reproduction of any text mining study. In order to achieve this, we prepared a checklist capturing all the information listed in the background to assess how reproducibility enabled are the 30 studies. The assessment is conducted to see if useful information is available in the studies regarding each of the aspects. A 'Y' is recorded if information is found, an 'N' if no (useful) information is found and an 'X' if the aspect is not relevant in the context of a particular study.

Unlike the mainstream machine learning studies on image classification where some benchmark datasets have been standardized and are easily retrievable through machine learning packages like 'keras' [1], text data (e.g. systematic review datasets) still exist in various forms and repositories (efforts of initiatives like the TREC[1] in the information retrieval domain is commendable and has been helpful at making shared corpora available for text mining research). Therefore, we tried to distinguish between the type of dataset information provided in a study, whether it is the raw data or the actual subset ((target dataset), if only part of a larger set) is used in the study.

4 Results

In this section, we present the outcome of the assessment exercise of the 30 studies based on each aspect. A compressed result of the assessment is presented in Table 1 with more details shown in Table 2. The summary from Table 1 is further presented in a grouped bar chart (Fig. 1) and a scatter plot (Fig. 2) to visually project the distribution and any correlation between (or across) the different entries of the aspects. The results according to each aspect is analysed below:

Dataset: The summary presented in Table 1 (with more details in Table 2 and Figure 1) shows that 26 (87%) of the studies provided information on the original location of the raw dataset they used but only 3 (10%) shared a local copy of the dataset while none of the studies made the subset, restructured or cleaned dataset they eventually used for their studies.

Preprocessing: The details regarding the conduct of the preprocessing activities which includes stopwords removal, stemming, feature representation etc. is found in 17 (57%) of the studies while 21 (90%) of the studies discussed their feature representation approach.

Dimensionality reduction: Though, dimensionality reduction is a key text mining process due to the generation of large but sparse feature vector during preprocessing but the typical benchmark datasets size in systematic reviews (particularly, the ones used in the studies reviewed) are relatively small compared to what obtains in image classification data. As a result, 25 (83%) of the studies did not report conducting any activity to reduce the dimension of their feature vector. But, five (17%) did reduce the dimension of their vector but only three (10%) gave an account of how they went about it. None of the studies made a copy of their final feature vector available for independent use while only one [2] provided intermediate preprocessing output that can be used for comparison.

Data partition: None of the studies provided any information on the portions of data used for either training or testing beyond basic ratio information.

[1] http://trec.nist.gov/

Model training: All the studies provided some details about the training of their models. However, of the 17 (57%) that proposed some new techniques, none of them provided access to their techniques code, four (13%) provided an algorithm of their techniques, only one (3%) made executable file available while 16(53%) provided only a textual description of their techniques.

Model assessment: All the studies were able to describe how their models were assessed.

Randomization control: 28 (93%) of the studies performed operations that involves some randomization in the algorithm execution. However, none of them provided any information on how this was handled.

Software information: The studies generally ($\sim$ 75%) provide the main software they used in their studies. Where they all fail (100%) is in providing the particular details of associated modules and packages as well as their respective version numbers.

5 Discussion

The assessment of available information in the 30 studies as summarized in Table 1 shows that the major points of reproducibility failure relate to:

1. Access to target dataset: The copy of the dataset(s) they used (Table 1, item 2). All the entries has zero value, consequently, no bar in Fig 1 while all the point overlay in Fig 2. The exact copy of dataset used in a study is particularly important as dataset host site or location may become inaccessible at any time.

2. Custom method: The new methods proposed in studies (Table 1, item 9). Providing access to the implementation or executable files of the proposed methods will go a long way to ensure that ambiguities and misinterpretations are eliminated during the reproduction process as against mere text description.

Table 1: Summary of the Assessment of 30 studies for essential reproduction information

Item No.	Elements	Yes	No	N/A
1	Original location of the raw dataset	26	4	0
2	Provided link to local copy of:			
	a. Raw dataset	3	27	0
	b. Target dataset	0	0	0
3	Pre-processing details	17	13	0
4	Feature representation technique	21	9	0
5	Feature selection technique	8	19	3
6	Dimensionality reduction technique	3	2	25
7	Final feature vector — download link	0	30	0
8	Training algorithm	30	0	0
9	Custom algorithm			
	a. Text	16	1	13
	b. Code	0	16	14
	c. Algorithm	4	12	14
	d. Executable file	1	15	14
10	Model assessment method	30	0	0
11	Detailed model assessment result	30	0	0
12	Randomization seed values	0	28	2
13	Training/test data partition available or indices provided			
	a. Link to data partitions provided	0	30	0
	b. (link to) data indices provided	0	30	0
	c. Seed value provided	0	30	0
14	Software information			
	a. Name provided	23	6	1
	b. Version details	0	29	1

Table 2: Assessment of 30 studies for essential reproduction information

Item No.	Elements	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Original location of the raw dataset	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N
	Provided link to local copy of:															
2	a. Raw dataset	N	N	N	N	Y	N	N	N	N	N	N	N	N	N	N
	b. Target dataset	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
3	Pre-processing details	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	N	N	N	Y	N
4	Feature representation technique	Y	Y	Y	N	Y	Y	Y	Y	Y	N	Y	N	N	Y	N
5	Feature selection technique	Y	N	X	X	X	X	X	Y	Y	X	X	X	N	X	X
6	Dimensionality reduction technique	X	N	X	X	X	X	Y	X	X	X	X	X	X	X	X
7	Final feature vector — download link	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
8	Training algorithm	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Custom algorithm															
	a. Text	Y	Y	X	Y	X	X	X	X	Y	X	Y	X	X	Y	Y
9	b. Code	N	N	X	N	X	X	X	X	X	X	N	X	X	N	N
	c. Algorithm	N	N	X	Y	X	X	X	X	X	X	N	X	X	N	N
	d. Executable file	N	N	X	N	X	X	X	X	X	X	N	X	X	N	N
10	Model assessment method	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
11	Detailed model assessment result	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
12	Randomization seed values	N	N	N	N	N	N	N	N	N	N	N	N	N	X	X
	Training/test data partition available or indices provided															
13	a. Link to data partitions provided	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	b. (link to) data indices provided	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	c. Seed value provided	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	Software information															
14	a. Name provided	N	Y	N	N	Y	Y	Y	Y	Y	Y	Y	N	N	Y	Y
	b. Version details	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

Table 2: (continued)

Item No.	Elements	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	Original location of the raw dataset	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y
	Provided link to local copy of:															
2	a. Raw dataset	N	N	N	N	N	N	N	N	N	N	Y	Y	N	N	N
	b. Target dataset	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
3	Pre-processing details	Y	Y	Y	N	N	N	Y	N	Y	N	Y	N	N	N	Y
4	Feature representation technique	Y	N	Y	Y	Y	Y	N	N	Y	Y	N	Y	Y	Y	Y
5	Feature selection technique	Y	Y	Y	Y	X	X	X	N	X	X	Y	X	X	X	X
6	Dimensionality reduction technique	X	X	Y	X	X	X	X	N	Y	X	X	X	X	X	X
7	Final feature vector — download link	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
8	Training algorithm	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Custom algorithm															
	a. Text	X	Y	N	X	Y	X	Y	Y	Y	X	Y	Y	X	Y	Y
9	b. Code	X	N	N	X	N	X	N	N	N	X	N	N	X	N	N
	c. Algorithm	X	N	N	X	N	X	N	N	N	X	Y	Y	X	N	Y
	d. Executable file	X	N	N	X	N	X	N	N	N	X	N	N	X	Y	N
10	Model assessment method	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
11	Detailed model assessment result	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
12	Randomization seed values	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	Training/test data partition available or indices provided															
13	a. Link to data partitions provided	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	b. (link to) data indices provided	N	N	N	N	N	N	Y	N	N	N	N	N	N	N	N
	c. Seed value provided	N	N	N	N	N	N	N	Y	N	N	N	N	N	N	N
	Software information															
14	a. Name provided	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	X	Y	Y	Y
	b. Version details	N	N	N	N	N	N	N	N	N	N	N	X	N	N	N

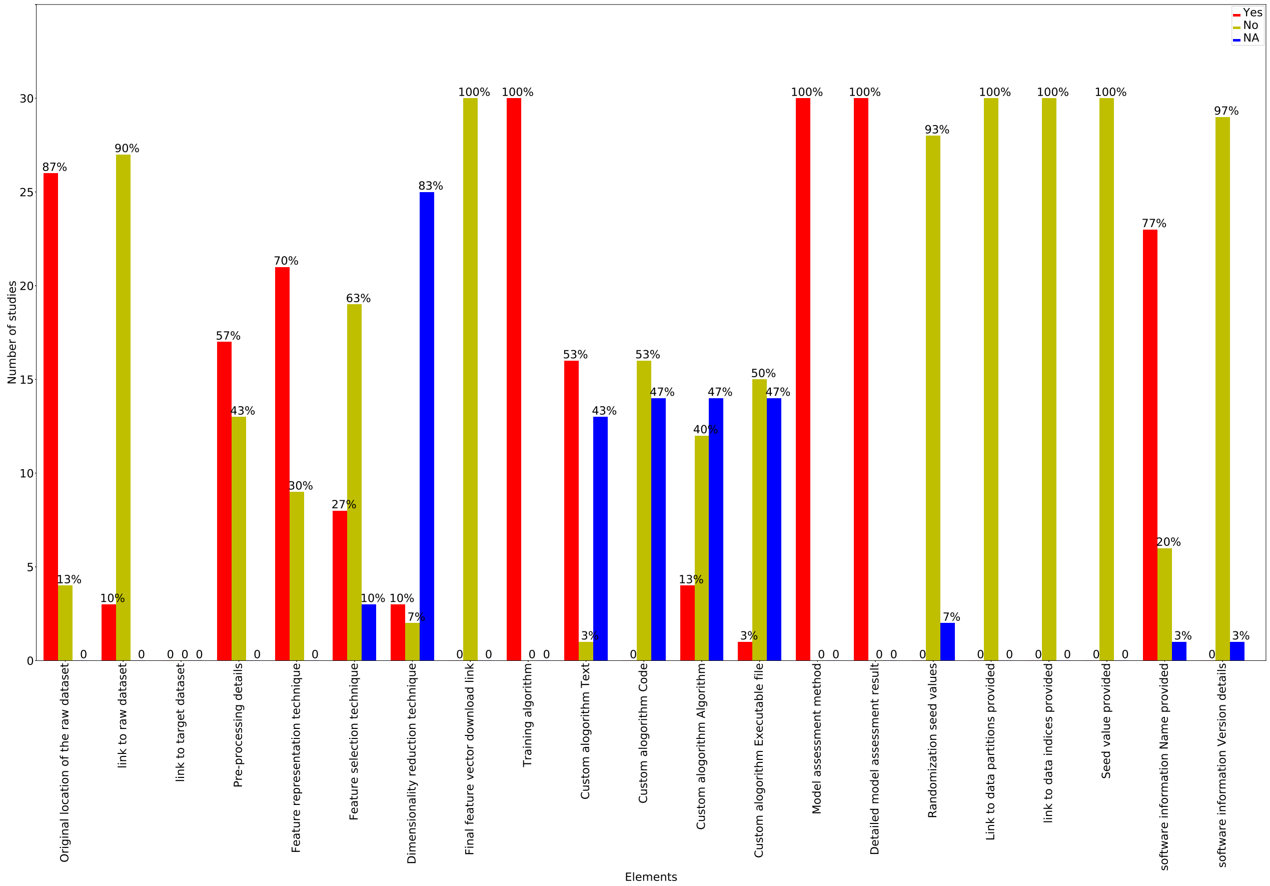

Figure 1: Distribution of studies containing information to support reproducibility

3. Randomization control: This refers to the seed values (or any other techniques used) to control randomization involved in the studies (Table 1, item 12). Even if every other piece of information required is provided, the presence of similar seed values (where necessary) as used in the original study is the only way to ensure the same process is repeated exactly as before.
4. Partioning information: The data partitions (Table 1, item 13) used for at different stages of the study. This is essential as found for example in image recognition datasets like the CIFAR 10 or MNIST datasets where the test set and train sets are provided for uniformity and comparability across experiments. Training a model with different sets of data has the potential to alter the outcome of what the model learned. Hence, difference in results.
5. the names and version numbers of the different modules and packages contained in the software environment used for the studies (Table 1, item 14b) of the table.

The assessment revealed that less attention is paid to the provision of datasets for replication use. Apart from access to the raw dataset, providing access to the different partitions used for training, evaluating or testing purposes had not been given proper attention. As an alternative, with sufficient information and access to ordered dataset, seed value information and algorithms used for the partition will be sufficient but it can be seen in Fig. 1 and Fig. 2 that the assessed studies failed to provide these essential information.

According to Table 1, researchers usually provide the name of the dataset or its host. It should be realized that providing the name of a popular dataset or that of its provider may sometimes be insufficient to have studies reproduced. Beyond the raw dataset, there may be need for extraction of part and even cleaning of the retrieved subset. Independent researchers should be able to get hold

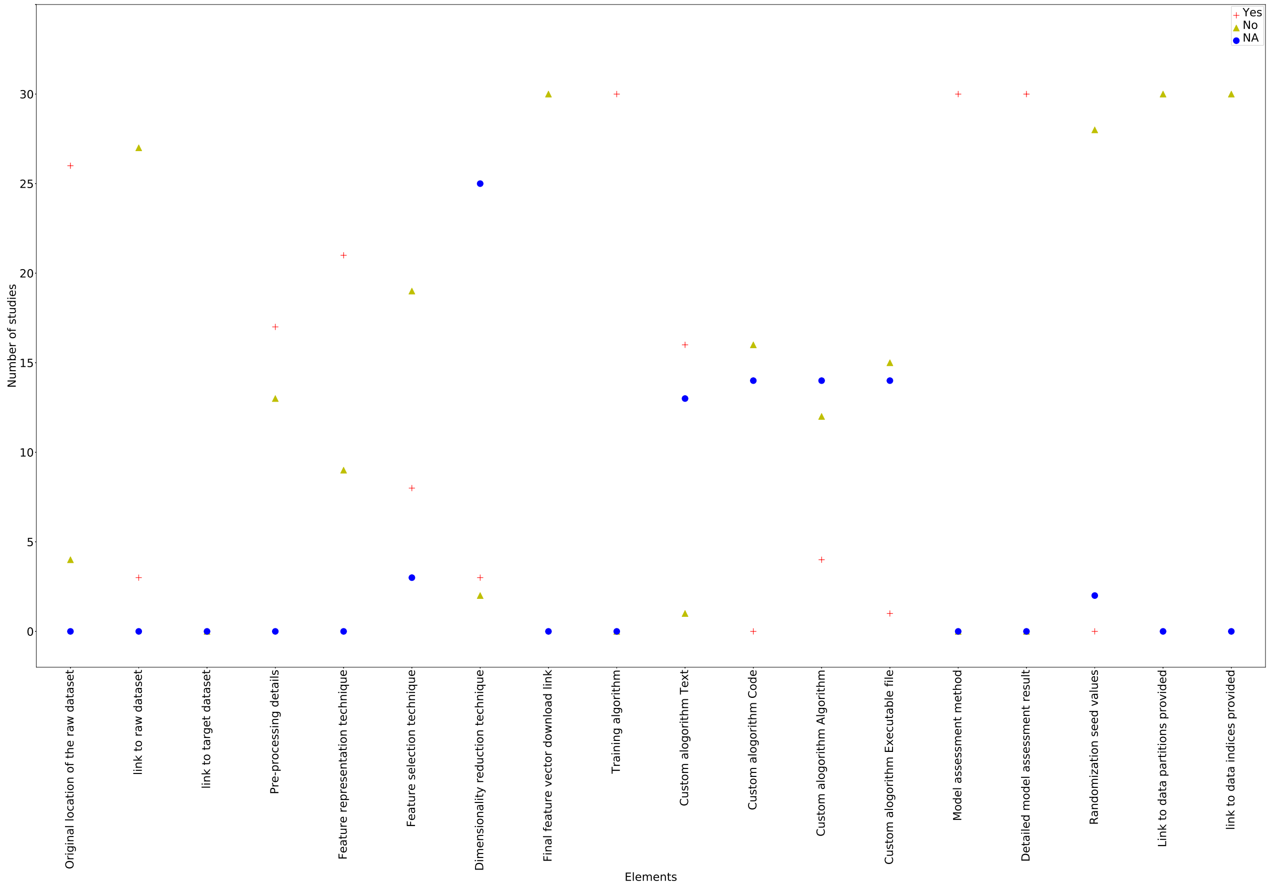

Figure 2: Scatter plot of studies containing information to support reproducibility

of the exact replica and in order, of the dataset used in studies else reproduction may be impossible. Therefore, we recommend that rather than give data or host name, it is more appropriate to provide access to the subset of the data that was used in particular experiments since most of the available dataset like the TREC are usually large and hardly used completely in a single experiment. Otherwise, a link to the raw dataset, access to the code used for extracting the portion used and details of the fields used will suffice.

Given the constant maintenance and updates of software packages, it is important to provided specific details of the software environment used during the course of a study [12]. A notable example is the deprecation of the module used for cross validation in python's sklearn (version 0.17), the *cross_validation* module was discontinued for the *model_selection* module in version 0.18 upwards to perform similar function but with different interface. It was a similar situation for the 'auto' option for the *class_weight* parameter (to cater for class imbalance) in most *sklearn's* classification modules which is now deprecated for the 'balanced' option. On the same dataset both *class_weight* options will produce different results. Other examples include the current changes in the various interfaces of keras 2.0 compared to previous versions.

Furthermore, reproducibility is adversely affected by the lack of detail about implementations of the proposed methods. In the context of citation screening automation, which is the focus of studies assessed in this work, information on dataset partitions and study parameters also contribute to an inability to reproduce these studies.

6 Conclusions

In this study, we highlight those aspects of text mining experiments where information is useful to the reproduction of the studies. In order to identify key factors responsible for the non-reproducible situation encountered in machine learning algorithm based studies, we assess the availability or otherwise of this information in 30 studies conducted on the use of text mining techniques to automate the citation screening stage of systematic reviews. The assessment shows that important explicit information concerning datasets, study parameters (particularly randomization control) and software environment are lacking in most studies and consequently hinder their reproducibility. It is also found that when researchers propose new methods, they only explain it in the study and at best provide some form of algorithms about it. Code implementations and/or executable files are usually not made available for the community's future use. The field thrives on the availability of public datasets; therefore, researchers should also do more by making their knowledge more accessible for easier development and advancement of the body of knowledge.

References

[1] F. Chollet et al. Keras, 2015.

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

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

[4] A. M. Cohen, K. Ambert, and M. McDonagh. A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review. In *AMIA Annual Symposium Proceedings*, volume 2010, page 121. American Medical Informatics Association, 2010.

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

[6] J. P. Higgins and S. Green. *Cochrane handbook for systematic reviews of interventions*, volume 4. John Wiley & Sons, 2011.

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

[8] B. A. Kitchenham, T. Dyba, and M. Jorgensen. Evidence-based software engineering. In *Proceedings of the 26th international conference on software engineering*, pages 273–281. IEEE Computer Society, 2004.

[9] S. Matwin, A. Kouznetsov, D. Inkpen, O. Frunza, and 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):446–453, 2010.

[10] B. K. Olorisade, P. Brereton, and P. Andras. Reproducibility of studies on text mining for citation screening in systematic reviews: Evaluation and checklist. *Journal of Biomedical Informatics*, 2017.

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

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

[13] 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):198, 2011. doi: 10.1186/1752--0509--5--198.

Assessed Papers

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

[2] S. Matwin, A. Kouznetsov, D. Inkpen, O. Frunza, and 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):446–453, 2010.

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

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

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

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

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

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

[9] A. M. Cohen. An effective general purpose approach for automated biomedical document classification. In *AMIA Annual Symposium Proceedings*, volume 2006, page 161. American Medical Informatics Association, 2006.

[10] A. M. Cohen. Optimizing feature representation for automated systematic review work prioritization. In *AMIA annual symposium proceedings*, volume 2008, page 121. American Medical Informatics Association, 2008.

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

[12] S. R. Dalal, P. G. Shekelle, S. Hempel, S. J. Newberry, A. Motala, and K. D. Shetty. A pilot study using machine learning and domain knowledge to facilitate comparative effectiveness review updating. *Medical Decision Making*, 33(3):343–355, 2013.

[13] A. M. Cohen, K. Ambert, and M. McDonagh. A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review. In *AMIA Annual Symposium Proceedings*, volume 2010, page 121. American Medical Informatics Association, 2010.

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

[15] K. R. Felizardo, N. Salleh, R. M. Martins, E. Mendes, S. G. MacDonell, and 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*, pages 77–86. IEEE, 2011.

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

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

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

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

[20] A. Kouznetsov and N. Japkowicz. Using classifier performance visualization to improve collective ranking techniques for biomedical abstracts classification. In *Canadian Conference on Artificial Intelligence*, pages 299–303. Springer, 2010.

[21] A. Kouznetsov, S. Matwin, D. Inkpen, A. H. Razavi, O. Frunza, M. Sehatkar, L. Seaward, and P. O'Blenis. Classifying biomedical abstracts using committees of classifiers and collective ranking techniques. In *Canadian Conference on Artificial Intelligence*, pages 224–228. Springer, 2009.

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

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

[24] M. Miwa, J. Thomas, A. O'Mara-Eves, and S. Ananiadou. Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics*, 51:242–253, 2014.

[25] I. Shemilt, A. Simon, G. J. Hollands, T. M. Marteau, D. Ogilvie, A. O'Mara-Eves, M. P. Kelly, and J. Thomas. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods*, 5(1):31–49, 2014.

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

[27] B. C. Wallace, K. Small, C. E. Brodley, and 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*, pages 176–187. SIAM, 2011.

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

[29] B. C. Wallace, K. Small, C. E. Brodley, J. Lau, and T. A. Trikalinos. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 819–824. ACM, 2012.

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