Conducting a Systematic Review: Trends in Machine Learning and Text Mining *

Mariana Falco[1][1234−5678−9012] and Ignacio Berdiñas[2]

[1] LIDTUA/CONICET, Engineering School, Austral University; Mariano Acosta 1611, Pilar, Buenos Aires, Argentina mfalco@austral.edu.ar
[2] Engineering School, Austral University; Mariano Acosta 1611, Pilar, Buenos Aires, Argentina ignacio.berdiñas@ing.austral.edu.ar

Abstract. The main goal of a Systematic Review is to identify, evaluate, and summarize the findings of all relevant studies over a topic or an issue, making the evidence accessible to decision makers. But the process of manually conducting a systematic reviews takes a lot of time and researchers often have to limit their procedures. With the recent technological advantages, machine learning (ML) and text mining (TM) became useful to aid the systematic review process. The objective of this study is to detect the main trends of these disciplines by carrying out an analysis of a set of relevant articles, identified with a scientific database search between 2015 and 2020. Our analysis showed that mostly ML and TM techniques were applied to three steps: search, screening and data extraction. Huge progresses have been made over the years, but full automation remains a distant goal at present.

Keywords: Systematic Reviews· Literature Reviews · Machine Learning · Text Mining.

1 Introduction

A Systematic Review (SR) can be defined as a method that allows to give meaning and identity to a large amount of information with a clear stated purpose, usually in the form of research studies [1] to answer a set of research questions combining evidence found on those studies [2]. The importance of SRs is that their main goal is to identify, evaluate, and summarize the findings of all relevant studies over a topic or issue, making the resulting evidence more accessible to decision makers [3].

Nevertheless, researchers are forced to limit their search procedures due to the time it takes to conduct a proper systematic review [4]. As pointed out by Zachary [5] reviewers regularly identify relevant searches by performing extensive searches and scanning contents, citations and references. Manually conducting a systematic review is no longer sustainable because practitioners and researchers

* Supported by a Research Grant from Universidad Austral

ISBN 978-987-4417-90-9

use enormous amount of time to perform the tasks of searching, screening, mapping and synthesizing within the process, reducing the hours to explode the creativity.

The process of systematic reviewing includes the following stages: searching, screening, mapping and synthesizing [7]. Text Mining makes possible to analyze collections of textual materials, in order to identify key concepts uncovering also hidden relationships within concepts, allowing the users to efficiently discover, interpret and curate knowledge [6]. Cohen et al. [8] demonstrated that machine learning techniques can reduce the labor required to update systematic reviews. Different authors have produced systematic reviews or reviews, but there are three to mention: Jonnalagadda et al. [9] analyzed methods to automate data extraction, Omara et al. [10] studied the screening phase, and Feng et al. [11] identified and classified text-mining techniques and tools to facilitate conducted a SLR, mostly focused on the SE domain. The biggest difference is that the mentioned articles carried out their analysis considering the studies published from 2014 downward.

The main objective of the present paper is to conduct an analysis of the main contributions of machine learning and text mining for each step within a systematic review process, from 2015 to June 2020 in the point of view of practitioners and researchers, in the industrial and academic context. In order to identify the contributions as well as the existing trends, we performed a comprehensive study based on searches on ACM, IEEE, Springer and Science Direct databases. The goal is to obtain trends over the past five years of technological progresses. The reminder of this paper is as follows: Section 2 presents the followed methodology while Section 3 summarizes the results obtained and theorems defined. Finally, Section 4 provides the conclusions and future work.

2 Methodological Process

Our methodology consisted on the following tasks: define the main goal, perform a search on scientific digital libraries, analyze the search results, identify relevant articles and extract data from these relevant articles.

As stated in the Introduction, our goal can be written as the following research question: *What are the main contributions to each step of conducting a systematic review, between 2015 and 2020?*. Also, the selected digital databases were ACM, IEEE, Springer and Science Direct. In order to make a feasible search, we considered the term extraction which improve the search strategy by creating metadata that can improve its accuracy [7]. In our case, we used the TerMine service which automatically extracts and ranks technical terms.

In this context, the identification of relevant articles were made with the inclusion criteria. Articles published between 2015 and June 2020 were included if they fulfilled the following topics: (a) the title and/or the keywords should contain "text mining applied to systematic reviews" or "machine learning applied to systematic reviews" or similar phrases, and (b) studies which were directed related to text mining and machine learning techniques, approaches, and imple-

ISBN 978-987-4417-90-9

mentations to help the process of a systematic review. Papers on the following topics were excluded: (a) studies on a non-English language; (b) duplicate or updated studies (we selected the most recent one); (c) journals with low impact factors; (d) books; (e) extended abstracts; (f) technical reports; g) doctoral dissertations; (h) thesis, and (i) non-direct application of TX or ML techniques to improve a step within the systematic review process.

From the searches, we obtained 105310 results while querying on ACM, 3005 results on Science Direct, 10 results on IEEE and 1920 results on Springer. After applying the inclusion and exclusion criteria defined above, we gathered the following studies per each database: eight (8) studies from ACM, five (5) studies from Springer, and three (3) studies from Science Direct summarizing sixteen (16) studies to be analyzed.

3 Description of Results and Theorems

As a general overview, we have identified the following publication years: 2015, 2016, 2018, 2019, and 2020; where the years with the highest amount of published studies are 2016 and 2018, followed by 2020, where the latter is still current so new contributions can be added in the following months.

Theorem 1. *The studies were mostly published in journals (8 studies) and conferences (6 studies).*

The studies were mainly published in journals and conferences, while only 2 studies where extracted from symposiums. While the conferences showed dispersion counting six different of them, the journals showed a more fixed distribution including three studies in *Systematic reviews*, and three in *Journal of biomedical informatics*. This is due to the fact that systematic reviews are being performed for different domains, and depending with the main goal of the study as well as the application field, the decision for the publication venue is made around this ideas.

Theorem 2. *The two journals with published analyzed studies were Systematic reviews, and three in Journal of biomedical informatics.*

Theorem 3. *The main steps covered by the studies are search, screening and data extraction.*

Theorem 4. *The step with the higher amount of contributions between 2015 and June 2020 is the screening step.*

Theorem 3 and 4 are condensed in Table 1, which describes the amount of contributions per each step while conducting a systematic review.

Theorem 5. ***Search*** *The studies proposed a set of approaches that includes: iterative methods to build the search string, an automatic query formulation, and an automated approach to extend a search.*

Table 1. Summary of findings.

Step	Studies	References
Search	5	(Cairo et al., 2019), (Lanera et al., 2018), (Marcos-Pablos et al., 2018), (Mergel et al., 2015), (Scells et al., 2020)
Screening	10	(Bannach-Brown et al., 2019), (Howard et al., 2016), (Hashimoto et al., 2016), (Kontonatsios et al., 2020), (Lee and Sun, 2018), (Lee et al., 2020), (Ouhbi et al., 2016), (Sellak et al., 2015), (Tsafnat et al., 2018)
Data extraction	3	(Blake and Lucic, 2015), (Bui et al., 2016), (Chatterjee et al., 2017)

Theorem 6. *Within the screening step, there are several contributions such as high-performing algorithms, approaches to active learning, new topic detection method, automatic text classification approach, approaches for semi-automating screening, and screening systems like SWIFT-Review [17] and SLR Toolkit [25].*

Theorem 7. *In the data extraction step, extraction technologies are still in formative stages.*

Theorem 8. *Methods for automating are still far away for current capacities of machine learning and text mining tools [26].*

3.1 Understanding the techniques

Table 2 summarizes the analyzed articles with respect to the technique used per author and per step.

Table 2: Extracted Techniques Applied on the Relevant Studies

Begin of Table			
Step	Year	Technique used	Evaluation-Performance
Search	2019	TF-IDF, CBOW and SkipGram	Recall and Workload
	2018	TF-IDF, Support vector machine	Area under the receiver operator characteristic curve (AUC)
	2018	TF-IDF, Multinomial Naive Bayes, Bernouli Naive Bayes, k-Nearest Neighbors, Support Vector Machines	F1 measure
	2015	TF-IDF, heatmap	Tool analysis with users
	2020	process to create binary questions	Recall, F1, F3 and Work Saved Over Sampling (WSS)

ISBN 978-987-4417-90-9

Continuation of Table 2

Step	Year	Technique used	Evaluation-Performance
	2019	TF-IDF, LDA, SVM	Recall, specificity, precision, accuracy, WSS@95%, positive likelihood ratio
Screening	2016	new topic detection model	yield, burden
	2020	novel neural network-based feature extraction method	Work Saved Over Sampling (WSS) at r% recall (WSS@r%)
	2018	seed-driven document ranking (SDR) model	Average precision, precision, and recall
	2020	multi-modal missing Data aware stacked auto-encoder	Work Saved Over Sampling (WSS)
	2016	TF-IDF, LDA, Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm	Work Saved over Sampling (WSS)
	2016	new algorithm called Rules7-hybrid feature selection (Rules7-HFSRM)	Precision, recall
	2015	novel Hybrid Feature Selection Method (HFSM) within a Class Association Rules (CARs) algorithm	Precision, recall
	2018	The algorithm was developed using the General Architecture for Text Engineering (GATE)	Precision, recall
	2015	Endpoint detection	Precision, recall
Search	2016	multi-pass sieve algorithm	Accuracy, recall, precision, F-measure
	2017	set of heuristics	Precision, recall

End of Table

As shown in Table 2, we have found different techniques within the studies, where the most applied is TF-IDF (Term Frequency-Inverse Document Frequency) which is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. By means of TF-IDF, knowledge of the research domain is expanded and improved [12].

Theorem 9. *The most applied technique is TF-IDF (Term Frequency-Inverse Document Frequency).*

The following items organize the techniques discovered in the studies.

– TF-IDF: six (6) studies.

- *Latent Dirichlet allocation (LDA), Support Vector Machines (SVM) and Hybrid Feature Selection Method (HFSM)*: two (2) studies per each.
- *Singular Value Decomposition (SVD), CBOW (Continuous Bag of Words), SkipGram, Multinomial Naive Bayes, Bernouli Naive Bayes, k-Nearest Neighbors, Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm, active learning, Heatmap, Neural network, and Seed-driven document ranking*: one (1) study per each.

With respect to evaluation and performance metrics, the most applied where, on the one hand, Precision, or the positive predictive value, refers to the fraction of relevant instances among the total retrieved instances; and on the other hand, Recall, also known as sensitivity, refers to the fraction of relevant instances retrieved over the total amount of relevant instances. In short, precision and recall are measurements of relevance.

Theorem 10. *The most applied evaluation and performance metrics were Precision and Recall.*

- *Recall*: nine (9) studies.
- *Precision*: seven (7) studies.
- *Work Saved over Sampling (WSS)*: four (4) studies.
- *F-measure*: four (4) studies.
- *Accuracy, Workload, WSS@95%, and Specificity*: two (2) studies per each.
- *Average precision, Positive likelihood ratio, Yield, area under the receiver operator characteristic curve (AUC), and Burden*: one (1) study per each.

3.2 Search Step

Techniques Following Table 2, devising an appropriate search string for a secondary study is not a trivial task and identifying suitable keywords has been reported in the literature as a difficulty faced by researchers. In this context, the ML algorithm TF-IDF was applied to different approaches for search string construction [12–14], and to extend a search on PubMed to clinical trials, as well a cross-validated support-vector machine (SVM) model as the classifier [15]. Also, the search string studies applied CBOW (Continuous Bag of Words) and SkipGram in [12]; Multinomial Naive Bayes, Bernouli Naive Bayes, and k-Nearest Neighbors in [13]. Mergel and others [14] have applied the heatmap for visualizing differences between features.

Evaluation and Performance Metrics The most applied performance metric is F1 [13, 22]. Cairo and others [12] used recall and workload, where the latter is used to measure workload in SRs, in the task of search strings. Also, Scells and others [22] measured recall, F1, F3, and Work Saved Over Sampling (WSS), while proposing a five-step approach to automatic query formulation, specific to boolean queries.

3.3 Screening Step

Techniques TF-IDF was used by Bannach-Brown and others [16] to identify potential errors made during the human screening process (including also Latent Dirichlet allocation (LDA), Support Vector Machines (SVM), and Singular Value Decomposition (SVD)), and by Howard and others [17] while introducing the characteristics of SWIFT (Sciome Workbench for Interactive computer-Facilitated Text-mining) a workbench to assist in the problem formulation and literature prioritization, which also includes LDA and the Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm for document prioritization [17].

Hashimoto and others [18] used a neural network-based vector space model to capture semantic similarities between documents; representing documents within the vector space, and cluster the documents into a predefined number of clusters. Ouhbi and others [19] proposed a new algorithm called Rules7-hybrid feature selection (Rules7-HFSRM) by combining the classical algorithm Rules7 and the Hybrid Feature election measure (HFSRM), for text classification. Also, Sellak and others [20] contributed to this line of work by proposing an alternative approach, not yet tested in this domain based on semantic rule-based classifiers. This approach involved applying a novel Hybrid Feature Selection Method (HFSM) within a Class Association Rules (CARs) algorithm.

Evaluation and Performance Metrics Bannach-Brown and others [16] assessed performance using recall (or sensitivity), specificity, precision, accuracy, WSS and the Positive likelihood ratio (LR+). They have obtained that the ML approaches reached 98.7% sensitivity based on learning from a training set of 5749 records, with an inclusion prevalence of 13.2%. The highest level of specificity reached was 86%. Hashimoto and others [18] evaluated performance of the active learning process, over different learning iterations, using two metrics, namely Yield (percentage of eligible studies identified by the active learner), and Burden (percentage of studies that are manually labelled). Kontonatsios and others [23] used WSS, recall and WSS@95%. The proposed method outperforms 10 baseline feature extraction methods by approximately 6% in terms of the WSS@95% metric.

3.4 Data Extraction

Techniques In the experiments performed by Blake and Lucic [21] they used a collection of more than 2 million sentences from three journals Diabetes, Carcinogenesis and Endocrinology and two machine learning algorithms, support vector machines (SVM) and a general linear model (GLM).

Evaluation and Performance Metrics In Blake and Lucic [21], F1 and accuracy measures for the SVM and GLM differed by only 0.01 across all three comparison facets in a randomly selected set of test sentences.

4 Conclusions

The amount of published studies have been growing over the years, and this volume of work lead to develop methods that aim to semi-automate different steps while conducting a systematic review, including machine learning and text mining techniques. Even though there is no unified methodology for applying this methods, or more than one method is valid, we believe that the contributions made a huge progress toward the semi-automation of the steps of a SR. The use of text mining as a second screener may also be used cautiously. The use of text mining to eliminate studies automatically should be considered promising, but not yet fully proven [10].

The present article performed a database search in order to obtain a set of relevant studies from 2015 to 2020, to find out the trends on machine learning and text mining when these disciplines are applied to each of the steps while conducting a systematic review. In this context, it was possible to define a set of theorems to show some trends within the studies. For example, Theorem 3 describes that the techniques were mostly applied to the steps of search, screening, and data extraction.

Most of the tools we encountered were written by academic groups involved in research of machine learning and text mining techniques, but very often the produced prototype were not fully maintainable or even thought to used for other practitioners. Nonetheless, for the pioneering systematic review team, many of the methods described can be used now. Users should expect to remain fully involved in each step of the review and to deal with some rough edges of the software. Data extraction tools are designed to assist the manual process, e.g. drawing the user's attention to relevant text or making suggestions to the user that they may validate, or change if needed.

As a conclusion, it it possible to point out that SRs require very high accuracy in their methods, which may be difficult for automation to attain. Yet accuracy is not the only barrier to full automation. In areas with a degree of subjectivity (e.g. determining whether a trial is at risk of bias), readers are more likely to be reassured by the subjective but considered opinion of an expert human versus a machine. As a side comment, peer reviewing is a standard process for assessing the quality of submissions at academic conferences and journals. We have found a generalized framework for fair reviewer assignment [24], which has been proved that it is superior to the current state-of-the-art.

As future work, we will perform a deeper analysis of the relevant studies in order to describe the hidden relationships as well as a bigger study of the techniques mentioned. Also, we will use abstractive summarization to conduct an evaluation of the included abstract.

References

1. Petticrew, M., and Roberts, H. (2006) Systematic Reviews in the Social Sciences: A Practical Guide, Oxford: Blackwell.

2. Millard, L. A., Flach, P. A., and Higgins, J. P. (2015). Machine learning to assist risk-of-bias assessments in systematic reviews. International journal of epidemiology, 45(1), 266-277.
3. Gopalakrishnan, S., and Ganeshkumar, P. (2013). Systematic reviews and meta-analysis: understanding the best evidence in primary healthcare. Journal of family medicine and primary care, 2(1), 9.
4. Boland, A., Cherry, G., and Dickson, R. (Eds.). (2017). Doing a systematic review: A student's guide. Sage.
5. Zachary, M. A., Gianiodis, P. T., Payne, G. T., and Markman, G. D. (2015). Entry timing: Enduring lessons and future directions. Journal of Management, 41(5), 1388-1415.
6. Ananiadou, S., and McNaught, J. Text mining for biology and biomedicine. Boston/London: Artech House; 2006
7. Ananiadou, S., Rea, B., Okazaki, N., Procter, R., and Thomas, J. (2009). Supporting systematic reviews using text mining. Social Science Computer Review, 27(4), 509-523.
8. Cohen, A., Hersh, W., Peterson, K., and Yen, P.Y.: Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. J Am Med Inform Assoc 2006, 13: 206–219. 10.1197/jamia.M1929
9. Jonnalagadda, S. R., Goyal, P., and Huffman, M. D. (2015). Automating data extraction in systematic reviews: a systematic review. Systematic reviews, 4(1), 78.
10. O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., and Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic reviews, 4(1), 5.
11. Feng, L., Chiam, Y. K., and Lo, S. K. (2017, December). Text-mining techniques and tools for systematic literature reviews: A systematic literature review. In 2017 24th Asia-Pacific Software Engineering Conference (APSEC) (pp. 41-50). IEEE.
12. Cairo, L., de F. Carneiro, G., Monteiro, M. P., and Abreu, F. B. (2019, September). Towards the Use of Machine Learning Algorithms to Enhance the Effectiveness of Search Strings in Secondary Studies. In Proceedings of the XXXIII Brazilian Symposium on Software Engineering (pp. 22-26).
13. Marcos-Pablos, S., and García-Peñalvo, F. J. (2018, October). Decision support tools for SLR search string construction. In Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'18), pp. 660-667. DOI:https://doi.org/10.1145/3284179.3284292
14. Mergel, G. D., Silveira, M. S., and da Silva, T. S. (2015, April). A method to support search string building in systematic literature reviews through visual text mining. In Proceedings of the 30th Annual ACM Symposium on Applied Computing (SAC '15), pp. 1594-1601.DOI:https://doi.org/10.1145/2695664.2695902
15. Lanera, C., Minto, C., Sharma, A., Gregori, D., Berchialla, P., and Baldi, I. (2018). Extending PubMed searches to ClinicalTrials. gov through a machine learning approach for systematic reviews. Journal of clinical epidemiology, 103, 22-30.
16. Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S., Ananiadou, S., Liao, J., and Macleod, M. R. (2019). Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. Systematic reviews, 8(1), 1-12.
17. Howard, B. E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M. R., ... and Macleod, M. (2016). SWIFT-Review: a text-mining workbench for systematic review. Systematic reviews, 5(1), 87.

18. Hashimoto, K., Kontonatsios, G., Miwa, M., and Ananiadou, S. (2016). Topic detection using paragraph vectors to support active learning in systematic reviews. Journal of biomedical informatics, 62, 59-65.
19. Ouhbi, B., Kamoune, M., Frikh, B., Moukhtar Zemmouri, E., and Behja, H. 2016. A hybrid feature selection rule measure and its application to systematic review. In Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services (iiWAS '16). Association for Computing Machinery, New York, NY, USA, 106–114. DOI:https://doi.org/10.1145/3011141.3011177
20. Sellak, H., Ouhbi, B., and Frikh, B. (2015, December). Using rule-based classifiers in systematic reviews: a semantic class association rules approach. In Proceedings of the 17th International Conference on Information Integration and Web-based Applications Services (iiWAS '15). Association for Computing Machinery, New York, NY, USA, Article 43, 1–5. DOI:https://doi.org/10.1145/2837185.2837279
21. Blake, C., and Lucic, A. (2015). Automatic endpoint detection to support the systematic review process. Journal of biomedical informatics, 56, 42-56.
22. Scells, H., Zuccon, G., Koopman, B., and Clark, J. (2020, April). Automatic Boolean Query Formulation for Systematic Review Literature Search. In Proceedings of The Web Conference 2020, (WWW '20), Association for Computing Machinery, New York, NY, USA, 1071–1081. DOI:https://doi.org/10.1145/3366423.3380185
23. Kontonatsios, G., Spencer, S., Matthew, P., and Korkontzelos, I. (2020). Using a Neural Network-based Feature Extraction Method to Facilitate Citation Screening for Systematic Reviews. Expert Systems with Applications: X, 100030.
24. Kou, N. M., U, L. H., Mamoulis, N., and Gong, Z. (2015, May). Weighted coverage based reviewer assignment. In Proceedings of the 2015 ACM SIGMOD international conference on management of data (pp. 2031-2046).
25. Götz, S. (2018, October). Supporting systematic literature reviews in computer science: the systematic literature review toolkit. In Proceedings of the 21st ACM/IEEE International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings (pp. 22-26).
26. Marshall, I. J., and Wallace, B. C. (2019). Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Systematic reviews, 8(1), 163.