Iberoamerican Journal of Science Measurement and Communication Vol. 1, No. 2, pp. 1-22. DOI: 10.47909/ijsmc.52

ORIGINAL ARTICLE

A roadmap toward the automatic composition of systematic literature reviews[1]

Eugênio Monteiro da Silva Júnior [2] , *Moisés Lima Dutra* [3]

How to cite (APA): da Silva Júnior, E.M.; & Dutra, M. L. (2021). A roadmap toward the automatic composition of systematic literature review. Iberoamerican Journal of Science Measurement and Communication; 1(2), 1-22. https://doi.org/10.47909/ijsmc.52

Received date: 17-05-2021

Accepted date: 22-07-2021

Handling editor : Carlos Luis González-Valiente

ABSTRACT

Objective. This paper presents an overview of existing artificial intelligence tools to produce systematic literature reviews. Furthermore, we propose a general framework resulting from combining these techniques to highlight the challenges and possibilities currently existing in this research area.

Design/Methodology/Approach. We undertook a scoping review on the systematic literature review steps to automate them via computational techniques.

Results/Discussion. The process of creating a literature review is both creative and technical. The technical part of this process is liable to automation. Based on the literature, we chose to divide this technical part into four steps: searching, screening, extraction, and synthesis. For each one of these steps, we presented practical artificial intelligence techniques to carry them out. In addition, we presented the obstacles encountered in the application of each technique.

Conclusion. We proposed a framework for automatically creating systematic literature reviews by combining and placing existing techniques in stages where they possess the greatest potential to be useful. Despite still lacking practical assessment in different areas of knowledge, this proposal indicates ways with the potential to reduce the time-consuming and repetitive work embedded in the systematic literature review process.

Originality/Value. The paper presents the current possibilities for automating systematic literature reviews and how they can work together to reduce researchers' operational workload.

---

1 This article is an extended version of the paper presented at the virtual conference DIONE 2021, celebrated on March 10–12, 2021.

2 *Federal University of Santa Catarina, Brazil. Email: eugeniomonteiro@hotmail.com, ORCID: 0000-0003-3510-8882.*

3 *Federal University of Santa Catarina, Brazil. Email: moises.dutra@ufsc.br, ORCID: 0000-0003-1000-5553.*

Keywords: Systematic literature review; Automation; Text mining; Framework; Bibliographic data analysis; Natural language processing

1 INTRODUCTION

Remarkably, scientific production keeps growing at an accelerated rate. According to Johnson et al. (2018), in August 2018, there were 33,100 active English-language peer-reviewed journals, which published together 3 million papers per year, resulting in annual growth of approximately 5%. Given this high amount of publications, writing literature reviews is extremely time-consuming since it requires analyzing a large number of texts. Although information technology tools have facilitated access to a myriad of journals worldwide and have made the search process more streamlined, the human effort to find potentially valuable information within a document when a large number of text files is retrieved is still too high. According to Wallace et al. (2010), an experienced reviewer can evaluate an average of two abstracts per minute. In the case of more complex topics, each abstract may require more minutes to be evaluated. It is noteworthy that sometimes it may be necessary to read hundreds or thousands of abstracts just to select the initial group of relevant papers. The evolution of artificial intelligence (AI) techniques observed in recent years, especially in the subarea known as natural language processing (NLP), allows us to envisage scenarios in which these modern techniques and their associated tools can be used to enhance the process of creating literature reviews, from an automatic composition approach (Silva Júnior & Dutra, 2021).

Through a scoping review (Munn et al., 2018), this paper aims to present an overview of the current scenario of applying AI techniques to create literature reviews automatically. Furthermore, we propose a framework resulting from combining these techniques to highlight the existing challenges and possibilities in this research area. The paper is structured as follows: in section 2, we discuss the automatic creation of systematic literature reviews. In section 3, we present and discuss related works and the paper's background and motivations. In section 4, we present the proposed framework. In section 5, we further depict the framework's internal structure. Finally, in section 6, some conclusions and perspectives for future work are given.

1.1 Automatic creation of systematic literature reviews (SLR)

Before automating a systematic literature review (SLR) process, it is necessary to know how it is manually conducted. Systematic reviews are a widely used method to reliably gather research results. As a research method, systematic reviews are undertaken according to explicit procedures. The term "systematic" distinguishes them from reviews undertaken without clear and accountable procedures (Gough et al., 2012).

According to Grant and Booth (2009), the SLR seeks to gather all available knowledge on a given topic to guarantee transparency in reporting their methods to facilitate other researchers to replicate that process (Jonnalagadda et al., 2015). SLR also helps identify research gaps to develop new ideas. Together with the SLR, there are other types of literature reviews whose objectives differ somehow. Grant and Booth (2009) provide descriptive insight into the 14 most common types of reviews, for example, scoping and state-of-the-art reviews. Scoping review provides a preliminary assessment of the potential size and scope of available research literature. It aims to identify the nature and extent of research evidence, usually including ongoing research (Grant & Booth, 2009). A state-of-the-art review mainly considers the most current research on a given area or topic. It often summarizes current and emerging

trends, research priorities, and standards in a particular field. This review aims to provide a critical survey of the extensive literature produced in recent years and synthesize the current thinking in the area. It may offer new perspectives on an issue or point out an area that needs more research (Dochy, 2006). In addition, there is the generic term “literature review,” i.e., a non-systematic literature review that aims to provide a summary or overview about a particular topic. In contrast, an SLR aims to answer a specific research question (Davis, 2016).

There is no consensus regarding how many steps the SLR can be divided into. Several different proposals in this regard can be found in the literature. While some authors propose only three steps, others, like Tsafnat et al. (2014), suggest 15 steps. This work considers the four steps described next. This decision proved to be capable of separating distinct tasks without generating an excessive number of steps or group tasks that are poorly related to each other within the same step. This facilitates the process of developing automated tools specific to each step. According to Ananiadou et al. (2009), the following steps are usually part of a review process:

1. **Searching:** Extensive searches are carried out to locate as much relevant research as possible according to a given query. These searches include scrutinizing electronic databases, scanning reference lists, and searching for published literature.
2. **Screening:** It narrows the search scope by reducing the collection to only the documents relevant to a specific review. The aim is to highlight key evidence and results that may impact the policy.
3. **Mapping:** The Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) [4] has pioneered the use of “research maps” to understand research activity in a given area, engage stakeholders, and identify priorities concerning the review focus.
4. **Synthesizing:** It correlates evidence from a plethora of resources and summarizes the results.

The process of preparing a systematic review is both creative and technical. It is worth mentioning that there is a natural dichotomy of tasks: creative tasks are performed to develop the core question to be answered and the protocol to be applied. In contrast, technical activities can be performed automatically following exactly the applied protocol (Tsafnat et al., 2014).

There are some standards for developing systematic reviews in a traditional way that can serve as guides for the automation process, such as the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement and the PICO (Patient, Intervention, Comparison, and Outcome) framework. PRISMA consists of a checklist with 27 items and a four-step flowchart to help authors improve the reporting of systematic reviews and undertake meta-analysis. It is focused on randomized trials, but it can also be used as a basis for reporting SLR from other types of research, particularly intervention evaluations. PRISMA may also be useful for the critical appraisal of published SLR (Moher et al., 2009). Regarding the PICO framework, according to Davis (2016), it can be used to develop a well-formulated research question with a clear statement of objectives. The JAMA User’s Guide to the Medical Literature (Guyatt et al., 2015) defines PICO as an expanded version of the original format:

4 A specialist center for (i) developing methods for systematic review and synthesis of research evidence; and (ii) developing methods for the study of the use research. Available at https://eppi.ioe.ac.uk

- P = Patients or Population. Who are the relevant patients?
- I = Intervention(s) or Exposure(s). What are the management strategies we are interested in comparing or the potentially harmful exposures we are concerned about?
- C = Comparator. There will always be both an experimental intervention or putative harmful exposure and control, alternative, or comparison intervention for issues or therapy, prevention, or harm.
- O = Outcome. What are the patient-relevant consequences of the exposures in which we are interested? We may also be interested in the consequences to society, including cost or resource use (Speckman & Friedly, 2019).

Some other standard models worth mentioning are PEO (Patient, Exposure, and Outcome) and PIO (Patient, Intervention, and Outcome). They are used to formulate the inclusion and exclusion criteria defined to select relevant studies to answer the research question (Davis, 2016).

1.2 Background and related works

Since the production of the SLR is both creative and technical, it is expected that all stages considered technical are subject to automation. Indeed, the idea of automating the steps of a systematic review is not exactly new. According to Jonnalagadda and Petitti (2013), the first paper that recommended the usage of machine learning (ML) for this purpose was published in 2005. From that year on, several works were published regarding computational techniques in each of the SLR stages. One good way to observe the evolution of this idea is by reading systematic reviews published on this subject. For example, in 2015, while Jonnalagadda et al. (2015) published a review that exclusively covers data extraction works in SLR, while O'Mara-Eves et al. (2015) dedicated to reviewing papers related to automatic identification of relevant studies.

The text mining methods considered most relevant to support systematic reviews are automatic term recognition (ATR), text clustering, text classification, and text summarization (Thomas et al., 2011). Moreover, there are also many software applications specifically developed to assist during the production of systematic reviews. The SR Toolbox [5] website provides, for example, a list of various available tools for supporting systematic literature reviews. Therefore, to look for gaps or opportunities to optimize processes and support the proposed framework, it is necessary to know the existing tools to assist in the automatic generation of SLRs. It is also worth emphasizing that it is out of the scope of this work to indicate whether the existing systems could be improved. In addition, we highlight that all the technical steps of the review process can be automated using computational techniques to reduce the human workload.

After defining the research and the inclusion criteria, the first technical stage of an SLR is the search for correlated studies. Ideally, 100% of the existing studies on the topic should be retrieved. Text mining can help by suggesting possible query terms. Even if the researcher has already found some documents that meet his/her inclusion criteria, he/she can always use a term recognition service that suggests new terms and concepts be used in a new query (Thomas et al., 2011). According to Ananiadou et al. (2009), term extraction improves the search strategy by creating additional metadata that can increase accuracy by automatically identifying key phrases, concepts, or technical terms, within the documents. The improvement in the set of search terms can expand the coverage of the results, which may be sufficient in

5 http://www.systematicreviewtools.com/

cases where the object of study is specific. Thus, the number of papers retrieved is relatively small, but the entire literature is covered. In some other cases, the number of papers retrieved is extremely high, making it even more challenging to find works relevant to the subject being searched. Consequently, it is possible to think about methods that can be applied to the set of retrieved papers to find those relevant.

In a systematic review, the term *screening* refers to the manual process of sifting through, at times, thousands of titles and abstracts that are retrieved from database searches. To improve reliability, the titles and abstracts are often screened by two people. This is a very labor-intensive task and adds considerably to the review's cost and time (Thomas et al., 2011). This is a stage where machine learning techniques can be very useful by filtering not only titles and abstracts but also full texts. Khabsa et al. (2016) say that the first study to consider this possibility was by Cohen et al. (2006). According to Thomas et al. (2011), there are two ways to use text mining to automate this step: the first aims to prioritize the list of items for manual screening so that the studies at the top of the list are those most likely to be relevant; the second one uses the studies manually labeled (included/excluded) as a training dataset so that the system can "learn" to classify the other works automatically. Conducting this step in a semi-automatic manner can bring many benefits to the researcher. However, Marshall and Wallace (2019) highlight that the main limitation of the automatic screening of abstracts is that it is unclear at what point it is "safe" for the reviewer to interrupt the manual screening. Even systems that, instead of providing a definitive and dichotomous classification, provide classifications based on probabilities are not free from the risk of loss. For example, a paper that has received a low probability may be relevant. If a researcher chooses to stop screening at a certain probability threshold, this paper may not be included in the results.

Another way to find relevant studies to a literature review is by citation mining. For example, the study of Belter (2016) proposes a method to systematically mine the various types of citation relations between papers to retrieve documents related to the topic searched by a specific systematic review. The author's proposal is conceptual and was conducted manually. This method, according to this author, despite having potential for automation, had some limitations related to the available database APIs that made it impossible to create a computational algorithm at that time. However, the existing database APIs provide more and more information about the indexed papers, making it possible to write algorithms that can automatically retrieve related papers through the citation mining technique. Moreover, Sarol et al. (2018) stress that it is possible to integrate the aforementioned content-based methods with the citation-based methods to create a more efficient model for retrieving relevant papers.

Once the relevant studies are identified and retrieved, the next step is to extract the useful information present in each one of them. According to Tsafnat et al. (2014), extracting data from texts is one of the most time-consuming tasks in a systematic review. Therefore, there are already several works whose objective is to automatically extract data from texts. According to Marshall and Wallace (2019), when considered specifically reviews of the randomized controlled trial (RCT), there are only a few platform prototypes that make these technologies available, such as ExaCT [6] (Kiritchenko et al., 2010) and RobotReviewer [7]. The NaCTeM (the United Kingdom National Center for Text Mining) has developed several

6 https://exact.cluster.gctools.nrc.ca/ExactDemo

7 https://www.robotreviewer.net/

systems that use structured models to extract concepts such as genes and proteins from texts for basic science reviews. Since the desired information can be present in several paper sections, extracting it can become a complex cognitive task. Consequently, even partial automation can reduce the time required to complete this task, as well as reduce errors and save time (Tsafnat et al., 2014).

One obstacle to achieving better data mining models is the lack of training data. ML systems need a dataset with manually assigned labels to adjust model parameters. However, associating labels with individual terms in documents to enable training data-extraction models is an expensive task. EXaCT, for example, was trained on a small set (132 in total) of full-text papers. RobotReviewer was trained using a much larger dataset, but the 'labels' were semi-automatically induced, using a strategy known as 'distant supervision.' This means the annotations used for training were imperfect, thus introducing noise to the model (Marshall & Wallace, 2019). Recently, Nye et al. (2018) released the EBM-NLP dataset, which comprises about 5000 abstracts of RCT reports manually annotated in detail. This may provide training data helpful for developing data extraction models (Marshall & Wallace, 2019).

The last SLR step that text mining techniques can assist is information synthesis. According to Marshall & Wallace (2019), although the software tools to support the synthesis of revision data have been around for a long time (especially for performing meta-analyses), the methods for automating it are beyond the capabilities of ML and NLP. Furthermore, it is also possible to think about ways to automatically summarize the texts selected for review by extracting information from the full texts of the papers and not just from their abstracts. For this, there is the technique that creates automatic text summaries. According to Mani (2001), this technique either generates a summary for a single document at once or for multiple documents together (MDS - multi-document summarization) by extracting the most relevant information found within the texts. Automatic summarization is quite important in systematic review processes, as it condenses the information that was discovered and classified and thus provides a solution to the information overload problem (Thomas et al., 2011).

The use of MDS methods offers the benefits of reducing the overwork on the reviewer, as well as enabling an overview of a body of research. However, the proper place and use of such summarization must be established to offer the greatest benefit regarding the current state of the art. This is partly an issue for a system designer and partly an issue of training and experience for the reviewer. Thus, running an MDS on a large collection of texts from many domains on many subjects would probably not be a useful exercise and would indicate a lack of understanding about getting the most from summarization. However, if the reviewer has previously produced a cluster or a classification of documents, it makes sense to apply the MDS since documents in a cluster or class can reasonably be expected to have something in common. Consequently, the results would be meaningful (Thomas et al., 2011).

Finally, natural language generation (NLG) technologies can automatically write review-specific paragraphs, such as describing the types of documents retrieved, evaluation results, and summary of the conclusions (Tsafnat et al., 2014). Currently, existing techniques are not able to produce perfect texts like those written by humans. However, the automatically generated text can serve as a basis for the text to be written manually by reviewers, e.g., avoiding errors in data transfers from multiple sources. Importantly, this kind of technology still has a lot to be improved. Thereby, researchers should be aware of the emerging new text generation methods in search of more integrated tools to automate systematic reviews.

Unlike the studies aimed to automate individual SLR steps, such as the works of Ros et al. (2017) and Marcos-Pablos and García-Peñalvo (2018), some others propose simultaneous

automation of multiple SLR steps. This second group of proposals is closer to ours, as they work with an overview of the entire process. Pulsiri and Vatananan-Thesenvitz (2018) presented the results of two systematic literature reviews. In one of them, the SLR and automation were researched, while in the other one the SLR and bibliometrics. They both centered on finding tools and methods to support the reviewing process by proposing a framework that divides SRL into four stages and integrates automation and bibliometric tools into this process. Although this work also considers four steps similar to ours, the proposed steps set is different from the one presented here. Besides, there is a difference in terms of focus and level of detail.

Van Dinter et al. (2021) performed a systematic review of the literature on studies about SLR automation to collect and summarize the current state of the art. Their study is the first systematic literature review on automating systematic literature reviews, focusing on all stages of the SLR, including NLP and ML techniques from a retrospective approach. Finally, Felizardo and Carver's work (2020) conducted a (non-exhaustive) literature survey to describe strategies and tools to support or automate the SLR process or its tasks. Notwithstanding, there are some differences between the two proposals. Felizardo and Carver (2020) do not specify the type of tool to be considered in their research, while our work focuses on AI-based tools. That is, our work is more generalist regarding the application area, while Felizardo and Carver's (2020) focuses on specific reviews in the software engineering area. Another difference has to do with the number of steps proposed for automating the SLR. Felizardo and Carver (2020) propose five; while in our framework four are defined. In the end, this study is more oriented to the supporting tools, thus it presents more technical details than the aforementioned proposals.

2 A FRAMEWORK FOR CREATING SYSTEMATIC LITERATURE REVIEWS

Figure 1 shows the proposed framework for combining several techniques used by different projects to automatically generate a systematic literature review. We aimed to indicate what should be the focus of researchers on AI when they are going to work within this theme. Disregarding the steps that naturally involve a creative process and, consequently, must be performed by humans, the next paragraphs focus on the operational tasks that are part of the reviewing process. This framework shows not only a sequence of technical steps required for creating an automatic literature review but also the respective AI techniques that can be useful in each step. In addition, Table 1 details and justifies the use of the techniques chosen to compose the proposed framework.

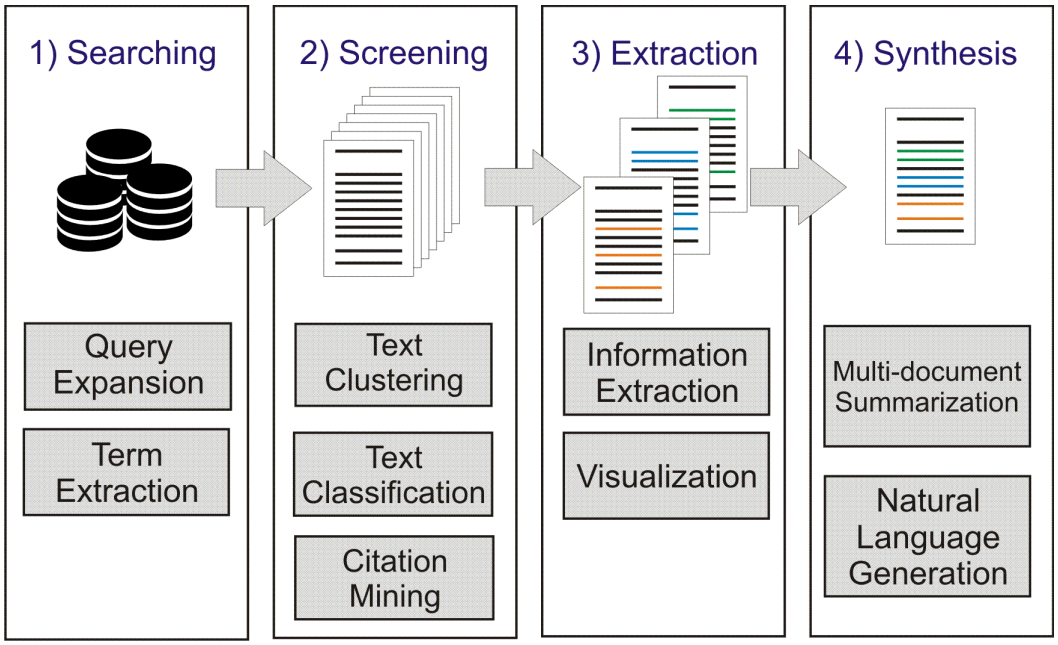

***Figure 1**. Main steps and associated methods for the automatic creation of systematic literature reviews.*

It is important to emphasize that this paper does not intend to provide strict rules of how an SLR should be automated nor to indicate specific tools or technologies for that purpose. Instead, the objective here is to highlight, based on the scientific literature on this topic, the stages of the literature review framework with the greatest potential for automation, namely: searching, screening, extraction, and synthesis.

In the searching step, computational techniques can help suggest terms to maximize the number of documents retrieved (increase recall) or minimize the number of unrelated papers (increase precision). However, the human operator remains essential to carry out the process. Thus, this step is considered to have a medium automation potential. Despite that, some works found in the literature propose techniques with great potential for increasing automation in this stage (Scells & Zuccon, 2018; Scells et al., 2020; Marcos-Pablos & García-Peñalvo, 2018). Machines usually perform this task better than humans do. Besides, scientific databases are increasingly providing structured data on references, which facilitates the automation process. In addition, it is possible to work on the direct extraction of references from papers (i.e., PDF files). Therefore, this is a step that deserves more attention and investment from researchers.

As for the screening stage, we consider that it still has a low potential for automation currently. Converting texts into vectors while preserving semantic relations is still incipient, among other causes, due to the small number of positive training examples available. Text classification methods are extremely dependent on a good conversion of texts into vectors since the extraction of text characteristics highly impacts their classification accuracy. Usually, ML-based classification methods depend on training data to 'learn' patterns. In systematic reviews, the number of papers labeled as 'included' is less than the size of the set of 'excluded' papers, making it difficult to properly adjust ML-based models due to this set imbalance. Furthermore, as previously mentioned, given the risk of losing potentially relevant papers during the screening stage, researchers may not feel secure in delegating the exclusion of much of the retrieved papers to an automatic classification process. Thus, we believe it is still necessary to develop new methods for extracting more precise text

characteristics so that it is possible to consider that automatic sorting as a secure time-saver for researchers.

The extraction and synthesis steps present a great potential to be automated. In these stages, computational techniques operate by extracting and organizing important information from texts. Various techniques for extracting certain data from texts are being developed and applied at this stage of the automatic creation of literature reviews. Especially for medical reviews, which are already more standardized, there is great potential for applying these techniques. Nevertheless, as natural language processing keeps evolving, it is possible to imagine for the near future the extraction of texts for automatically creating literature reviews to be applied to some other areas of knowledge, such as the social sciences.

Technique	**Inputs**	**Outputs**	**Main Approaches**	**Contribution to Automating SLRs**
Query expansion	Search terms	Optimized Boolean query	Conceptual or statistic approach	Used to increase recall (more comprehensive search) or increase precision (better-targeted search) of search results
Term extraction	Seed documents	Technical terms	Statistical measures	Improves the search strategy by creating additional metadata that can increase accuracy by automatically identifying key phrases, concepts, or technical terms, within the documents
Text clustering	Abstracts or full texts	Texts grouped by similarity	General clustering algorithms and Topic Modeling	Facilitates the identification of papers that have similarities to each other
Text classification	Abstracts or full texts	Papers classified as “relevant” or “irrelevant”	Machine learning-based classifiers	Selection of abstracts or full texts relevant to the review. A part of the database must be labeled to train the classifier
Citation mining	Reference metadata	Citation graph (or network)	Named Entity Recognition, Built-in citation mining tools (e.g., Google Scholar API), Citation mining databases (e.g., CrossRef API)	Finds more relevant papers based on citations in the relevant articles already identified
Information extraction	Full text of relevant papers	Important information from each paper	Rule-based approach, Machine learning, Deep learning, and Hybrid approaches	Extracts relevant information from scientific papers
Visualization	Full text of	Important information	Visual Text	Presents the information in a visual

Technique	Inputs	Outputs	Main Approaches	Contribution to Automating SLRs
	relevant papers	highlighted or displayed in graphic form	Mining	form, as this way the user more quickly evaluates it
Multi-document summarization	Full text of relevant papers	Summaries with the most important information from each paper	Multi-objective optimization	Produces summaries with the most important information contained in the papers considered most relevant to the review
Natural language generation	Full text of relevant papers	Human-understandable texts	RNN and LSTM	Writes specific review paragraphs, such as describing the types of the retrieved documents, evaluation results, and summary of the conclusions

Table 1. Conceptual details to justify the use of the techniques chosen to compose the proposed framework.

3 DISCUSSION

This section aims to give further details on the computational techniques applied to automate each step of the proposed framework.

3.1 Searching

At this stage, computational techniques are used to increase recall (more comprehensive search) or increase precision (better-targeted search) of search results (Tsafnat et al., 2014). In addition, an automated system can assist the researcher by suggesting synonyms or related keywords to expand the coverage of the results.

As previously mentioned, many studies focus on reducing the workload on the screening step by applying machine learning techniques. However, according to Scells and Zuccon (2018), the search query can have a much more significant effect on screening workload reduction simply by reducing the number of studies retrieved. However, developing queries is challenging and time-consuming (Bullers et al., 2018; Golder et al., 2008). SLR seeks to identify and synthesize all relevant studies available on a given topic. To find these papers, queries are usually made in scientific databases through Boolean operators (AND, OR, NOT) searching for terms present in the title, abstract, and keywords. However, a query is far from the optimal result when it retrieves more papers than necessary, i.e. when the number of false positives (irrelevant papers) is much greater than the number of true positives (relevant papers) (Scells et al., 2020). The automation of query generation in SLR has not been widely explored. Recently, efforts have been made to automate information specialists' processes to develop Boolean queries (Scells et al., 2020). Scells et al. (2020) present fully automatic computational adaptations of two manual approaches developing queries that information specialists employ. A first manual approach is a conceptual approach (Hausner et al., 2012), where a query is developed by identifying high-level concepts and finding synonyms for those concepts. The objective approach is the second one (Clark, 2013), where a query is developed by identifying and classifying terms using a statistical approach.

The conceptual approach is the most commonly used approach to develop Boolean queries

for SLR-like researches. Under this approach, several high-level concepts are identified, either from seed studies or through initial searches, representing the research question of the review. These concepts are often categorized using the PICO question scheme. Once the topic experts have identified the high-level concepts used to develop the research, they use their experience and a range of tools to help them identify synonyms and keywords related to their high-level concepts (Scells et al., 2021).

The objective approach is a relatively recent approach to developing Boolean queries for SLRs. The steps in this query development method are more well-defined than the conceptual approach, and, therefore, it is easier to simulate this method computationally. It involves the use of statistical methods to identify which terms should be added to the query. This approach uses seed papers to identify terms and evaluate the effectiveness of the terms (Scells et al., 2021).

The work Scells et al. (2021) compared computational implementations of the two aforementioned approaches (conceptual and objective). The results indicated that no automatic formulation method performed better than the queries elaborated manually, which shows that the elaboration of automatic queries still has a lot to evolve.

3.2 Screening

The screening step is one of the most explored in the literature (van Dinter et al., 2021). At this stage, the main idea is to train a machine learning model to classify papers as relevant or irrelevant (binary classification) to a particular research question. Several classifiers can be used, and they all depend on a reliable feature extraction to perform well. Furthermore, as this process is a classification of textual data, feature extraction of these texts must be carried out so that the classifier can assess the similarity between them. Therefore, before addressing the screening itself, it is necessary to know the basics of text feature extraction methods.

Most existing semi-automatic citation screening methods adopt unsupervised document representation techniques, such as bag-of-words (BoW), to address an inherently supervised classification task (Kontonatsios et al., 2020). BoW is one of the most well-known methods of selecting textual features that have already been widely adopted by semi-automatic citation sorting methods (Cohen et al., 2006; Wallace et al., 2010). In the BoW model, each document is represented as a sparse and high-dimensional feature vector. Each dimension of the vector corresponds to words or phrases that occur in the document (Kontonatsios et al., 2020). A limitation of the BoW model is that the resulting feature vector consists of a large number of words, so the model is associated with increased memory and computational costs when applied to large-scale systematic review datasets (Forman, 2003). In a multi-document scenario, a large number of vector positions are set to zero, which also contributes to increasing the computational cost. Despite being easy to implement and interpret, one-hot-encoding models (such as BoW) are known for not considering the order or semantics of words (Le & Mikolov, 2014). An alternative is to use the paragraph vector (PV) method. The PV model is a feature extraction method based on a neural network that follows a distributional semantics approach to better account for the words and document semantics. More specifically, the PV model trains a shallow neural network, consisting of a hidden layer, maximizing the conditional probability of a word given the context and the document in which it appears (Kontonatsios et al., 2020).

Once defined the method for extracting features from the texts, the next step is to classify the papers as relevant or irrelevant for the review using one of the available classifiers. The system works by receiving abstracts of papers classified as relevant or irrelevant by the

human operator to build a training base for the automatic classifier. Thus, the system “learns” to classify the other abstracts that the user has not yet read. Figure 2 presents a generic scheme of how the screening stage works using the active learning approach. Figure 3 shows a more specific operating scheme: an automatic paper prioritization system using the SVM classifier.

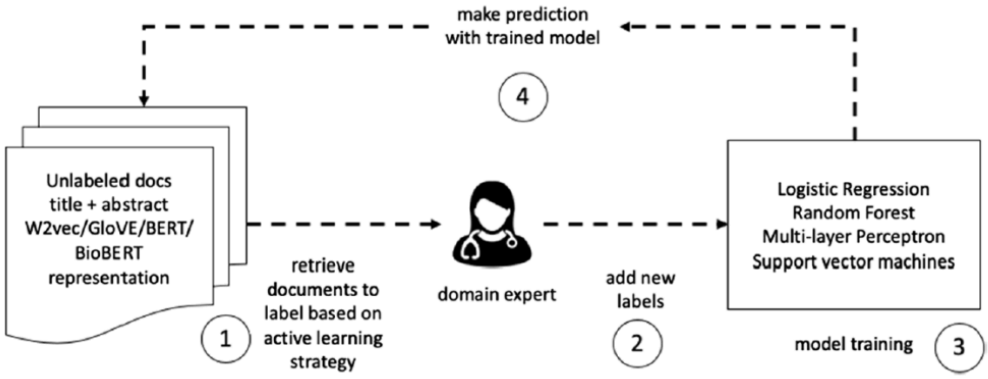

Figure 2. Representation of the active learning approach (Carvallo et al., 2020).

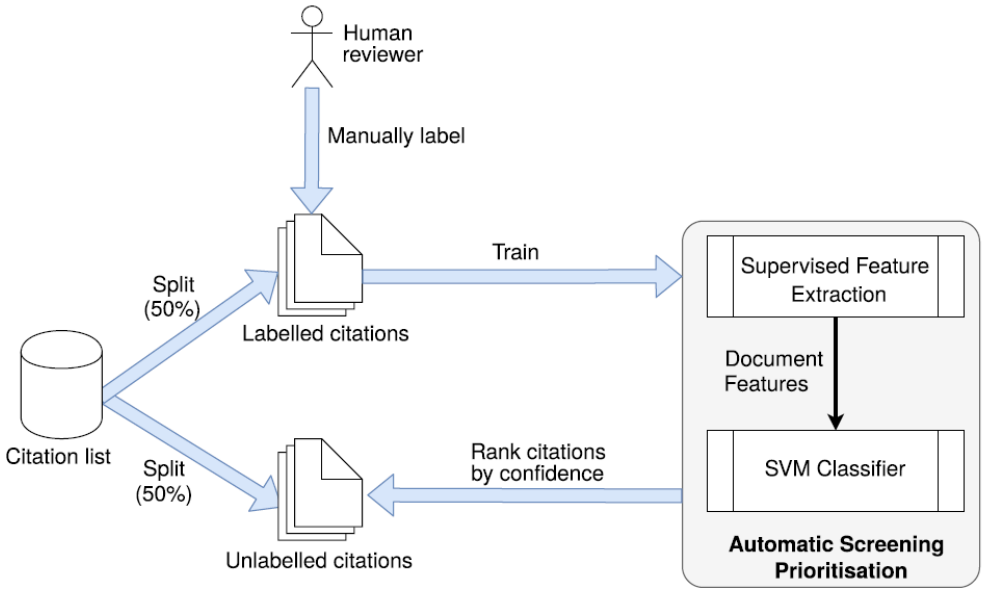

Figure 3. Representation of the active learning approach (Kontonatsios et al., 2020).

A problem in this type of classification is that there is an imbalance of classes, i.e., the number of relevant papers is always much less than the number of irrelevant papers. Therefore, to work with automatic screening, the system designer needs to consider methods to promote class balancing to avoid loss of classification accuracy.

In addition to the supervised and unsupervised learning paradigms, there is a third paradigm known as reinforcement learning (RL). This kind of learning refers to an actor or agent that interacts with its environment and modifies its actions, or control policies, based on stimuli received in response to its actions. This is based on evaluative information from the environment and could be called action-based learning. RL implies a cause-and-effect

relationship between actions and reward or punishment (Lewis & Vrabie, 2009). It implies goal-directed behavior at least insofar as the agent understands reward versus lack of reward or punishment. Reinforcement learning problems have three characteristics that distinguish them from others: (1) the problem is closed-loop; (2) the learner does not have a tutor to teach it what to do, but it should figure out what to do through trial-and-error; and (3) actions influence not only the short-term results but also the long-term ones (Sutton & Barto, 2018). We can mention industry automation, robotics, and self-driving vehicles as examples of this type of problem. The problems encountered in automating systematic reviews do not possess these characteristics. We also could not find applications of reinforcement learning literature in this research area.

3.3 EXTRACTION

The data extraction step is also time-consuming. In this way, the idea of building a model to automate this task is handy. According to Jonnalagadda et al.(2015), there is still no unified framework for extracting SLR data. In particular, NLP techniques have not been fully or partially exploited to automate this task (Aliyu et al., 2018). Despite this, it is possible to define PICO elements, which originated in the medical field, as potentially relevant information to be extracted from scientific papers. This information set can be adapted to other knowledge areas and serve as a basis for information extraction systems. Methods for extracting information from texts can be stratified into four categories: rule-based, traditional machine learning (non-deep learning variations), deep learning, or hybrid approaches (Fu et al., 2020). One particular advantage of using rule-based approaches is that this solution provides reliable results in a timely and low-cost manner, given the benefit of not needing to manually annotate a large number of training examples (Wang et al., 2018). Based on specific tasks, the combination of rules and well-curated dictionaries can lead to promising performance. Creating the ruleset is iterative and requires manual effort (Fu et al., 2020).

According to Fu et al.(2020), advances in processing methods have revitalized interest in statistical machine learning approaches for NLP, for which the non-deep-learning variants are typically referred to as “traditional” machine learning. Although feature engineering can be complex, the ability to process and learn from large document corpora greatly reduces the need to manually review documents and provides the possibility of developing more accurate models. However, in contrast to deep learning methods that learn from the text in the sequential format in which it is stored, traditional machine learning approaches require more human intervention feature engineering.

Deep learning is a subfield of machine learning that focuses on the automatic learning of features in multiple levels of abstract representations (LeCun et al., 2015). According to Fu et al. (2020), the algorithms are largely focused on neural networks such as recurrent neural networks (RNN), convolutional neural networks (CNN), and transformers (Vaswani et al., 2017). Nevertheless, there are a few other niche approaches. In contrast to traditional machine learning paradigms, deep learning minimizes the need to engineer explicit data representations such as bag-of-words or n-grams.

Finally, hybrid approaches combine rule-based approaches and machine learning in the same system, offering the advantages of both and minimizing their respective weaknesses. These two architectures are named either terminal hybrid approaches or supplemental hybrid approaches, depending on how these traditional machine learning approaches have been leveraged. In a terminal hybrid approach, rule-based systems are used for feature extraction. The outputs became features used as input for the ML system, and the ML system is then

a terminal step that selects optimal features. In supplemental hybrid approaches, machine learning approaches are used to patch deficiencies in extracting entities that have poor performance when extracted by purely rule-based approaches (Fu et al., 2020).

Another way to facilitate the extraction of relevant information from scientific papers is by using information visualization. This is considered an emerging multidisciplinary field that comprises visual representations of abstract data to facilitate communication and assist in exploring and analyzing information (Steele & Iliinsky, 2010). The benefits of information visualization applications are related to the exploration of human vision. One of the strongest visualization points is the human ability to visually process information much faster than verbally. It is possible to condense a larger amount of data in a single view, and the visualization process involves the human sense with greater capacity to capture information per unit of time (Few, 2009).

Due to the complex tasks involved in mining textual corpora, information visualization techniques should play a central role in improving the performance of the text analysis process. In visual text mining (VTM), multi-disciplinary approaches are gathered to allow users to understand general structure and local trends in complex sets of documents (Lopes et al., 2007). In an SLR, the information visualization can be used to graphically present documents grouped according to the similarity between their contents. It can also be useful in presenting the citation network by indicating papers that cite and are cited by others. In this way, the author can browse through the information presented and use his/her judgment criteria to select the relevant information.

3.4 Synthesis

Information synthesis is the last potential SLR step for automation, which document summarization can carry out. Summarization can be classified in several ways. Based on the number of documents, single and multi-document summarization are considered two important summarization categories. In the single-document summarization process, the summary of only one document is generated, while in the multi-document summarization, many documents are used to generate the summary. Consequently, multi-document summarization is an extension of single-document summarization. However, summarizing many documents is more complex than summarizing just one document, mainly due to redundancy. Some systems face this problem by initially selecting the sentence at the beginning of the paragraph and then measuring the similarity with the next sentence. If that sentence consists of relevant and new content, only then that sentence is selected (Sarkar, 2010).

Summarization can also be classified as extractive or abstractive (Gambhir & Gupta, 2017). A summary generated by the extractive method is formed by a set of relevant sentences extracted from the original document. The summary size depends on the compression ratio. It is a simple and robust way of summarizing. Scores are assigned to each sentence, and those that receive the highest values are selected to compose the summary. On the other hand, abstractive summarization produces an abstract that includes words and phrases different from those present in the original document. Therefore, this summary consists of ideas and concepts from the original document reinterpreted and presented differently. This method requires extensive natural language processing. Thus, the abstractive summary is much more complex than the extractive. Consequently, due to its greater computational viability, extractive summarization has become a standard in document summarization (Gambhir & Gupta, 2017). According to Gambhir & Gupta (2017), summaries are classified into indicative and informative summaries based on the output style. Indicative summaries tell what the

document is about. They provide information about the document topic. On the other hand, while covering the topics, informative summaries provide the whole information in elaborated form.

The summary task can be supervised or unsupervised. The supervised method requires data for model training, i.e., large amounts of previously labeled data. These systems deal with the problem as a binary classification at the sentence level. That is, the sentences belonging to the summary are classified as positive samples, and the sentences that do not belong to the summary are classified as negative samples (Song et al., 2011; Chali & Hasan, 2012). Several classifiers can perform the classification, such as SVM, decision trees, or neural networks. On the other hand, the unsupervised method does not require data for training, and the summary is generated only based on the target documents. These models apply heuristics to extract the most relevant sentences to generate summaries. Clustering is the technique used in unsupervised systems (Gambhir & Gupta, 2017).

There are several extractive approaches for document summarization. In Gambhir and Gupta's work (2017), it is possible to find several of these approaches, such as statistics, topic-based, graph-based, and speech-based. However, the focus of this work is the summary approaches that involve AI to be applied to systematic reviews. The SLR is inherently a multi-document summarization since it needs to extract information from various scientific papers. Due to its greater viability, it is possible to consider the extractive summarization approach. Therefore, it is considered the extractive, informative, and unsupervised multi-document summarization in the specific case of SLRs. Because in most situations, it is not possible to use previously labeled data to train the models. In this case, according to Alguliev et al. (2013), document summarization, especially of multiple documents, is in its essence a multi-objective optimization problem, i.e., it requires the optimization of more than one objective function. It is expected that a good summary, as a whole, possesses extensive coverage of the key contents presented in the documents, minimal redundancy, and smooth connection among sentences. Huang et al. (2010) consider the four objectives: information coverage, information significance, information redundancy, and text cohesion. Thus, optimization algorithms can be used to perform the creation of multiple-document summaries. Among these, we can highlight Genetic Algorithms (GA) (Neduncheli et al., 2012), Differential Evolution (DE) (Mishra et al., 2021), and Particle Swarm Optimization (PSO), among others (Rautray & Balabantaray, 2017).

Another useful option for the information synthesis step is the natural language generation (NLG). Reiter and Dale (1997, p. 1) state that: "NLG is the subfield of AI and computational linguistics that is concerned with the construction of computer systems that can produce understandable texts in English or other human languages from some underlying non-linguistic representation of information." According to Reiter and Dale (1997), it is possible to say that there are six basic types of activities that need to be carried out from the input data to the final output text in an NLG system:

1. Content determination, which is deciding what information the text will communicate.
2. Discourse planning, which consists of imposing order and structure on the set of messages to be transmitted.
3. Sentence aggregation, which is the process of grouping messages into sentences.
4. Lexicalization, which decides which specific words and phrases should be selected to express the domain concepts and relationships that appear in the messages.
5. Referring-expression generation, which is selecting words or phrases to identify domain entities.
6. Linguistic realization, which consists of applying grammatical rules to produce a final text

that is correct from an orthographic, morphological, and syntactic point of view.

There are several ways to build an NLG system that performs the tasks described above. According to Das and Verma (2020), the simplest approach is to create a module for each task and connect them via a one-way pipeline. However, present a pragmatic architecture that groups the tasks presented in three modules: text planning, sentence planning, and linguistic realization (Figure 4).

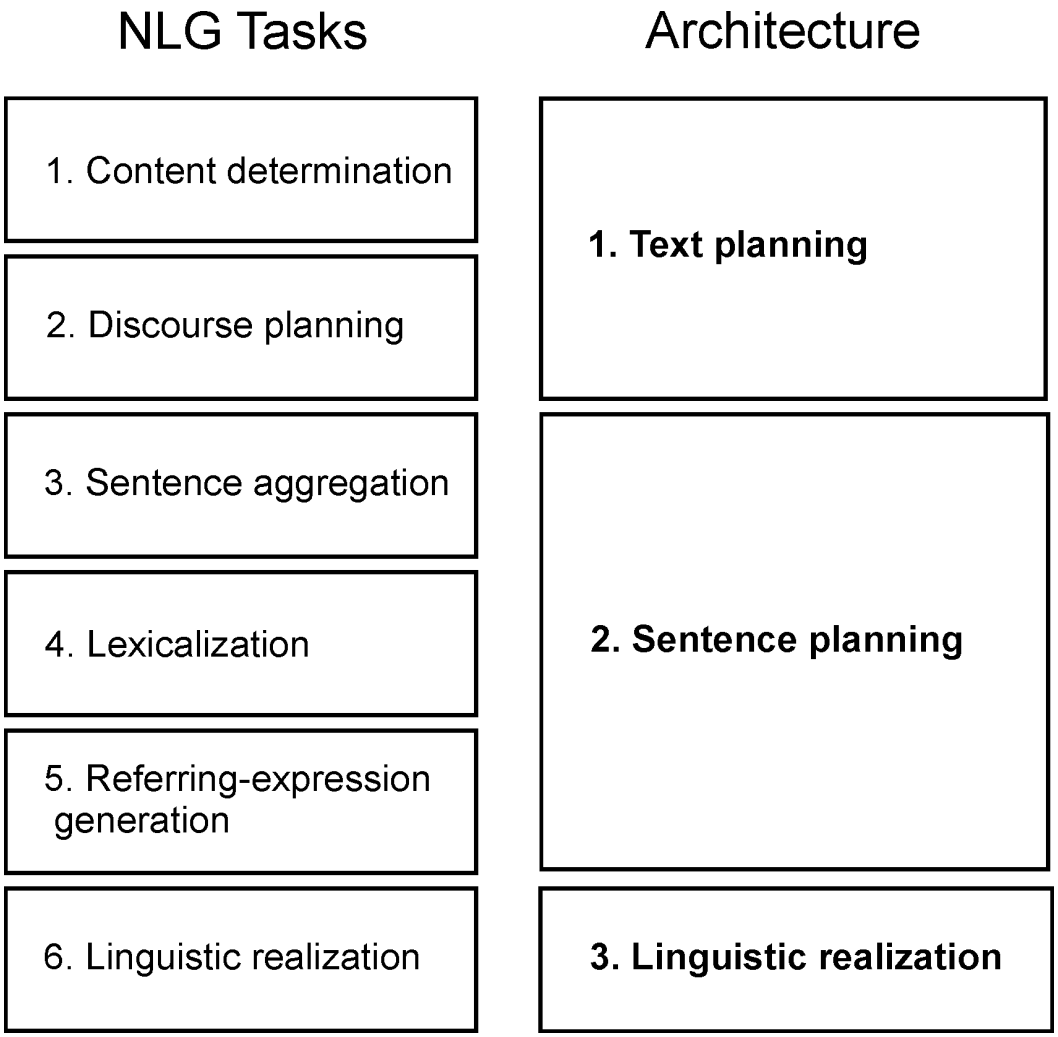

Figure 4. NLG tasks and architecture according to Reiter and Dale (1997).

According to Das and Verma (2020), automated long content generation is a difficult task, as maintaining consistency becomes more challenging with increasing text length. Like RNN (Recurrent Neural Networks) and LSTM (Long Short-Term Memory), neural network architectures are widely used for a content generation due to their ability to learn in the entire textual context. Islam et al. (2019) consider that the LSTM network is a particular type of RNN. A recurrent neural network is an artificial neural network that attempts to model sequence or time-dependent in regular behavior. It is worth emphasizing that neural network architectures are not unique to be used with unsupervised learning. In an SLR, texts from papers that are part of the review can be used as input to train models capable of predicting coherent word sequences to assist the SLR author in writing specific text parts.

4 CONCLUDING REMARKS

Automating systematic literature reviews is a promising field of research since the number of scientific papers published grows every year. This growing total of available texts makes the human work of writing systematic reviews of the scientific literature quite challenging. For this reason, the development of computational tools to help researchers in this task continues to arouse the scientific community's interest. However, it is essential to highlight that, due to the numerous limitations of existing computational techniques, there are still no definitive/ standardized tools to help the automatic creation of systematic literature reviews. We believe that the development of computational models that will reduce the human workload, especially during the operational stages of the SLR, can provide more agility to generating scientific knowledge. Aiming for such a goal and through a scoping review, this work has identified some existing initiatives and brought them together in a framework representing the most common steps and their associated techniques toward the automatic creation of an SLR. Our purpose was to propose an action plan that could be used as a reference by AI researchers.

Although some techniques to facilitate the reviewer's work have been identified, the literature found only presents specific/partial solutions for certain steps in constructing a systematic review. For example, despite being quite handy at some steps, supervised methods face a lack of data for training; consequently, these techniques have less potential for automating SLRs. As for unsupervised methods, these are more promising. Moreover, summarizing, visualizing, and clustering documents are examples of tasks that can help researchers deal with a large number of publications without relying on previously labeled databases for training their models. Document clusters also facilitate the task of automatically creating summaries through specific techniques, such as MDS. With the help of visualization tools, the researcher can achieve better performance in the intermediate tasks of the review, such as identifying main and associated topics, paragraphs or key sections of the documents, excerpts with relevant contributions, as well as details about citations, authors, co-authors, affiliations, among others.

Despite the existing obstacles identified throughout this research, we believe that this is a path that has the potential to alleviate the repetitive work of researchers and direct them toward the tasks of this field more suited to the human intellect: creativity and sensitivity in the analysis of scientific knowledge.

As for future work, the computational implementation of the proposed framework will be carried out. Ideally, this implementation will mostly use unsupervised methods to avoid relying on training data, which remains very scarce. In addition, we intend to use existing algorithms to gather, extract and synthesize the information available in the literature that best suits the work scenario. Our final goal is to achieve a complete solution to automate the operational steps of a systematic literature review. As a subsequent step to this prototype, which is currently under development, we intend to test it in application scenarios from different areas of knowledge, with variations in the technique combinations, and make it available for researchers specialized in these areas to qualitatively evaluate the results obtained.

Conflicts of interest

The authors declare that there is no conflict of interest.

Contribution statement

Writing – original draft, Writing – review &editing: Eugênio Monteiro da Silva Júnior.

Writing– review & editing: Moisés Lima Dutra.

Statement of data consent

The data generated during the development of this study has been included in the manuscript.

REFERENCES

Aliyu, M. B., Iqbal, R., & James, A. (2018). The Canonical Model of Structure for Data Extraction in Systematic Reviews of Scientific Research Articles. *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 264–271. https://doi.org/10.1109/SNAMS.2018.8554896

Ananiadou, S., Rea, B., Okazaki, N., Procter, R., & Thomas, J. (2009). Supporting Systematic Reviews Using Text Mining. *Social Science Computer Review*, *27*(4), 509–523. https://doi.org/10.1177/0894439309332293

Belter, C. W. (2016). Citation analysis as a literature search method for systematic reviews. *Journal of the Association for Information Science and Technology*, *67*(11), 2766–2777. https://doi.org/10.1002/asi.23605

Bullers, K., Howard, A. M., Hanson, A., Kearns, W. D., Orriola, J. J., Polo, R. L., & Sakmar, K. A. (2018). It takes longer than you think: Librarian time spent on systematic review tasks. *Journal of the Medical Library Association*, *106*(2). https://doi.org/10.5195/JMLA.2018.323

Carvallo, A., Parra, D., Lobel, H., & Soto, A. (2020). Automatic document screening of medical literature using word and text embeddings in an active learning setting. *Scientometrics*, *125*(3), 3047–3084. https://doi.org/10.1007/s11192-020-03648-6

Chali, Y., & Hasan, S. A. (2012). Query-focused multi-document summarization: Automatic data annotations and supervised learning approaches. *Natural Language Engineering*, *18*(1), 109–145. https://doi.org/10.1017/S1351324911000167

Clark, J. (2013). Systematic Reviewing. In S. A. R. Doi & G. M. Williams (Eds.), *Methods of Clinical Epidemiology* (pp. 187–211). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37131-8_12

Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P.-Y. (2006). Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. *Journal of the American Medical Informatics Association*, *13*(2), 206–219. https://doi.org/10.1197/jamia.M1929

Das, A., & Verma, R. M. (2020). Can Machines Tell Stories? A Comparative Study of Deep Neural Language Models and Metrics. *IEEE Access*, *8*, 181258–181292. https://doi.org/10.1109/ACCESS.2020.3023421

Davis, D. (2016). A practical overview of how to conduct a systematic review. *Nursing Standard*, *31*(12), 60–71. https://doi.org/10.7748/ns.2016.e10316

Felizardo, K. R., & Carver, J. C. (2020). Automating Systematic Literature Review. In M. Felderer & G. H. Travassos (Eds.), Contemporary Empirical Methods in Software Engineering (pp. 327–355). Springer International Publishing. https://doi.org/10.1007/978-3-030-32489-6_12

Forman, G. (n.d.). *An Extensive Empirical Study of Feature Selection Metrics for Text Classification*. 17.

Fu, S., Chen, D., He, H., Liu, S., Moon, S., Peterson, K. J., Shen, F., Wang, L., Wang, Y., Wen, A., Zhao, Y., Sohn, S., & Liu, H. (2020). Clinical concept extraction: A methodology review. *Journal of Biomedical Informatics*, *109*. https://doi.org/10.1016/j.jbi.2020.103526

Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: A survey. *Artificial Intelligence Review*, *47*(1), 1–66. https://doi.org/10.1007/s10462-016-9475-9

Golder, S., Loke, Y., & McIntosh, H. M. (2008). Poor reporting and inadequate searches were apparent in systematic reviews of adverse effects. *Journal of Clinical Epidemiology*, *61*(5), 440–448. https://doi.org/10.1016/j.jclinepi.2007.06.005

Gough, D., Thomas, J., & Oliver, S. (2012). Clarifying differences between review designs and methods. *Systematic Reviews*, *1*(1), 28. https://doi.org/10.1186/2046-4053-1-28

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies: A typology of reviews, *Maria J. Grant & Andrew Booth*. *Health Information & Libraries Journal*, *26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Guyatt, G., Rennie, D., Meade, M., & Cook, D. (Eds.). (2015). *Users' guides to the medical literature. Essentials of evidence-based clinical practice* (Third edition). McGraw-Hill Education Medical.

Hausner, E., Waffenschmidt, S., Kaiser, T., & Simon, M. (2012). Routine development of objectively derived search strategies. *Systematic Reviews*, *1*(1), 19. https://doi.org/10.1186/2046-4053-1-19

Huang, L., He, Y., Wei, F., & Li, W. (2010). Modeling Document Summarization as Multi-objective Optimization. *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, 382–386. https://doi.org/10.1109/IITSI.2010.80

Islam, Md. S., Sharmin Mousumi, S. S., Abujar, S., & Hossain, S. A. (2019). Sequence-to-sequence Bangla Sentence Generation with LSTM Recurrent Neural Networks. *Procedia Computer Science*, *152*, 51–58. https://doi.org/10.1016/j.procs.2019.05.026

Jonnalagadda, S., & Petitti, D. (2013). A new iterative method to reduce workload in systematic review process. *International Journal of Computational Biology and Drug Design*, *6*(1/2), 5. https://doi.org/10.1504/IJCBDD.2013.052198

Jonnalagadda, S. R., Goyal, P., & Huffman, M. D. (2015). Automating data extraction in systematic reviews: A systematic review. *Systematic Reviews*, *4*(1), 78. https://doi.org/10.1186/s13643-015-0066-7

Khabsa, M., Elmagarmid, A., Ilyas, I., Hammady, H., & Ouzzani, M. (2016). Learning to

identify relevant studies for systematic reviews using random forest and external information. *Machine Learning*, *102*(3), 465–482. https://doi.org/10.1007/s10994-015-5535-7

Kiritchenko, S., de Bruijn, B., Carini, S., Martin, J., & Sim, I. (2010). ExaCT: Automatic extraction of clinical trial characteristics from journal publications. *BMC Medical Informatics and Decision Making*, *10*(1), 56. https://doi.org/10.1186/1472-6947-10-56

Kontonatsios, G., Spencer, S., Matthew, P., & Korkontzelos, I. (2020). Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews. *Expert Systems with Applications: X*, *6*, 100030. https://doi.org/10.1016/j.eswax.2020.100030

Le, Q., & Mikolov, T. (n.d.). *Distributed Representations of Sentences and Documents*. 9.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lewis, F. L., & Vrabie, D. (2009). Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE Circuits and Systems Magazine*, *9*(3), 32–50. https://doi.org/10.1109/MCAS.2009.933854

Lopes, A. A., Pinho, R., Paulovich, F. V., & Minghim, R. (2007). Visual text mining using association rules. *Computers & Graphics*, *31*(3), 316–326. https://doi.org/10.1016/j.cag.2007.01.023

Marcos-Pablos, S., & García-Peñalvo, F. J. (2018). Decision support tools for SLR search string construction. Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality, 660–667. https://doi.org/10.1145/3284179.3284292

Marshall, I. J., & Wallace, B. C. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, *8*(1), 163, s13643-019-1074–1079. https://doi.org/10.1186/s13643-019-1074-9

Mishra, S. K., Saini, N., Saha, S., & Bhattacharyya, P. (2021). Scientific document summarization in multi-objective clustering framework. *Applied Intelligence*. https://doi.org/10.1007/s10489-021-02376-5

Munn, Z., Peters, M.D.J., Stern, C. et al. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 18, 143 (2018). https://doi.org/10.1186/s12874-018-0611-x

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Neduncheli, R., . R. M., & . E. S. (2012). Text Summarization for Multi Documents Using Genetic Algorithm. *International Journal of Soft Computing*, *7*(1), 20–23. https://doi.org/10.3923/ijscomp.2012.20.23

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current

approaches. *Systematic Reviews*, *4*(1), 5. https://doi.org/10.1186/2046-4053-4-5

Pulsiri, N., & Vatananan-Thesenvitz, R. (2018). Improving Systematic Literature Review with Automation and Bibliometrics. 2018 Portland International Conference on Management of Engineering and Technology (PICMET), 1–8. https://doi.org/10.23919/PICMET.2018.8481746

Rautray, R., & Balabantaray, R. C. (2017). Bio-inspired approaches for extractive document summarization: A comparative study. *Karbala International Journal of Modern Science*, *3*(3), 119–130. https://doi.org/10.1016/j.kijoms.2017.06.001

Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, *3*(1), 57–87. https://doi.org/10.1017/S1351324997001502

Ros, R., Bjarnason, E., & Runeson, P. (2017). A Machine Learning Approach for Semi-Automated Search and Selection in Literature Studies. Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering, 118–127. https://doi.org/10.1145/3084226.3084243

Scells, H., & Zuccon, G. (2018). Generating Better Queries for Systematic Reviews. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 475–484. https://doi.org/10.1145/3209978.3210020

Scells, H., Zuccon, G., & Koopman, B. (2021). A comparison of automatic Boolean query formulation for systematic reviews. *Information Retrieval Journal*, *24*(1), 3–28. https://doi.org/10.1007/s10791-020-09381-1

Scells, H., Zuccon, G., Koopman, B., & Clark, J. (2020). Automatic Boolean Query Formulation for Systematic Review Literature Search. *Proceedings of The Web Conference 2020*, 1071–1081. https://doi.org/10.1145/3366423.3380185

Silva Júnior, E. M. da, & Dutra, M. L. (2021). A Roadmap for Composing Automatic Literature Reviews: A Text Mining Approach. In E. Bisset Álvarez (Ed.), *Data and Information in Online Environments* (Vol. 378, pp. 229–239). Springer International Publishing. https://doi.org/10.1007/978-3-030-77417-2_17

Song, W., Cheon Choi, L., Cheol Park, S., & Feng Ding, X. (2011). Fuzzy evolutionary optimization modeling and its applications to unsupervised categorization and extractive summarization. *Expert Systems with Applications*, *38*(8), 9112–9121. https://doi.org/10.1016/j.eswa.2010.12.102

Speckman, R. A., & Friedly, J. L. (2019). Asking Structured, Answerable Clinical Questions Using the Population, Intervention/Comparator, Outcome (PICO) Framework. *PM&R*, *11*(5), 548–553. https://doi.org/10.1002/pmrj.12116

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second edition). The MIT Press.

Thomas, J., McNaught, J., & Ananiadou, S. (2011). *Applications of text mining within systematic reviews*. 14.

Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., & Coiera, E. (2014).

Systematic review automation technologies. *Systematic Reviews*, *3*(1), 74. https://doi.org/10.1186/2046-4053-3-74

van Dinter, R., Tekinerdogan, B., & Catal, C. (2021). Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology*, *136*, 106589. https://doi.org/10.1016/j.infsof.2021.106589

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*. http://arxiv.org/abs/1706.03762

Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., & Schmid, C. H. (2010). Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics*, *11*(1), 55. https://doi.org/10.1186/1471-2105-11-55

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., & Liu, H. (2018). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, *77*, 34–49. https://doi.org/10.1016/j.jbi.2017.11.011