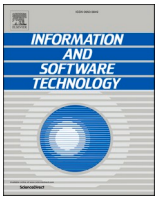
Information and Software Technology 136 (2021) 106589

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof



Automation of systematic literature reviews: A systematic literature review

Raymon van Dinter [a], Bedir Tekinerdogan [a], Cagatay Catal [b,*]

[a] Information Technology Group, Wageningen University & Research, Wageningen, Netherlands
[b] Department of Computer Science and Engineering, Qatar University, Doha, Qatar

ARTICLE INFO

Keywords:
Systematic literature review (SLR)
Automation
Review
Text mining
Machine learning
Natural language processing

ABSTRACT

Context: Systematic Literature Review (SLR) studies aim to identify relevant primary papers, extract the required data, analyze, and synthesize results to gain further and broader insight into the investigated domain. Multiple SLR studies have been conducted in several domains, such as software engineering, medicine, and pharmacy. Conducting an SLR is a time-consuming, laborious, and costly effort. As such, several researchers developed different techniques to automate the SLR process. However, a systematic overview of the current state-of-the-art in SLR automation seems to be lacking.

Objective: This study aims to collect and synthesize the studies that focus on the automation of SLR to pave the way for further research.

Method: A systematic literature review is conducted on published primary studies on the automation of SLR studies, in which 41 primary studies have been analyzed.

Results: This SLR identifies the objectives of automation studies, application domains, automated steps of the SLR, automation techniques, and challenges and solution directions.

Conclusion: According to our study, the leading automated step is the *Selection of Primary Studies*. Although many studies have provided automation approaches for systematic literature reviews, no study has been found to apply automation techniques in the planning and reporting phase. Further research is needed to support the automation of the other activities of the SLR process.

1. Introduction

The number of papers published in academic databases is proliferating. The scientific database ScienceDirect grants access to more than 16 million papers from 2500 journals and provides insights into breakthrough innovations to more than 25 million researchers every month [1]. Pubmed is another search engine maintained by the National Center for Biotechnology Information (NCBI) and contains over 30 million citations and summaries of biomedical literature [2]. Due to the rapid growth of publications in these scientific databases, a timely review and systematic overview of the state-of-the-art in a particular research domain are more challenging.

According to the European Patent Office [3], up to 30% of the R&D investment is wasted due to redeveloping existing literature information. Also, pertinent literature is critical for proposals submitted to the funding agencies such as the National Science Foundation (NSF) and National Institutes of Health (NIH), and failing to provide the pertinent literature causes the fail of the research proposal. Traditional survey/review articles do not cover all the published papers in a particular domain systematically, and new project ideas based on these traditional review papers might sometimes be misleading. Different techniques exist in the literature to address these concerns, and one of them is conducting a Systematic Literature Review (SLR) study.

An SLR is a means of identifying, evaluating, and synthesizing all available research relevant to a particular research question, topic area, or phenomenon of interest [4]. An SLR's goal is a trustworthy method to gain clear, reasonable, and unbiased information on a research topic [5]. Kitchenham and Charters [4] proposed a Systematic Literature Review (SLR) method for the software engineering domain in 2007. This process provides a robust framework to find relevant literature systematically with low bias and high rigor. Since 2007, systematic reviews as proposed by Kitchenham and Charters have been widely used in the software engineering field. However, the collection, extraction, and synthesizing of the required data for systematic reviews are known to be highly manual, error-prone, and labor-intensive tasks in the software engineering domain and other fields such as medicine [6]. The time from the

* Corresponding auhor.
E-mail addresses: raymon.vandinter@wur.nl (R. van Dinter), bedir.tekinerdogan@wur.nl (B. Tekinerdogan), ccatal@qu.edu.qa (C. Catal).

https://doi.org/10.1016/j.infsof.2021.106589
Received 24 October 2020; Received in revised form 22 March 2021; Accepted 27 March 2021
Available online 4 April 2021

last search to publication takes commonly over 1 year for an SLR study, and for a primary study, it takes 2.5 - 6.5 years before it is incorporated in an SLR study [7,8]. Additionally, 23% of all SLR studies are outdated within 2 years of publication, as reviewers failed to incorporate novel evidence on their subject of interest [8,9].

The sub-branch of Artificial Intelligence (AI) called Natural Language Processing (NLP) is nowadays used increasingly to gain insights from these huge volumes of textual data. NLP is a research area that aims to understand and manipulate natural language text or speech [10]. Natural language refers to the language used in everyday communication, e.g., human language and studies on NLP can vary from counting words to generating sentences or classification of textual data. NLP for Robotic Process Automation is an important tool to improve operational efficiency across all industries, including the academic sector, as most academics need to process large amounts of documents during research.

Several researchers recently developed different approaches to automate steps of the SLR process by using machine learning and NLP techniques. This paper performs a systematic literature review (SLR) on the automation of SLR studies to collect and summarize the current state-of-the-art that is needed to define a framework for further research activities. Table 1 lists the steps in the systematic review process, as proposed by [4]. Synonyms that were used in the literature were noted for consistency. As shown in this table, there are 12 steps that researchers must follow during an SLR study, which are in practice time-consuming and expensive, considering the project budget.

We have found 41 studies that focused on the automation of one or more selected steps in the SLR process, mainly for the Software Engineering and Medical domain. However, no synthesis of these has been reported to provide a comprehensive overview of the current state of SLRs' automation support. Hence, in this paper, we aim to identify and synthesize the current studies that have focused on the automation of SLR and herewith identify the objectives, the application domains, the automated steps of the SLR, the automation approaches, and the corresponding challenges and solution directions. The research method that we have applied for this is an SLR itself, in which we reviewed 41 research papers.

This SLR study is the most up-to-date study on SLR studies' automation and covers all the related literature to the best of our knowledge. This study paves the way for further research on the use of automation techniques for different SLR process stages. It is considered that new technological advancements in big data, deep learning, and text mining fields provide many opportunities for the full automation of SLR studies; however, there are many challenges to address.

Table 1
Steps in the systematic review process as proposed by Kitchenham and Charters [4].

ID	Category	Step	Synonyms
SLR1	Need for a review	Commissioning a review	
SLR2		Specifying the research question(s)	
SLR3		Developing a review protocol	
SLR4		Evaluating the review protocol	
SLR5	Conducting the review	Identification of research	Literature Search, Search String Development
SLR6		Selection of primary studies	Citation Screening
SLR7		Study quality assessment	Selection Review
SLR8		Data extraction and monitoring	
SLR9		Data synthesis	
SLR10	Reporting the review	Specifying dissemination mechanisms	
SLR11		Formatting the main report	
SLR12		Evaluating the report	

The following sections are organized as follows: Section 2 presents the related work and systematic literature review approach. Section 3 explains the research methodology. Section 4 presents the results. Section 5 shows the discussion. Section 6 discusses the conclusion and future work.

2. Related work

James Lind is the first medical researcher to conduct an SLR [11]. In his article *Treatise of the Scurvy* (1753), he conducted systematic clinical trials of potential cures for scurvy-trials in which oranges and lemons came out as decisive winners [11]. From that point, SLR became an extensively used practice to support evidence-based medicine. The success of SLR in evidence-based medicine triggered various other research areas to adopt similar SLR approaches [5]. In 2007, Kitchenham attempted to construct guidelines for performing SLR that satisfied the needs of software engineering researchers [4]. Since this moment, software engineering researchers widely use the SLR method to conduct unbiased research.

Nearly all studies found acknowledge that their purpose is to cut down the cost for systematic reviews. Furthermore, researchers aim to improve the SLR process by maximizing precision while maintaining a high recall, as the current approach often lacks a high precision. In an SLR study on the *Selection of Primary Studies* (SLR6) automation, most citations suggested that a workload saving of 30- 70% can be achieved, accompanied by the loss of 5% of relevant citations (i.e., a 95% recall) [12]. Researchers are also attempting to reduce human error since most of *Conducting the Review* category steps are highly repetitive [6]. The articles by primary authors K.R. Felizardo and J.C. Maldonado [13–16] attempted to shift the SLR process from a repetitive, error-prone approach to Visual Text Mining, which enables users to find related articles through unsupervised learning. The major downside is that users must be familiar with machine learning and statistics.

While conducting this study, we also identified related research regarding the automation of systematic reviews. In Table 2, we have listed 15 related studies that we have found. In this section, we explain each of these articles in more detail.

In 2011, Thomas et al. [17] published a report that lists the application of text mining techniques to automate the systematic literature review process. In total, we have found 5 studies that reported text mining techniques and tools to automate a – part of – the systematic review process [12,17–20]. Tsafnat et al. [21] describe each step in the systematic review process, its automation potential, and current tools that have already been developed. Jonnalagadda et al. [8] collected 26 published reports and lists their methods to automate data extraction for various data points. These data points vary from PICO to the number of trial participants extracted. O'Mara-Eves et al. [12] and Olorisade et al. [20] performed a systematic review on text mining in the automation of the *Selection of Primary Studies* (SLR6) step. In their study, they also describe that: *"Given that an experienced reviewer can take between 30 s and several minutes to evaluate a citation [22], the work involved in screening 10,000 citations is considerable (and the screening burden in some reviews is considerably higher than this)"*. Paynter et al. [23] published a report that lists the application of text mining techniques to automate the selection, extraction, and updating steps of the systematic review process. Feng et al. [18] conducted a systematic review to find evidence regarding text mining techniques currently used in the systematic literature review process. In their study, Shakeel et al. [24] highlight all threats that one could find when creating their automation tool. Jaspers et al. [25] published an in-depth report that discusses useful machine learning techniques to automate systematic reviews. Besides, Beller et al. [19] listed automation tools that can be used to speed up the systematic literature review process and set 8 guidelines for creating a systematic review tool. O'Connor et al. [26] state barriers to why

Table 2
Related studies.

No.	Title	SLR steps	Reference
1	A critical analysis of studies that address the use of text mining for citation screening in systematic reviews	An SLR on primary study selection. In related work, they state that Jonnalagadda et al. have found 20 articles in the data extraction automation of SLR	Olorisade et al. [20]
2	(Automated) literature analysis: threats and experiences	Highlight all threats that reviewers could encounter when automating the SLR process	Shakeel et al. [24]
3	Text-Mining Techniques and Tools for Systematic Literature Reviews: A Systematic Literature Review	Highlight text mining techniques that are currently used in SLRs	Feng et al. [18]
4	EPC Methods: An Exploration of the Use of Text-Mining Software in Systematic Reviews AHRQ Methods for Effective Health Care	A report that lists the application of TM techniques to automate the selection, extraction, and update steps of the SLR process	Paynter et al. [23]
5	A full systematic review was completed in 2 weeks using automation tools: a case study	Completed an SLR in 2 weeks using multiple tools	Clark et al. [30]
6	The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials	Calculated the economic cost and total time estimate and called for the development of automated tools for SLRs	Michelson and Reuter [27]
7	Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR)	List tools that can be used, and set 8 guidelines for automating SLRs	Beller et al. [19]
8	Automating data extraction in systematic reviews: a systematic review	Lists methods to use data extraction for various data items found from 26 published reports	Jonnalagadda et al. [8]
9	Toward systematic review automation: a practical guide to using machine learning tools in research synthesis	Lists tools that are useful for systematic reviews	Marshall and Wallace [28]
10	A question of trust: can we build an evidence base to gain trust in systematic review automation technologies?	States barriers why people don't use systematic review automation tools	O'Connor et al. [26]
11	Using text mining for study identification in systematic reviews: a systematic review of current approaches	SLR on text mining in the automation of SLR6	O'Mara-Eves et al. [12]
12	Systematic review automation technologies	Describe each step in the SLR process, its automation potential, and current tools	Tsafnat et al. [21]
13	Machine learning techniques for the automation of literature reviews and systematic reviews in EFSA	A report that discusses possible machine learning techniques for the automation of SLRs	Jaspers et al. [25]
14	Applications of text mining within systematic reviews	A report that lists the application of TM techniques to automate the SLR process	Thomas et al. [17]
15	Usage of automation tools in systematic reviews	A survey that concludes not many researchers are using an SLR tool	van Altena et al. [29]

researchers don't use systematic review automation tools to speed up the process. Michelson and Reuter [27] calculated the financial cost and provided a total time estimate to complete a systematic review. A systematic review was found to take 1.72 years for a single scientific reviewer. A single review would cost $141,194.80. On average, the total cost of all SLRs per year to each of the ten major academic institution amounts to $18,660,304.77, and for each pharmaceutical company is $16,761,234.71. They also called for action to develop automation tools to speed up since the high time and cost of a systematic review may pose a barrier to their consistent application to assess the promise of studies carefully [27].

Marshall and Wallace [28] list tools that are useful for systematic reviews. They also state that a systematic review is estimated to cost around 67 weeks to produce from start to end [28]. Van Altena et al. [29] conducted a survey that concludes that not many researchers use a systematic review tool. When tools were used, participants often learn about them from their environments, such as colleagues, peers, or organization. Tools were often chosen based on user experience, either by experience or from colleagues or peers. Last, in this year, Clark et al. [30] published a study that announces that they have completed a systematic review – from start to end – within two weeks. They accomplished this outstanding result using automation tools for each of the steps of the systematic review process.

To our knowledge, Feng et al. [18] is the only systematic literature review on the automation of systematic literature reviews with a focus on all systematic literature review steps and Text Mining. However, it is outdated since the paper's timespan was just until 31st December 2014. Furthermore, the paper does not thoroughly discuss NLP preprocessing and representation techniques and does not split all results into the systematic review steps proposed by Kitchenham and Charters [4]. Therefore, this study represents the first systematic literature review on the automation of systematic literature reviews focusing on all systematic literature review steps and machine learning and natural language processing techniques.

3. Research methodology

As we have seen in the Related Work, there is no up-to-date overview of systematic literature studies' current automation techniques. As the field of Artificial Intelligence is developing rapidly (e.g., embeddings in NLP and Neural Networks in deep learning), we aim to provide an overview of the current trends of these techniques in the automation of systematic literature reviews. Furthermore, we see that most other secondary studies focus on automation techniques as a part of the systematic literature review process. This study aims to provide an overview of all steps of the process. By doing so, this study can act as an accelerator for future primary studies in this domain. To gather all relevant primary studies, we perform a systematic literature review.

The systematic review follows the guidelines reported by Kitchenham and Charters [4]. Kitchenham and Charters [4] describe that a predefined, strictly followed protocol reduces bias among researchers and increases rigor and reproducibility. Therefore, we constructed a review protocol before conducting the review, based on the guidelines from Kitchenham and Charters [4], Ali et al. [31], and Gurbuz and Tekinerdogan [5]. Fig. 1 shows the adopted review protocol.

Section 3.1 provides a table of examined research questions for the systematic literature review. Sections 3.1.1 to 3.1.4 describe the systematic literature review's scope, search methods, and used search strings. Section 3.2 describes the inclusion/exclusion criteria of the retrieved literature, and Section 3.3 adds quality assessment criteria to the retrieved literature.

3.1. Research questions

We aim to find all relevant information regarding the development of automation systems for SLR studies from a Software Engineering

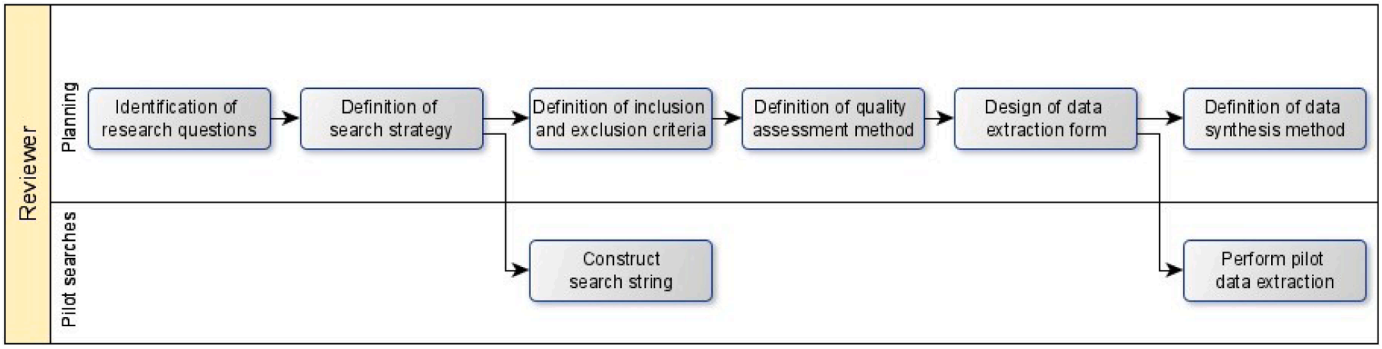


Fig. 1. Review protocol adapted from Ali et al. [31] and Gurbuz and Tekinerdogan [5].

perspective. When building a model for text classification, developers should always know (1) its objective and application domain, (2) where to get relevant data from, (3) what data to evaluate the model on, (4) which algorithms/techniques to use for the model, and (5) which challenges to expect. We constructed a table of research questions to which the systematic literature review should provide answers. Research Questions are given in Table 3.

3.1.1. Search strategy

The purpose of this review is to collect as many related primary studies on the automation of systematic literature reviews as possible (a high *recall*) while also keeping out irrelevant studies (a high *precision*). A well-developed search strategy is essential to achieve a high recall as well as a high precision. This section elaborates on the review's search strategy, which consists of the scope (publication period and publication venue), search method (automatic search or manual search), and search string.

3.1.2. Search scope

The systematic literature review scope consists of two dimensions, namely, publication period and publication venues. Concerning the publication period, this systematic literature review includes papers published from January 2000 to June 2020. The year 2000 was taken as the starting point since, from this year, research in text mining started growing. We conducted the literature search in June 2020, accepting only studies until that point. From a publication venue perspective, we searched for studies in the following databases; ScienceDirect, PubMed, ACM Digital Library, IEEE Xplore, Wiley Interscience, and Springer Link. Publication venues were required to have an academic discipline in Software Engineering or Medicine.

Table 3
Research questions.

No.	Research Question (RQ)
RQ1	What is the objective of automation?
RQ2	On which application domains are automated systematic literature review tools evaluated?
RQ3	Which are the used databases for the automation of systematic literature reviews?
RQ3a	Which databases have been used for automated metadata extraction?
RQ3b	Which databases have been used for the retrieval of studies to evaluate a tool?
RQ4	Which are the automated steps of systematic literature reviews?
RQ5	Which are the adopted machine learning-based automation techniques?
RQ5a	Which parts of the documents are used to automate systematic reviews?
RQ5b	Which NLP preprocessing and representation techniques are used to automate systematic reviews?
RQ5c	Which machine learning techniques, tasks, evaluation approach, evaluation metrics, and algorithms are used to automate systematic reviews?
RQ6	What are the open challenges and solution directions?

3.1.3. Search method

For this systematic study, we used an automated literature search. Automatic search refers to scanning for the search strings of electronic databases. For each publishing location, we used an automated scan. We used an automatic search for each publication venue. We also supported our search with snowballing (i.e., backward snowballing and forward snowballing), which means that we also utilized from manual search.

3.1.4. Search string

To find relevant articles regarding the automation of SLR studies, which are often about text or data mining, we have constructed a search string containing Boolean operators. The search strings were adjusted iteratively to optimize the literature search's precision and recall through several pilot searches. Additionally, the search strings were adapted to fit each database. As a result, we used the following general search string:

("Automation" OR "Automating" OR "Automated" OR "Automatic" OR "Automates" OR "Mining")

AND

("Systematic review" OR "Systematic Literature Review")

Table 4 lists the results of the search query. A total of 1291 papers were found via the automated search. ScienceDirect was the source with the most results ($n = 493$). Adversely, PubMed was the source with the lowest amount of results ($n = 19$).

3.2. Study selection criteria

The search string purposely has an extensive reach because we did not want to skip any study of concern. To add, the terms "Systematic review" and "Systematic literature review" provided a vast number of secondary studies. We identified the relevant studies using the study selection criteria provided in Table 5.

The selection criteria were applied by reading the title and abstract, which reduced the number of included studies to 59. The second column of Table 4 lists the number of articles that passed the selection criteria.

Table 4
Overview of search results and study selection.

Source	After Automated Search	After Selection Criteria (Abstract)	After Selection Criteria (Full-Text)	After Quality Assessment
ScienceDirect	493	23	19	19
ACM Digital Library	97	14	10	10
IEEE Xplore	348	12	8	8
Springer	220	5	3	3
Wiley	114	4	1	1
PubMed	19	1	0	0
Total	1291	59	41	41

Table 5
Study selection criteria.

No.	Criterion
EC1	Papers without full text available
EC2	Duplicate publications
EC3	Papers not written in English
EC4	Papers that do not relate to the automation of systematic literature reviews
EC5	Papers that do not discuss the automation of systematic literature reviews using Machine Learning and/or Natural Language Processing techniques
EC6	Studies that do not present any empirical result

Finally, we retrieved and read the entire study and applied the selection criteria, reducing the number of papers to 41.

3.3. Study quality assessment

In addition to the exclusion criteria, the quality of the included literature was assessed as well. Quality criteria have been derived to determine if factors could bias study results. Table 6 shows the quality criteria. While developing the quality assessment, the summary quality checklist for quantitative studies and qualitative studies has been adopted, as proposed by Kitchenham and Charters [4] and Gurbuz and Tekinerdogan [5]. We chose the study quality assessment criteria based on their impact on the quality of this SLR.

While reading each study's full text, points were granted to each of the eight assessment criteria. These points were granted based on a scale from 1 to 0. As Tummers, Kassahun and Tekinerdogan [32] describe in their study, a full point should be provided for Q1 if the study's goal was specified in the introduction (expected place), and no point (0) should be provided if the study's intent was not mentioned in the report. A half-point (0.5) should be given if the objective was stated vaguely, or not at the expected location. Studies with a score lower than 4 out of 8 were excluded. As a result, studies with a higher score were maintained to keep only high-quality input for our study. However, there were no studies found with such a low score, as shown in Fig. 2. Therefore, the number of included articles remained at 41 (Table 4, fourth column).

3.4. Data extraction

With our data extraction form, we read the full text of the 41 primary studies and extracted the essential data for our study. We first established a base extraction form using the research questions from Table 3. Afterward, we performed a pilot extraction to update the data extraction form and iteratively updated the data extraction form by adding more papers. Following several iterations of selecting a limited number of studies and changing the data extraction process, we arrived at the final form, as seen in Table 7. In addition to the seven data extraction elements, this form also includes general metadata such as year of publication and title. This data form was implemented in MS Excel, allowing further data synthesis to identify patterns.

Table 6
Quality checklist.

No.	Question
Q1	Are the aims of the study clearly stated?
Q2	Are the scope and context and experimental design of the study clearly defined?
Q3	Are the variables in the study likely to be valid and reliable?
Q4	Is the research process documented adequately?
Q5	Are all the study questions answered?
Q6	Are the negative findings presented?
Q7	Are the main findings stated clearly? Regarding creditability, validity, and reliability?
Q8	Do the conclusions relate to the aim of the purpose of the study? Are they reliable?

3.5. Data synthesis

Data synthesis is how the extracted data from SLR8 is gathered and interpreted appropriately for answering the research questions of a systematic literature review. Since studies identify their objectives, steps, and algorithms with slightly different names, we first synthesized synonyms. We achieved this by defining core terms, which enabled us to gain insights into data patterns.

4. Results

In the results section, we first describe the main statistics of the 41 primary studies we found. Afterward, we present the results corresponding to each research question.

4.1. Main statistics

Table 8 lists the 41 primary studies that we included in this review. The publication year of these studies ranges from 2006 to 2020. The year-wise distribution can be seen in Fig. 4. Cohen, Hersh, Peterson and Yen [33] were the first to automate a part of the systematic review, specifically the *Selection of Primary Studies* (SLR6).

As shown in Fig. 3, most studies were published by J.C. Maldonado and S. Ananiadou. Maldonado often collaborated with K. R. Felizardo, and Ananiadou often collaborated with G. Kontonatsios and J. Thomas. Also, S. Jonnalagadda, D. D. A. Bui, and G. Del Fiol collaborated for publications. With 219 citations, the article of Cohen et al. [33] was the article most cited. To add, it was also the oldest article included, being published in 2006. The year 2016 had an exceptionally high number of publications in this field ($n = 8$). Out of these publications, two were from the same authors who aimed at the automation of *Data Extraction* (SLR8) in the systematic review process [34,35].

As shown in Table 4, ScienceDirect and ACM Digital Library were the most popular databases for primary studies, with 19 and 10 studies directly found. The most famous publication channel is the *Journal of Biomedical Informatics,* with 9 primary studies, as listed in Table 9. This journal is classified as high-influence due to its 5-year impact factor of 3.765 in 2019 [70]. The 5-year impact factor indicates the average number of times articles from a journal have been cited in the past five years. The second most popular publication channel is a conference named *Evaluation and Assessment in Software Engineering* with three primary studies: 2 are from ACM, and 1 is from IEEE.

When looking at the study type, 2% of studies were book chapters, 38% were conference papers, and over 60% of the studies were categorized as journal articles, indicating that most studies in this systematic review are peer-reviewed.

4.2. RQ-1: objective of automation

Nearly all studies acknowledge that the primary purpose of automating systematic reviews is to cut down the cost of systematic reviews. Furthermore, we observed that each study had one or more sub-objectives, which are further categorized. The categorized objectives, and their corresponding number of studies, are shown in Fig. 5. Besides, we incorporated studies that had the automation of a systematic review step as their objective. The distribution of numbers of goals relating to these steps' automation is consistent with the findings in Section 3.4.5. Not all studies that automated a step have also discussed it as an objective, as total numbers do not match.

The articles with primary authors K.R. Felizardo and J.C. Maldonado [13–16] fully account for the objective of Incorporate Visual Text Mining in SLR. Their studies describe the development of a tool that uses Visual Text Mining (VTM) to visually perform the Selection of Primary Studies and Quality Assessment steps (SLR6 and SLR 7). Two articles did not provide any objective regarding the systematic review process automation, as they both were systematic reviews using text mining [65,

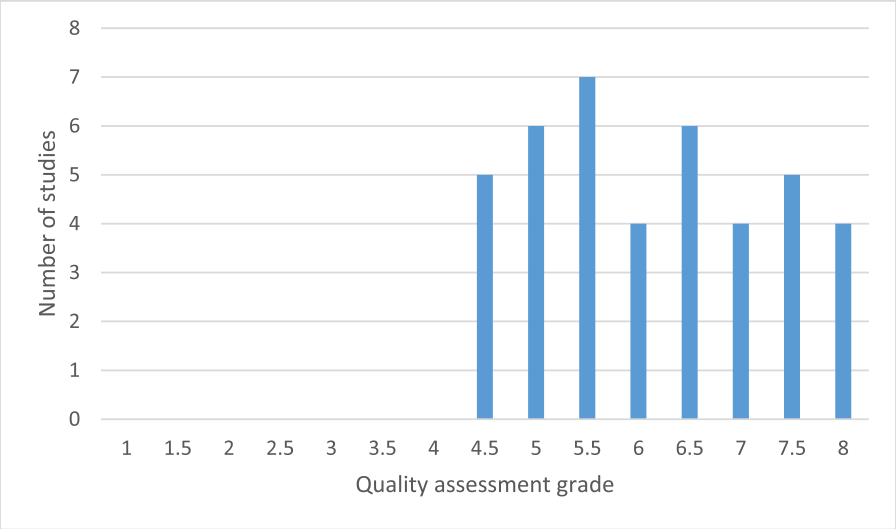


Fig. 2. Histogram of the quality assessment grades.

Table 7
The data extraction form.

No.	Extraction Element
1	**ID**
2	**Title**
3	**Passed inclusion criteria**
4	**Date of extraction**
5	**Year of publication**
6	**Authors**
7	**Repository of extraction**
8	**Publication title**
9	**Type**
10	**Volume**
11	**Issue**
12	**Pages**
13	**DOI**
14	**URL**
15	**Keywords**
16	**Abstract**
17	**Times cited on extraction**
1	Objective
2	Application domain
3	Databases used
4	Document features
5	Automated steps
6	Level of automation per step
7	Open challenges and solution direction

69].

4.3. RQ-2: application domains

Studies often evaluated their model or tool using data from an already completed systematic review from a specific application domain (e.g., *Evidence-Based Medicine*). The application domains of these completed systematic reviews are listed in Fig. 6. If no evaluation set with a specific application domain was mentioned, either the application domain of interest was inserted or *Not Mentioned*. The sole application domains used for model evaluation are *Software Engineering* (40%) and *Medicine* (60%). Specialization domains within the medical field (e.g., *Pharmacy* and *Public health*) were noted separately.

4.4. RQ-3: databases

We have split this research question into two sub-questions, as listed in Table 3. The results of each sub-question have been presented to answer the main research question fully.

4.4.1. Which databases have been used for automated metadata extraction?

For automated metadata extraction, we have found just one study. In this study, González-Toral, Freire, Gualán and Saquicela [40] describe a Python-based algorithm that automates metadata extraction from articles published by IEEE, ACM, Springer, Scopus, and Semantic Scholar. They extracted metadata through each repositories' REST API using a customized search query, consisting of Boolean operators, filters, and specific metadata. Since ACM and Semantic Scholar do not have an API that suits this challenge, González-Toral, Freire, Gualán and Saquicela [40] developed a web-scraping algorithm using the Selenium Python library. The collected metadata M was stored in the following scheme:

$$M = \{title;\ abstract;\ keywords;\ year;\ authors;\ doi\}$$

Once the metadata was stored, the title, abstract, and keywords were processed using NLP and unsupervised machine learning techniques to identify which papers are the most relevant for citation screening.

4.4.2. Which databases have been used for the retrieval of studies to evaluate a tool?

We found that 26 articles had a form of article extraction from a database to evaluate their tool. Fig. 7 highlights databases that included more than one study. As shown in the figure below, there is a mix between Software Engineering and Medical databases. A high number of databases are from the Medical domain, as it is the main SLR domain. The publication venue and used databases do not seem to relate to each other, as ScienceDirect does not have the highest number of applications, and PubMed is used relatively often.

4.5. RQ-4: automated steps of the SLR

As described in Table 1, all automated steps are from the *Conduct of review* category (SLR5–9). Most studies were about *Selecting Primary Studies* (SLR6); the least studies were about the *Study Quality Assessment* (SLR7). Three studies automated steps SLR5 and SLR6 at the same time.

The *Selection of primary studies* (SLR6) was automated most often, as researchers agree that this is one of the most time-consuming steps [58, 67,71].

The second most automated step is *Identifying research* (SLR5), with its sub-step *Formulating the search query* being the only automated. Several studies [39,45,46] describe that formulating a search query for a

Table 8
The 41 primary studies used as input for the systematic review.

ID	Title	Year	Reference
1	A clustering approach for topic filtering within systematic literature reviews	2020	Weißer et al. [36]
2	A hybrid feature selection rule measure and its application to systematic review	2016	Ouhbi et al. [37]
3	A Machine Learning Approach for Semi-Automated Search and Selection in Literature Studies	2017	Ros et al. [38]
4	A method to support search string building in systematic literature reviews through visual text mining	2015	Mergel et al. [39]
5	A ranking-based approach for supporting the initial selection of primary studies in a Systematic Literature Review	2019	González-Toral et al. [40]
6	A semi-supervised approach using label propagation to support citation screening	2017	Kontonatsios et al. [41]
7	A visual analysis approach to update systematic reviews	2014	Felizardo et al. [14]
8	A visual analysis approach to validate the selection review of primary studies in systematic reviews	2012	Felizardo et al. [13]
9	A Visual Text Mining approach for Systematic Reviews	2007	Malheiros et al. [16]
10	Active learning for biomedical citation screening	2010	Wallace et al. [42]
11	Advanced analytics for the automation of medical systematic reviews	2016	Timsina et al. [43]
12	An SVM-based high-quality article classifier for systematic reviews	2014	Kim and Choi [44]
13	Automatic Boolean Query Formulation for Systematic Review Literature Search	2020	Scells et al. [45]
14	Automatic Boolean Query Refinement for Systematic Review Literature Search	2019	Scells et al. [46]
15	Automatic endpoint detection to support the systematic review process	2015	Blake and Lucic [47]
16	Automatic text classification to support systematic reviews in medicine	2014	García Adeva et al. [48]
17	Automatically finding relevant citations for clinical guideline development	2015	Bui et al. [49]
18	Automation in systematic, scoping and rapid reviews by an NLP toolkit: a case study in enhanced living environments	2019	Zdravevski et al. [50]
19	Building systematic reviews using automatic text classification techniques	2010	Frunza et al. [51]
20	Cross-Topic Learning for Work Prioritization in Systematic Review Creation and Update	2009	Cohen et al. [52]
21	Data Sampling and Supervised Learning for HIV Literature Screening	2016	Almeida et al. [53]
22	Discriminating between empirical studies and nonempirical works using automated text classification	2018	Langlois et al. [54]
23	Enhancing academic literature review through relevance recommendation: Using bibliometric and text-based features for classification	2016	Rúbio and Gulo [55]
24	Exploiting the systematic review protocol for classification of medical abstracts	2011	Bui et al. [49]
25	Extracting PICO sentences from clinical trial reports using supervised distant supervision	2016	Wallace et al. [56]
26	Extractive text summarization system to aid data extraction from full text in systematic review development	2016	Bui et al. [34]
27	Linked data approach for selection process automation in systematic reviews	2011	Tomassetti et al. [57]
28	Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error	2019	Bannach-Brown et al. [58]
29	Novel text analytics approach to identify relevant literature for human health risk assessments: A pilot study with health effects of in utero exposures	2020	Cawley et al. [59]
30	PDF text classification to leverage information extraction from publication reports	2016	Bui et al. [35]
31	Reducing systematic review workload through certainty-based screening	2014	Miwa et al. [60]
32	Reducing Workload in Systematic Review Preparation Using Automated Citation Classification	2006	Cohen et al. [33]
33	Screening nonrandomized studies for medical systematic reviews: A comparative study of classifiers	2012	Bekhuis and Demner-Fushman [61]
34	The Canonical Model of Structure for Data Extraction in Systematic Reviews of Scientific Research Articles	2018	Aliyu et al. [62]
35	The use of bibliography enriched features for automatic citation screening	2019	Olorisade et al. [63]
36	Topic detection using paragraph vectors to support active learning in systematic reviews	2016	Hashimoto et al. [64]
37	Twitter and Research: A Systematic Literature Review Through Text Mining	2020	Karami et al. [65]
38	Using a Neural Network-based Feature Extraction Method to Facilitate Citation Screening for Systematic Reviews	2020	Kontonatsios et al. [66]
39	Using rule-based classifiers in systematic reviews: a semantic class association rules approach	2015	Sellak et al. [67]
40	Using Visual Text Mining to Support the Study Selection Activity in Systematic Literature Reviews	2011	Felizardo et al. [68]
41	Whole field tendencies in transcranial magnetic stimulation: A systematic review with data and text mining	2011	Dias et al. [69]

systematic review is a challenging task, as researchers want to keep as many relevant studies as possible (high recall) while excluding as many irrelevant studies as possible (high precision) [72–74]. Powerful search queries take several weeks, if not months, to be developed according to the approach mentioned above [45,75,76]. As a result, systematic reviews can take several months to complete and cost upwards of a quarter of a million dollars [45,46,77]. The screening phase attributes to most of these costs, where there are typically many false positives (low precision), causing a significant impact on the time spent on the study. A significant impact on both the total cost of a review and the time required to produce a review can be achieved with even small increases in precision [33,46,78].

In their study, Felizardo, Andery, Paulovich, Minghim and Maldonado [13] automated the *evaluation of the selection of primary studies*, categorized as the only study performing a *Study quality assessment* (SLR7). They based their study and methodology on the procedures as proposed by Kitchenham [79]. Felizardo, Andery, Paulovich, Minghim and Maldonado [13] describe that in these procedures, *Conducting the Review* part has two steps; *Selection execution* and *Information extraction*. The *Selection execution* step also has three sub-steps, where the last one, the *Selection review* step (i.e., the *Study quality assessment* step, SLR7), is their focus of work. They mention that if necessary, reviewers perform this step based on quality criteria to ensure that relevant studies are not initially eliminated.

At last, the *Data extraction and monitoring* step (SLR8) has been automated in five studies. The rationale to automate this step is that the data extraction step is usually highly manual [62,80]. Research has found a high prevalence of errors in the manual data extraction process due to human factors such as time and resource constraints, inconsistency, and tediousness-induced mistakes [34,81,82].

4.6. RQ-5: automation techniques

We have split this research question into two sub-questions, as listed

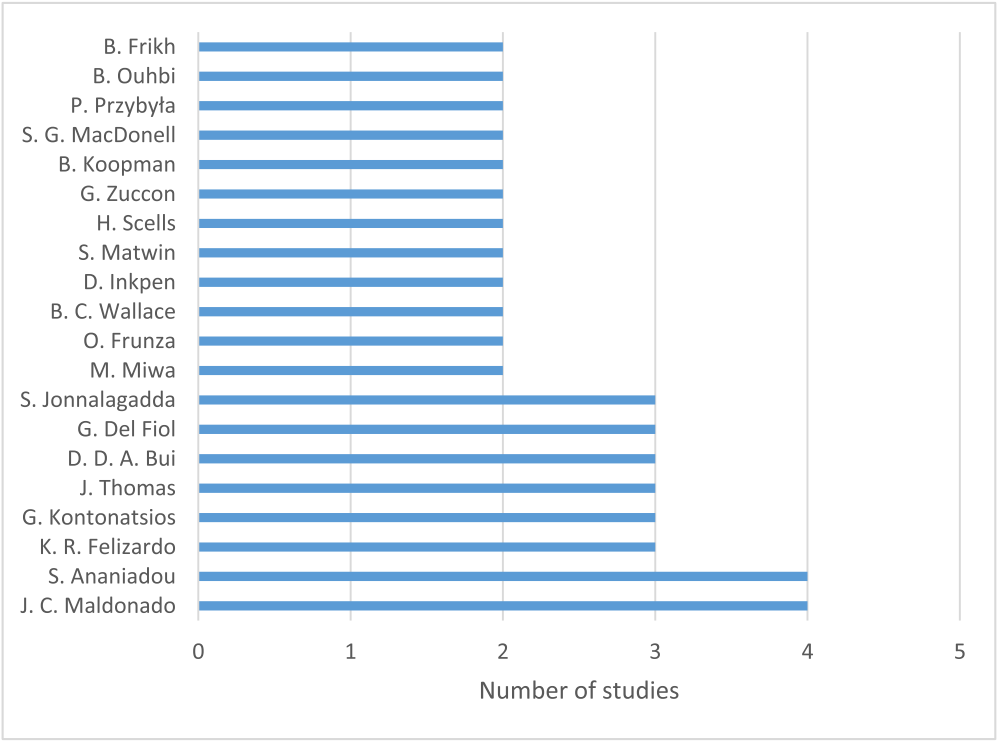


Fig. 3. Authors that have published more than one relevant article.

Table 9
Number of inclusions per publication title.

Publication Channel	Publication Source	Type	No. of Studies
Journal of Biomedical Informatics	ScienceDirect	Journal	9
Evaluation and Assessment in Software Engineering	ACM & IEEE	Conference	3
Artificial Intelligence in Medicine	ScienceDirect	Journal	2
Journal of the American Medical Informatics Association	ScienceDirect	Journal	2
Empirical Software Engineering and Measurement	IEEE	Conference	2
Expert Systems with Applications	ScienceDirect	Journal	2
Information Integration and Web-based Applications and Services	ACM	Conference	2
The World Wide Web Conference	ACM	Conference	2
Computational Linguistics: Posters	ACM	Conference	1
Applied Computing	ACM	Conference	1
IEEE Transactions on NanoBioscience	IEEE	Journal	1
Information Systems and Technologies (CISTI)	IEEE	Conference	1
Social Networks Analysis, Management and Security (SNAMS)	IEEE	Conference	1
Latin American Computing Conference (CLEI)	IEEE	Conference	1
IEEE Access	IEEE	Journal	1
Asian Journal of Psychiatry	ScienceDirect	Journal	1
Information and Software Technology	ScienceDirect	Journal	1
Environment International	ScienceDirect	Journal	1
Expert Systems with Applications: X	ScienceDirect	Journal	1
MethodsX	ScienceDirect	Journal	1
Knowledge discovery and data mining	ACM	Conference	1
Information Systems Frontiers	ScienceDirect	Journal	1
J. Mach. Learn. Res.	ACM Digital	Journal	1
Research Synthesis Methods	Wiley	Journal	1
Systematic reviews	Springer	Journal	1
Enhanced Living Environments	Springer	Chapter	1

in Table 3. The results of each sub-question have been presented to answer the main research question fully.

4.6.1. *Which parts of the documents are used to automate systematic reviews?*

All the *Conduct of the Systematic Review* (SLR5 to SLR9) steps have an interaction between the document and algorithm. This interaction means that the algorithm must use text for its automation task. However, it is essential to consider which parts of the documents to choose per step of the systematic review. We have created a bubble chart, as shown in Fig. 8, which displays which part of the document (e.g., Document Features) has been used in which step of the systematic review process. SLR5–8 are the only steps automated, as elaborated in Section 4.5 and therefore, the other steps are excluded from the following figures.

We first provide a more profound elaboration on the document features that are not usually found in the metadata.

Reference number describes the number of references an author made; this can indicate the quality of a study: often, more is better. Reference information describes the references a study made, which indicates that if two studies made the same reference, they are possibly similar and essential to read. At last, we have subject headings with a medical focus. Both Emtree and Medical Subject Headings (MeSH) cover the same focus here. If two studies have the same subject heading, e.g., *asthma*, possibly they are similar.

We have found that the full-text is often avoided for the *Selection of Primary Studies* (SLR6). [13,14,83] describe that this is because titles and abstracts are significantly different from full-texts. As quoted from Cohen, Johnson, Verspoor, Roeder and Hunter [83]: *"Full-text articles can also present other challenges, such as the recognition and clean-up of embedded tags, non-ASCII characters, tables and figures, and even the need to convert from PDF to textual format. Access to full text is an especially troublesome issue"* [83]. These challenges are easily avoided by using, for instance, the title and abstract. Dieste and Padua [84] confirm that using only the title and abstract are a better option than using full-text in the *Selection of Primary Studies*. Opposingly, for the *Data Extraction and Monitoring* step (SLR8), the full-text is used almost exclusively because

this step needs all information available from the document.

Which NLP preprocessing and representation techniques are used to automate systematic reviews?

First, we collected the main NLP preprocessing steps, as pointed out by [85–87]. We've added the following steps in addition to these critical steps:

- Bi-normal separation
- Language detection
- Introduction, Methods, Results, and Discussion (IMRAD) detection
- Word to Number, e.g., from 'seven' to '7'.
- Linguistic normalization (i.e., synonym/hypernym/acronym normalization), e.g., from 'SLR' to 'systematic literature review,' or 'identity' and 'ID' to 'identification.'

As Fig. 9 shows, the usage of the NLP preprocessing step was highest at the automation of the *Selection of Primary Studies* (SLR6). The removal of stop words and stemming were used most often. When researchers provided more in-depth information on the stemming algorithm, they used the Porter stemmer.

Second, we've collected the main NLP representation techniques, as pointed out by [88]. We've added the following techniques in addition to their key steps:

- Latent semantic analysis
- Singular value decomposition
- Text vector (word/paragraph/document vector)

As Fig. 10 shows, the primary usage of NLP representation techniques was at the automation of the *Selection of Primary Studies* (SLR6). *Bag of Words* (BoW) and *Term Frequency-Inverse Document Frequency* (TF-IDF) have been used most often. When making use of n-grams, researchers often used a combination of unigrams and bigrams. In one case, also trigrams were used.

Which machine learning techniques, tasks, evaluation approach, evaluation metrics, and algorithms are used?

Fig. 11 shows the machine learning techniques used on the left side of the vertical axis and machine learning tasks used on the axis's right side. The machine learning task categories were used as listed by [89]. We found that supervised machine learning is the primary technique for the automation of systematic reviews. Furthermore, classification is the main task in both the *Selection of Primary Studies* and *Data Extraction* (SLR6 and 8). For the *Identification of Research* (SLR5), the ranking was the primary task. For the *Study Quality Assessment* (SLR7), clustering was the only task highlighted by [13].

After developing their machine learning-powered tool, researchers need to evaluate the tool to display its power and relevance. Therefore, we have collected the evaluation approaches and metrics used by researchers to evaluate machine learning tools [90,91]. The results can be observed in Fig. 12. To save space on paper, we have taken out all metrics used just one time. These metrics were all used in the *Selection of Primary Studies* (SLR6). However, the last item was used in the *Data Extraction* step. The metrics kept out of the figure are as follows:

- Sum of Squared Errors
- Sensitivity
- Specificity
- Positive likelihood relation
- Net reclassification index
- Coverage
- Matthews correlation coefficient
- Normalized Discounted Cumulative Gain

The main evaluation metrics used are precision, recall, and F-measure. The main evaluation approaches are cross-validation and train/test sets. When using cross-validation, 11 studies adopted the 10-fold or 5 × 2-fold approach [33,37,38,43,44,48,52,61,63,66,67]. When using train/test sets, 2 studies opted for the 70/30 approach [55,63]. One study used a 70/30/30 approach [46] when using a train/-test/validation set, meaning they took 30% from the full set as the test set and took 30% from the train set as the validation set.

Furthermore, also, non-machine learning metrics were used. These metrics all tried to measure the systematic review process's improvement through one or more case studies. As it is used 9 times, the primary metric in this field is Work Saved over Sampling (WSS), founded by [33]. As they describe, *"We define the work saved as the percentage of papers that meet the original search criteria that the reviewers do not have to read (because they have been screened out by the classifier). A recall of 0.95 can be obtained with a 0.95 random sampling of the data, and this process would save the reviewers 5% of the work of reading the papers. Clearly, for the classifier system to provide an advantage, the work saved must be greater than the work saved by simple random sampling. Therefore, we measure the work saved over and above the work saved by simple sampling for a given level of recall"* [33].

At last, we have collected all machine learning algorithms discussed per paper. If a paper discussed multiple algorithms and evaluated them to find the best solution, all algorithms were still included. We have split the machine learning algorithms into deep learning and shallow learning algorithms. For shallow learning techniques, SVM and Bayesian Networks have been used the most. SVM has also been used to support active learning. Focusing on Bayesian Networks, studies used (Complement) Naïve Bayes most often.

It has taken until early 2020 to find at least one paper that discusses Deep Learning algorithms to automate systematic reviews [66]. In their paper, [66] describes that they have used a denoising autoencoder combined with a deep neural network for document feature extraction. Consecutively, they used a flattened vector resulting from the last feed-forward layer as input for an SVM to select primary studies' relevance. We want to point out that still to this day, no study has been published that uses deep learning for other purposes, document classification. No study has been published with highly potential deep learning techniques like CNN, LSTM, and RNN.

4.7. RQ-6: challenges and solution directions

We aimed to find the open challenges and solution direction for the automation of systematic reviews. In this process, we collected limitations explicitly formulated in the article body and the Limitations sections. Table 10 lists the studies' challenges, how many studies discuss these challenges, and the solutions found by the studies.

Most of the studies discuss challenge 2 regarding class imbalance since the *Selection of Primary Studies* (SLR6) step often must deal with a skewed distribution of a high number of negatives and a small number of positives. Such a skewed distribution causes classification problems, as most classifiers tend to maximize overall accuracy.

Challenge 1 discusses the need for a system that can retrieve all full-text articles from various databases automatically. Studies that discussed challenge 3 needed to select the best features for their models. Other studies that discussed challenge 4 used PDF's and needed some type of conversion to be used as input for their model. They also suggested a tool that could extract tables and images from PDF files. Challenge 6 proposes that an Active Learning model is not always best, as reviewers like to hold control over the results [52]. Therefore, Cohen et al. [52] suggested using a ranking model instead. At last, challenge 7 points out that a canonical model is not understandable by machines, and therefore should be translated to machine-understandable data [62].

5. Discussion

In the following sub-sections, we discuss the results of our study. In Section 5.1, we provide a critical reflection on the results. In Section 5.2

Table 10
Challenges and their solution directions.

ID	Challenge	SLR Step	# of Studies	Reference	Solution and Corresponding Study
1	Complete literature retriever is missing	SLR6	2	[54,55]	Develop a web crawler that retrieves full-text articles from the primary databases through their APIs [40]
2	Class imbalance is challenging for a model to train	SLR6	11	[41–44, 53,58,61, 63,64,66, 92]	Cost assignment [66] Data resampling [41,42,58,64] SMOTE [43,53] Feature enrichment [43,63] Cost-sensitive classifiers [43,53, 61,66] Precision@95% recall [43]
3	Finding and developing optimal features	SLR5, 6, 8	7	[45,46,51, 53,56,63, 92]	Add MeSH terms [45,56] Add Reference information [63] Include full-text for SLR6 [92] Filter out noisy and barely discriminative features [53]
4	Document feature extraction is not optimal	SLR7,8	2	[13,34]	Develop a method to extract text, images, and tables without much cleaning [34,35,54, 62].
5	Precision and/or recall can be improved	SLR5, 6	7	[16,37,38, 49,58,59, 67]	Use a rule-based algorithm [37,67] Iteratively refine search string to increase the recall [38] Include only reviewers with expert domain knowledge [16] Use automated document classification techniques instead of ranking [49] Use the technique as a literature scoping approach [59] Don't use the technique when there is a class imbalance [58]
6	Active learning is reversing control	SLR6	1	[52]	Ranking would be a more attractive option for reviewers, as they hold control [52]
7	A canonical model is not understandable by machines	SLR8	1	[62]	Translate the model into machine-understandable data using AI [62]

we discuss the results in relation to the related work. In Section 5.3, we discuss the threats to the validity of the present study and how we tried to address them.

5.1. General discussion

To the best of our knowledge, this study represents the first systematic literature review on the automation of systematic literature reviews with a focus on all systematic literature review steps and NLP and ML techniques. In this respect, we identified over a thousand papers from which we identified 41 high-quality primary studies. From the results, we can identify several interesting observations, which we will highlight per the research question.

5.2. Main statistics

We have included 41 primary studies in this review, whereof 60% of studies were categorized as journal articles. Since 2006, a stable number of high-quality papers have been published (see Fig. 4), whereby the focus has been on the automation of the *Conduct of the Review*, with *Selection of Primary Studies* as the key automated step.

5.3. What is the objective of automation?

We want to note that most researchers did not mention their objectives explicitly. However, most of the mentioned objectives were straightforward and aimed at automating a specific step in the systematic review process, such as the (semi) automation of the *Primary Study Selection* (SLR6) step. Another objective that was often implicitly mentioned is to reduce human workload in the SLR process.

5.4. On which application domains are automated systematic literature review tools evaluated?

We have found that there were two main domains used for the automation of SLR studies: Software Engineering and Medicine. For the Medical domain, we have found four sub-domains. The fact that we could not identify primary studies for other domains than Software Engineering and Medical might indicate that these domains are not explicitly described in the literature.

5.5. Which are the used databases for the automation of systematic literature reviews?

We have found that scientific databases are being used in the evaluation of SLR tools, and for automated metadata extraction. The following paragraphs will discuss the results per topic.

5.6. RQ3a which databases have been used for automated metadata extraction?

For automated metadata extraction, we have found just one study. In this study, González-Toral, Freire, Gualán and Saquicela [40] describe an algorithm that automates metadata extraction from articles

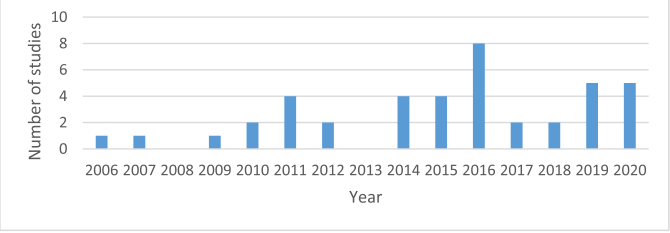


Fig. 4. Year-wise distribution of the included articles.

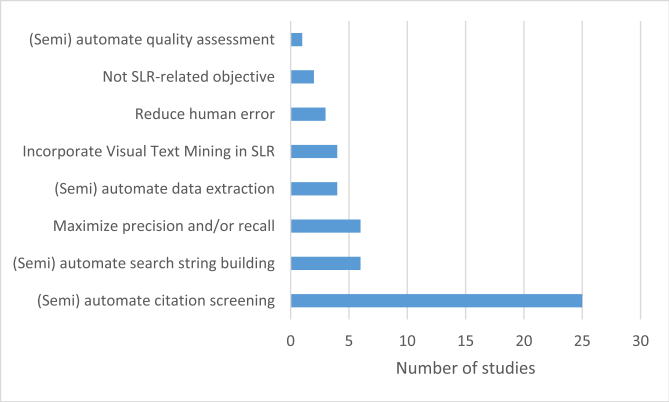


Fig. 5. Objective of automation.

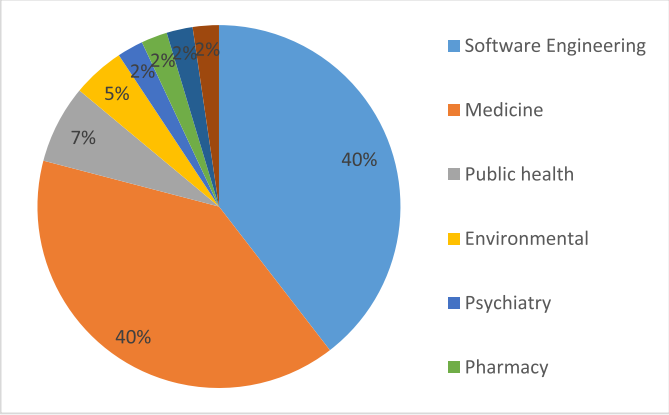


Fig. 6. Application domain.

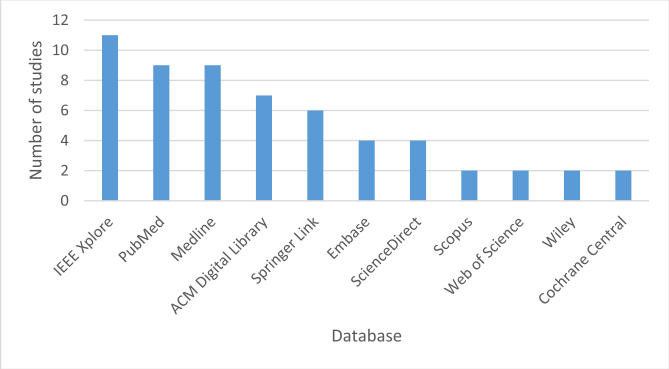


Fig. 7. Databases that were used for extracting articles for evaluation of a tool.

published by IEEE, ACM, Springer, Scopus, and Semantic Scholar. The fact that we did not find other articles mentioning metadata extraction of scientific literature might indicate that these domains are not explicitly described in the literature, or that we should have widened our search query.

5.7. RQ3b which databases have been used for the retrieval of studies to evaluate a tool?

We found that 26 articles had a form of article extraction from a database to evaluate their tool. IEEE Xplore and PubMed were used most often to gather document features automatically. The main application domain also reflected on the databases used: most had a discipline in Medicine.

5.8. RQ4 which are the automated steps of systematic literature reviews?

All automated steps are from the *Conduct of review* category (SLR5–9), where most studies focused on *Selecting Primary Studies* (SLR6). We think that most studies are focused on the *Conduct of review*, as the *Need for a review* and *Reporting the review* are categories that require human creativity and insight and are yet too difficult to be automated.

5.9. RQ5 which are the adopted machine learning-based automation techniques?

The following paragraphs will discuss the use of machine learning automation techniques to support the systematic literature review process.

5.10. RQ5a which parts of the documents are used to automate systematic reviews?

Throughout the steps, the most-used features are text from the documents. For the *Data Extraction and Monitoring* step (SLR8), the full-text is used almost exclusively because this step needs all information available from the document. The other steps mainly prefer using the title and abstract as features. In the medical domain, also MeSH features are used often. However, this also makes automation systems domain-specific.

5.11. RQ5b which NLP preprocessing and representation techniques are used to automate systematic reviews?

We have seen that the NLP preprocessing step contains many techniques. As the SLR process is currently automated through shallow machine learning techniques, features need to be hand-crafted and finetuned for each domain or even per dataset. Furthermore, also many NLP representation techniques are currently deemed legacy, as text classification currently mainly makes use of word embeddings. Word embeddings tend to capture the meaning of words, which eliminates the need for thorough text cleaning and finetuning hand-crafted features.

RQ5c Which machine learning techniques, tasks, evaluation approach, evaluation metrics, and algorithms are used to automate systematic reviews?

The automation of SLR studies is most often done by classification, which is a supervised machine learning task. Furthermore, most models are evaluated using Cross-Validation, with the main metrics being Work Saved over Sampling (WSS), Recall, Precision, and F-measure. As WSS enraptures the intention of Precision, Recall, and F-measure, WSS should be used as the main metric for the automation of SLR studies.

Furthermore, SVM and Bayesian Networks, such as Naïve Bayes classifiers, are mostly used for the automation of systematic literature reviews, across most steps. Based on our results, we can state that there seems to be a noticeable lack of evidence on Deep Learning techniques for the automation of systematic literature reviews.

5.12. RQ6 what are the open challenges and solution directions?

For each of the automated steps, we have found at least one challenge. Most of the papers that mentioned a challenge also came up with a solution. Therefore, we categorized the solutions, and we were able to present at least one solution per challenge. The most reoccurring challenge that researchers encounter is class imbalance in the *Primary Study Selection* (SLR6) step. This challenge resulted in 6 solutions that have been widely used across multiple studies.

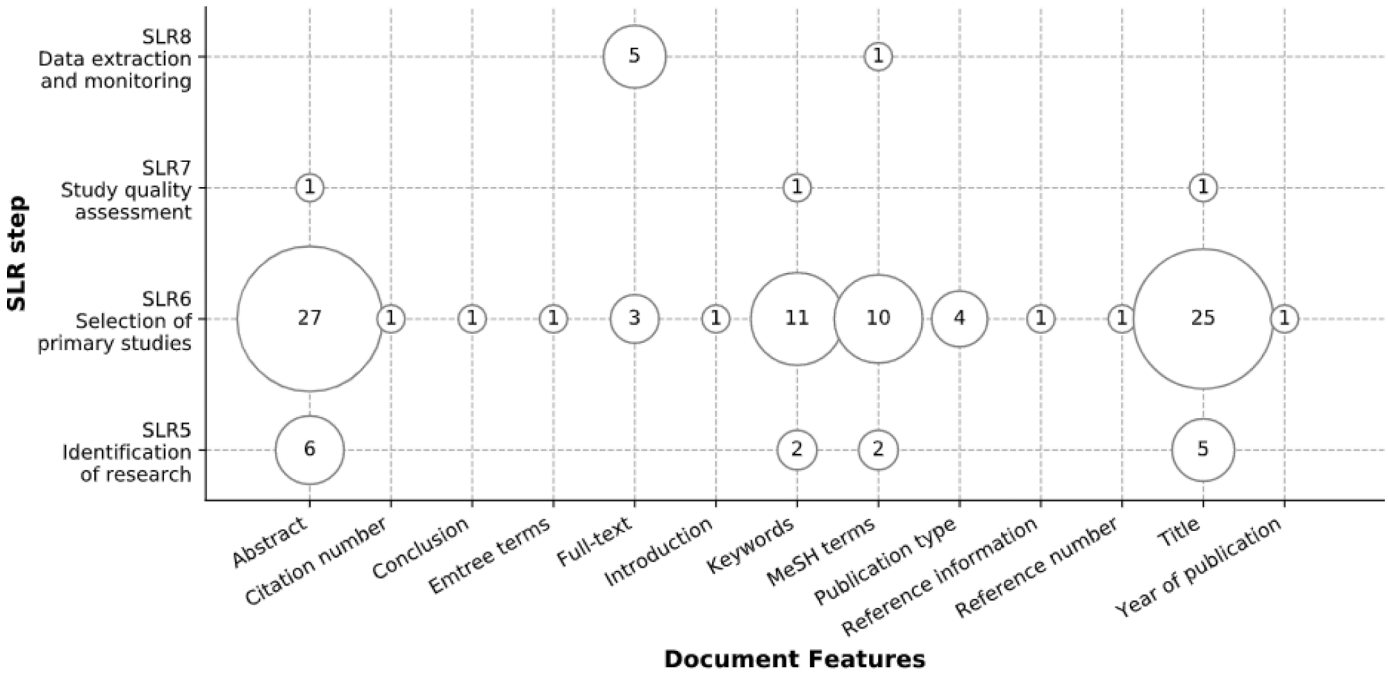


Fig. 8. Document features used as input for the automation tool.

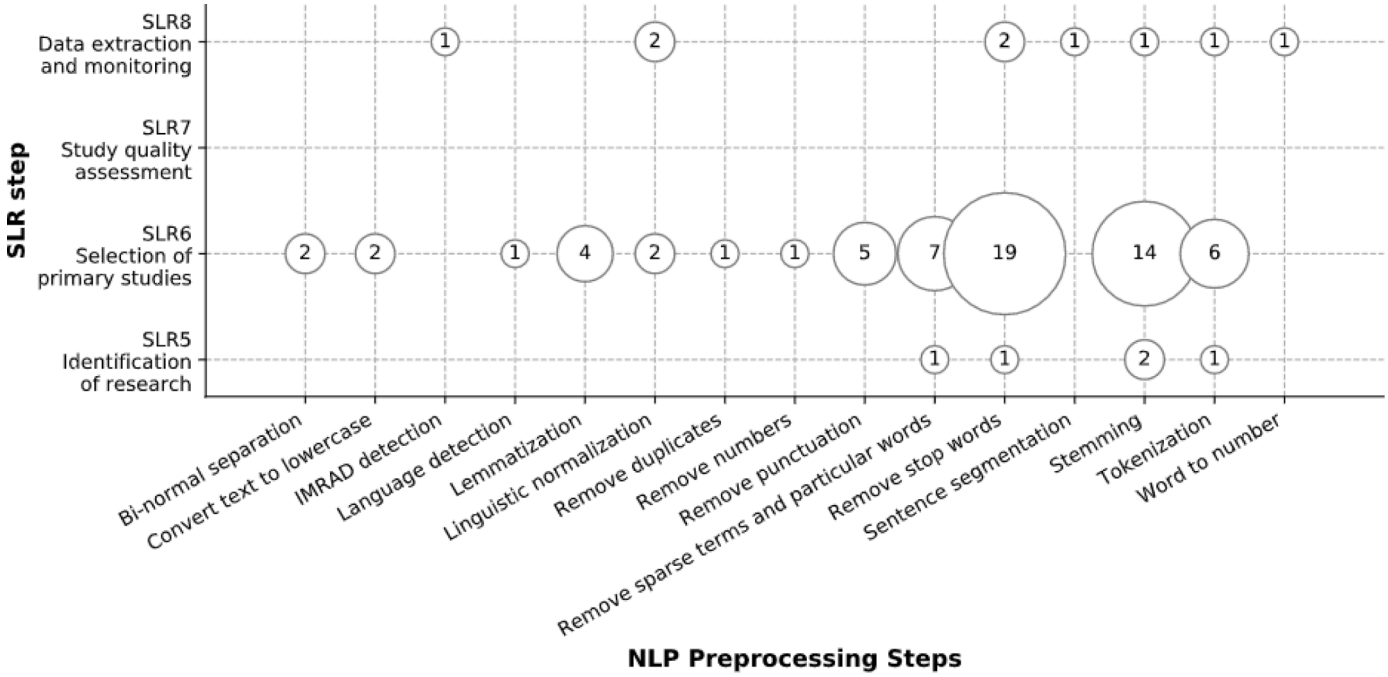


Fig. 9. NLP preprocessing steps.

5.13. Results in relation to related work

Our study's main difference with the related work is that we have explicitly adopted a systematic literature review protocol that is widely accepted and used in the Evidence-Based Medicine and Software Engineering communities. Based on the SLR protocol, we have searched and identified primary studies from a broad set of over a thousand studies from which we selected 41 primary studies. This systematic approach differs from [17,23,25], which published a report on systematic reviews' automation. Besides, [12,17–20] studied tools to automate a – part of – the systematic review process, not always the full process. Feng et al. [18] is the only outdated paper that discusses Text Mining techniques to automate the full systematic review process. However, the study did not discuss NLP techniques, databases for extraction, objectives, open challenges, and split them into the systematic review steps proposed by Kitchenham and Charters [4]. Therefore, we think we have developed a systematic review that is significantly different, with new and relevant insights.

5.14. Threats to validity

Construct validity: Construct validity assesses whether the SLR

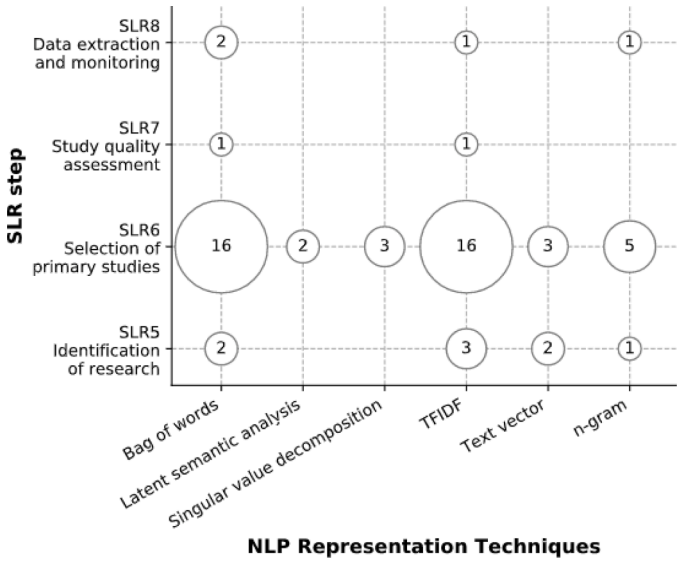


Fig. 10. NLP representation techniques.

represents the degree to which it measures what it asserts. We attempted to analyze the information from the current literature and used automated search queries in multiple databases to protect this purpose. Even though a database is a robust literature search tool, it is also sensitive to a query's phrasing, where even the slightest change of words can cause very different search results.

The query's phrasing is our first threat. Each database has different search query formulation options, the use of Boolean operators, and search fields, which meant that we slightly modified the general query for each database. Consequently, the modified search query might have missed a related study. Therefore, the query design has been thoroughly discussed among the authors and checked through several trials to prevent these risks. We manually checked the abstract of the trial articles to ensure the search query was correct. If several of the results returned were insignificant, the query was updated, and the trial was executed again.

The second threat relates to the selection of primary studies, with publication and selection bias in specific. Publication bias is the phenomenon that authors are more likely to publish positive results than the negative results of their research [4]. We believe we covered the threat of publication bias by applying the study quality assessment. However, we did not find any low-quality papers after study applying the selection criteria. The threat of selection bias was covered by defining the study selection criteria after screening a primary studies pilot set. All selection criteria were discussed among the co-authors to ensure their quality. Although the selection is based on predefined selection criteria and quality assessment questions, it is impossible to remove personal and subjective decisions during the scoring. Since the systematic review domain's automation is vast, it requires a broad spectrum of expert knowledge. Therefore, we carefully reviewed the studies, but we might have misidentified some of the studies.

The third threat is the data extraction step. Even though the data extraction form was predefined, it is highly likely that some useful data fields were not included in this extraction form. We updated the data extraction form multiple times to ensure that we extracted all relevant data from the included studies. We added new data fields if the data could be extracted from most studies and if it was useful to answer the research questions.

Internal Validity: Internal validity shows the incomplete relationship between results, which may lead to structural errors. We formulated all research questions carefully to identify the required techniques for the automation of systematic literature reviews. As these automation techniques were well-defined, the synergy of research questions and objectives were described adequately.

External Validity: This systematic review only investigated published studies that applied machine learning or natural language processing techniques to automate the systematic literature review process. It is likely that a new machine/deep learning or natural language processing algorithm has not been applied yet in the automation of systematic literature reviews. As these studies have not been published, they have not been discussed regardless of their potential.

Conclusion Validity: The conclusion validity measures the reproducibility of the systematic literature review. Our study followed the protocol proposed by [4]. The research question design, search process, screening criteria, and quality evaluation were performed based on this widely used protocol. Our systematic literature review process was also discussed among the authors to minimize individual bias. We derived all conclusions from the extracted and synthesized data based on the tables and figures to avoid subjective interpretation of the results among researchers.

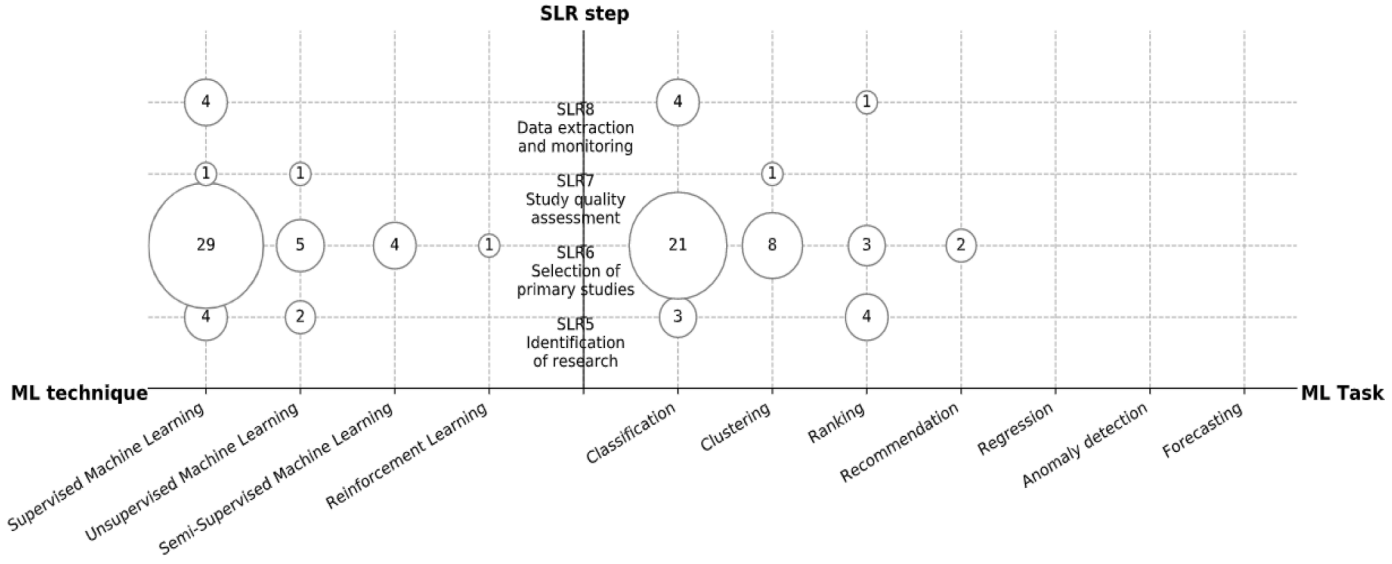


Fig. 11. Machine Learning automation techniques and tasks per SLR step.

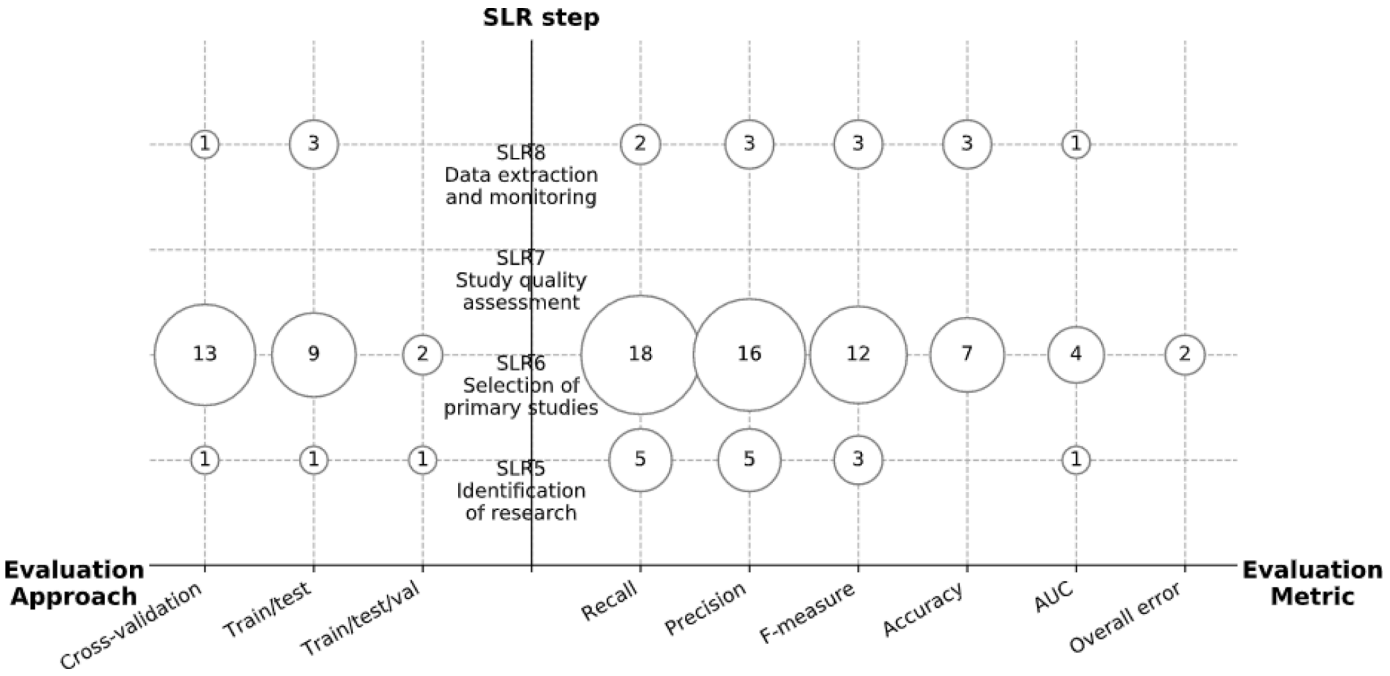


Fig. 12. Model evaluation approaches and metrics.

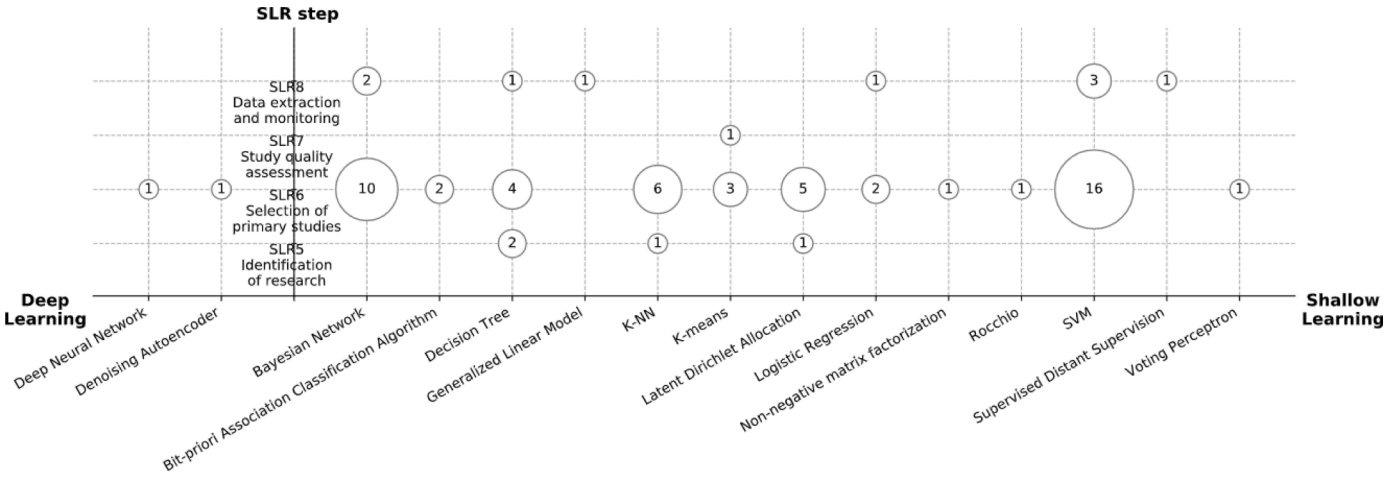


Fig. 13. Deep and shallow learning techniques.

6. Conclusion and future work

In this study, we have systematically searched the scientific literature of the past twenty years to identify the features, challenges, and solution directions of SLRs' automation through machine learning and NLP. We addressed 41 studies in this systematic literature review that capture state-of-the-art strategies to automate systematic literature reviews. To the best of our knowledge, this is the first SLR of its kind and the first to review machine learning and NLP techniques per step of the SLR process. Our choice to adopt an SLR as an instrument to answer our key research questions appeared to be very useful and led us to the critical insights that could benefit both practitioners and researchers.

This study has led to novel insights into the current literature on the automation of SLRs. Our bubble charts enable researchers to comfortably find the key features and algorithms to use when developing their tool to automate SLRs. We have found that time reduction is the primary goal of automating a systematic literature review since manual execution is labor-intensive, time-consuming, and vulnerable to errors. As a result, the leading automated step in SLRs is the Selection of Primary Studies (SLR6), which is also the most time-consuming step. Although many studies have provided automation approaches for systematic literature reviews, no study has been found to apply automation techniques in the planning and reporting phase. In the Evidence-Based Medicine and Software Engineering domains, machine learning techniques, such as SVMs and Naïve Bayes, have been widely applied to automate the Selection of Primary Studies (SLR6). Another important insight from this SLR is the overview of challenges and solution directions. We have categorized the challenges that researchers mentioned based on their kind and included multiple solutions per challenge. Based on this systematic literature review, there is a research gap in developing novel methods to support the systematic literature review process through Deep Learning techniques.

Our analysis observed that not all steps of the SLR process had been automated yet. There is a focus on most of the *Conducting the Review* steps of the SLR process. However, some of the steps have not been addressed in scientific studies. Some steps are relatively easier to automate than others due to the nature of the underlying problem, but other steps require technical challenges. Current ML and NLP techniques are

not sufficient to cope with some of these problems and additional techniques and technologies are required. There is also a lack of tool support.

Automating the SLR process pays off because it dramatically reduces the required time and effort; this automation objective will be more and more critical in the future in many different fields. As for future research, we plan to develop a deep learning tool to automate the Selection of Primary Studies (SLR6) in the systematic review process.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Elsevier, Content & features, in, n.d.
[2] PubMed, About, in, n.d.
[3] E.P. Office, Why Researchers Should Care About Patents, 2007.
[4] B. Kitchenham, S. Charters, Guidelines For Performing Systematic Literature Reviews in Software Engineering, Keele University, 2007.
[5] H.G. Gurbuz, B. Tekinerdogan, Model-based testing for software safety: a systematic mapping study, Softw. Qual. J. 26 (2018) 1327–1372.
[6] C. Marshall, Tool Support for Systematic Reviews in Software Engineering, Keele University, 2016.
[7] J.H. Elliott, T. Turner, O. Clavisi, J. Thomas, J.P. Higgins, C. Mavergames, R. L. Gruen, Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap, PLoS Med. 11 (2014), e1001603.
[8] S.R. Jonnalagadda, P. Goyal, M.D. Huffman, Automating data extraction in systematic reviews: a systematic review, Syst. Rev. 4 (2015) 78.
[9] K.G. Shojania, M. Sampson, M.T. Ansari, J. Ji, S. Doucette, D. Moher, How quickly do systematic reviews go out of date? A survival analysis, Ann. Intern. Med. 147 (2007) 224–233.
[10] G.G. Chowdhury, Natural language processing, Annu. Rev. Inf. Sci. Technol. 37 (2003) 51–89.
[11] M. Bartholomew, James Lind's treatise of the scurvy (1753), Postgrad. Med. J. 78 (2002) 695–696.
[12] A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, S. Ananiadou, Using text mining for study identification in systematic reviews: a systematic review of current approaches, Syst. Rev. 4 (2015) 5.
[13] K.R. Felizardo, G.F. Andery, F.V. Paulovich, R. Minghim, J.C. Maldonado, A visual analysis approach to validate the selection review of primary studies in systematic reviews, Inf. Softw. Technol. 54 (2012) 1079–1091.
[14] K.R. Felizardo, E.Y. Nakagawa, S.G. MacDonell, J.C. Maldonado, A visual analysis approach to update systematic reviews, in: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, London, England, United Kingdom, Association for Computing Machinery, 2014. Article 4.
[15] K.R. Felizardo, M. Riaz, M. Sulayman, E. Mendes, S.G. MacDonell, J.C. Maldonado, Analysing the use of graphs to represent the results of systematic reviews in software engineering, in: 2011 25th Brazilian Symposium on Software Engineering, 2011, pp. 174–183.
[16] V. Malheiros, E. Hohn, R. Pinho, M. Mendonca, J.C. Maldonado, A visual text mining approach for systematic reviews, in: First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007), 2007, pp. 245–254.
[17] J. Thomas, J. McNaught, S. Ananiadou, Applications of text mining within systematic reviews, Res. Synth. Methods 2 (2011) 1–14.
[18] L. Feng, Y.K. Chiam, S.K. Lo, Text-mining techniques and tools for systematic literature reviews: a systematic literature review, in: 2017 24th Asia-Pacific Software Engineering Conference (APSEC), 2017, pp. 41–50.
[19] E. Beller, J. Clark, G. Tsafnat, C. Adams, H. Diehl, H. Lund, M. Ouzzani, K. Thayer, J. Thomas, T. Turner, Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR), Syst. Rev. 7 (2018) 77.
[20] B.K. Olorisade, E.d. Quincey, P. Brereton, P. Andras, A critical analysis of studies that address the use of text mining for citation screening in systematic reviews, in: Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering, Limerick, Ireland, Association for Computing Machinery, 2016. Article 14.
[21] G. Tsafnat, P. Glasziou, M.K. Choong, A. Dunn, F. Galgani, E. Coiera, Systematic review automation technologies, Syst. Rev. 3 (2014) 74.
[22] B.C. Wallace, T.A. Trikalinos, J. Lau, C. Brodley, C.H. Schmid, Semi-automated screening of biomedical citations for systematic reviews, BMC Bioinformatics 11 (2010) 1–11.
[23] R. Paynter, L.L. Bañez, E. Berliner, E. Erinoff, J. Lege-Matsuura, S. Potter, S. Uhl, E. P.C. Methods, An exploration of the use of text-mining software in systematic reviews. EPC Methods: an Exploration of the Use of Text-Mining Software in Systematic Reviews, Agency for Healthcare Research and Quality (US), Rockville (MD), 2016.
[24] Y. Shakeel, J. Krüger, I.v. Nostitz-Wallwitz, C. Lausberger, G.C. Durand, G. Saake, T. Leich, Automated) literature analysis: threats and experiences, in: Proceedings of the International Workshop on Software Engineering for Science, Gothenburg, Sweden, Association for Computing Machinery, 2018, pp. 20–27.
[25] S. Jaspers, E. De Troyer, M. Aerts, Machine Learning Techniques For the Automation of Literature Reviews and Systematic Reviews in EFSA, 15, EFSA Supporting Publications, 2018, 1427E.
[26] A.M. O'Connor, G. Tsafnat, J. Thomas, P. Glasziou, S.B. Gilbert, B. Hutton, A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? Syst. Rev. 8 (2019) 143.
[27] M. Michelson, K. Reuter, The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials, Contemp. Clin. Trials Commun. 16 (2019), 100443.
[28] I.J. Marshall, B.C. Wallace, Toward systematic review automation: a practical guide to using machine learning tools in research synthesis, Syst. Rev. 8 (2019) 163.
[29] A.J. van Altena, R. Spijker, S.D. Olabarriaga, Usage of automation tools in systematic reviews, Res. Synth. Methods 10 (2019) 72–82.
[30] J. Clark, P. Glasziou, C. Del Mar, A. Bannach-Brown, P. Stehlik, A.M. Scott, A full systematic review was completed in 2 weeks using automation tools: a case study, J. Clin. Epidemiol. 121 (2020) 81–90.
[31] M.S. Ali, M.A. Babar, L. Chen, K.-J. Stol, A systematic review of comparative evidence of aspect-oriented programming, Inf. Softw. Technol. 52 (2010) 871–887.
[32] J. Tummers, A. Kassahun, B. Tekinerdogan, Obstacles and features of farm management information systems: a systematic literature review, Comput. Electron. Agric. 157 (2019) 189–204.
[33] A.M. Cohen, W.R. Hersh, K. Peterson, P.-Y. Yen, Reducing workload in systematic review preparation using automated citation classification, J. Am. Med. Inform. Assoc. 13 (2006) 206–219.
[34] D.D.A. Bui, G. Del Fiol, J.F. Hurdle, S. Jonnalagadda, Extractive text summarization system to aid data extraction from full text in systematic review development, J. Biomed. Inform. 64 (2016) 265–272.
[35] D.D.A. Bui, G. Del Fiol, S. Jonnalagadda, PDF text classification to leverage information extraction from publication reports, J. Biomed. Inform. 61 (2016) 141–148.
[36] T. Weißer, T. Saßmannshausen, D. Ohrndorf, P. Burggräf, J. Wagner, A clustering approach for topic filtering within systematic literature reviews, MethodsX 7 (2020), 100831.
[37] B. Ouhbi, M. Kamoune, B. Frikh, E.M. Zemmouri, H. Behja, 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, Singapore, Singapore, Association for Computing Machinery, 2016, pp. 106–114.
[38] R. Ros, E. Bjarnason, P. Runeson, A machine learning approach for semi-automated search and selection in literature studies, in: Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering, Karlskrona, Sweden, Association for Computing Machinery, 2017, pp. 118–127.
[39] G.D. Mergel, M.S. Silveira, T.S.d. Silva, 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, Salamanca, Spain, Association for Computing Machinery, 2015, pp. 1594–1601.
[40] S. González-Toral, R. Freire, R. Gualán, V. Saquicela, A ranking-based approach for supporting the initial selection of primary studies in a systematic literature review, in: 2019 XLV Latin American Computing Conference (CLEI), 2019, pp. 1–10.
[41] G. Kontonatsios, A.J. Brockmeier, P. Przybyła, J. McNaught, T. Mu, J. Y. Goulermas, S. Ananiadou, A semi-supervised approach using label propagation to support citation screening, J. Biomed. Inform. 72 (2017) 67–76.
[42] B.C. Wallace, K. Small, C.E. Brodley, T.A. Trikalinos, Active learning for biomedical citation screening, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, Washington, DC, USA, 2010, pp. 173–182.
[43] P. Timsina, J. Liu, O. El-Gayar, Advanced analytics for the automation of medical systematic reviews, Inf. Syst. Front. 18 (2016) 237–252.
[44] S. Kim, J. Choi, An SVM-based high-quality article classifier for systematic reviews, J. Biomed. Inform. 47 (2014) 153–159.
[45] H. Scells, G. Zuccon, B. Koopman, J. Clark, Automatic Boolean query formulation for systematic review literature search, in: Proceedings of The Web Conference 2020, Taipei, Taiwan, Association for Computing Machinery, 2020, pp. 1071–1081.
[46] H. Scells, G. Zuccon, B. Koopman, Automatic Boolean query refinement for systematic review literature search, in: The World Wide Web Conference, San Francisco, CA, USA, Association for Computing Machinery, 2019, pp. 1646–1656.
[47] C. Blake, A. Lucic, Automatic endpoint detection to support the systematic review process, J. Biomed. Inform. 56 (2015) 42–56.
[48] J.J. García Adeva, J.M. Pikatza Atxa, M. Ubeda Carrillo, E. Ansuategi Zengotitabengoa, Automatic text classification to support systematic reviews in medicine, Expert Syst. Appl. 41 (2014) 1498–1508.
[49] D.D.A. Bui, S. Jonnalagadda, G. Del Fiol, Automatically finding relevant citations for clinical guideline development, J. Biomed. Inform. 57 (2015) 436–445.
[50] E. Zdravevski, P. Lameski, V. Trajkovik, I. Chorbev, R. Goleva, N. Pombo, N. M. Garcia, Automation in systematic, scoping and rapid reviews by an NLP toolkit: a case study in enhanced living environments. Enhanced Living Environments, Springer, 2019, pp. 1–18.
[51] O. Frunza, D. Inkpen, S. Matwin, Building systematic reviews using automatic text classification techniques, in: Proceedings of the 23rd International Conference on

Computational Linguistics: Posters, Beijing, China, Association for Computational Linguistics, 2010, pp. 303–311.
[52] A.M. Cohen, K. Ambert, M. McDonagh, Cross-topic learning for work prioritization in systematic review creation and update, J. Am. Med. Inform. Assoc. 16 (2009) 690–704.
[53] H. Almeida, M. Meurs, L. Kosseim, A. Tsang, Data sampling and supervised learning for HIV literature screening, IEEE Trans. Nanobiosci. 15 (2016) 354–361.
[54] A. Langlois, J.-Y. Nie, J. Thomas, Q.N. Hong, P. Pluye, Discriminating between empirical studies and nonempirical works using automated text classification, Res. Synth. Methods 9 (2018) 587–601.
[55] T.R.P.M. Rúbio, C.A.S.J. Gulo, Enhancing academic literature review through relevance recommendation: using bibliometric and text-based features for classification, in: in: 2016 11th Iberian Conference on Information Systems and Technologies (CISTI), 2016, pp. 1–6.
[56] B.C. Wallace, J. Kuiper, A. Sharma, M. Zhu, I.J. Marshall, Extracting PICO sentences from clinical trial reports using supervised distant supervision, J. Mach. Learn. Res. 17 (2016) 4572–4596.
[57] F. Tomassetti, G. Rizzo, A. Vetro, L. Ardito, M. Torchiano, M. Morisio, Linked data approach for selection process automation in systematic reviews, in: 15th Annual Conference on Evaluation & Assessment in Software Engineering (EASE 2011), 2011, pp. 31–35.
[58] A. Bannach-Brown, P. Przybyła, J. Thomas, A.S. Rice, S. Ananiadou, J. Liao, M. R. Macleod, Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error, Syst. Rev. 8 (2019) 1–12.
[59] M. Cawley, R. Beardslee, B. Beverly, A. Hotchkiss, E. Kirrane, R. Sams, A. Varghese, J. Wignall, J. Cowden, Novel text analytics approach to identify relevant literature for human health risk assessments: a pilot study with health effects of in utero exposures, Environ. Int. 134 (2020), 105228.
[60] M. Miwa, J. Thomas, A. O'Mara-Eves, S. Ananiadou, Reducing systematic review workload through certainty-based screening, J. Biomed. Inform. 51 (2014) 242–253.
[61] T. Bekhuis, D. Demner-Fushman, Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers, Artif. Intell. Med. 55 (2012) 197–207.
[62] M.B. Aliyu, R. Iqbal, A. James, The canonical model of structure for data extraction in systematic reviews of scientific research articles, in: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2018, pp. 264–271.
[63] B.K. Olorisade, P. Brereton, P. Andras, The use of bibliography enriched features for automatic citation screening, J. Biomed. Inform. 94 (2019), 103202.
[64] K. Hashimoto, G. Kontonatsios, M. Miwa, S. Ananiadou, Topic detection using paragraph vectors to support active learning in systematic reviews, J. Biomed. Inform. 62 (2016) 59–65.
[65] A. Karami, M. Lundy, F. Webb, Y.K. Dwivedi, Twitter and research: a systematic literature review through text mining, IEEE Access 8 (2020) 67698–67717.
[66] G. Kontonatsios, S. Spencer, P. Matthew, I. Korkontzelos, Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews, Expert Syst. Appl. (2020), 100030.
[67] H. Sellak, B. Ouhbi, B. Frikh, 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, Brussels, Belgium, Association for Computing Machinery, 2015. Article 43.
[68] K.R. Felizardo, N. Salleh, R.M. Martins, E. Mendes, S.G. MacDonell, J. C. Maldonado, Using visual text mining to support the study selection activity in systematic literature reviews, in: 2011 International Symposium on Empirical Software Engineering and Measurement, 2011, pp. 77–86.
[69] Á.M. Dias, C.G. Mansur, M. Myczkowski, M. Marcolin, Whole field tendencies in transcranial magnetic stimulation: a systematic review with data and text mining, Asian J. Psychiatr. 4 (2011) 107–112.
[70] 5. Elsevier, Year Impact Factor & Ranking, 2020.
[71] G. Tsafnat, P. Glasziou, G. Karystianis, E. Coiera, Automated screening of research studies for systematic reviews using study characteristics, Syst. Rev. 7 (2018) 64.
[72] T. Dyba, T. Dingsoyr, G.K. Hanssen, Applying systematic reviews to diverse study types: an experience report, in: First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007), IEEE, 2007, pp. 225–234.
[73] M. Riaz, M. Sulayman, N. Salleh, E. Mendes, Experiences conducting systematic reviews from novices' perspective, in: 14th International Conference on Evaluation and Assessment in Software Engineering (EASE), 2010, pp. 1–10.
[74] J. Biolchini, P.G. Mian, A.C.C. Natali, G.H. Travassos, Systematic Review in Software Engineering, 679, System Engineering and Computer Science Department COPPE/UFRJ, 2005, p. 45. Technical Report ES.
[75] S. Karimi, J. Zobel, S. Pohl, F. Scholer, The challenge of high recall in biomedical systematic search, in: Proceedings of the Third International Workshop on Data and Text Mining in Bioinformatics, 2009, pp. 89–92.
[76] S. Reeves, I. Koppel, H. Barr, D. Freeth, M. Hammick, Twelve tips for undertaking a systematic review, Med. Teach. 24 (2002) 358–363.
[77] J. McGowan, M. Sampson, Systematic reviews need systematic searchers, J. Med. Libr. Assoc. 93 (2005) 74.
[78] I. Shemilt, N. Khan, S. Park, J. Thomas, Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews, Syst. Rev. 5 (2016) 140.
[79] B. Kitchenham, Procedures for Performing Systematic Reviews, 33, Keele University, 2004, pp. 1–26.
[80] M.B. Elamin, D.N. Flynn, D. Bassler, M. Briel, P. Alonso-Coello, P.J. Karanicolas, G. H. Guyatt, G. Malaga, T.A. Furukawa, R. Kunz, Choice of data extraction tools for systematic reviews depends on resources and review complexity, J. Clin. Epidemiol. 62 (2009) 506–510.
[81] P.C. Gøtzsche, A. Hróbjartsson, K. Marić, B. Tendal, Data extraction errors in meta-analyses that use standardized mean differences, JAMA 298 (2007) 430–437.
[82] A.P. Jones, T. Remmington, P.R. Williamson, D. Ashby, R.L. Smyth, High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews, J. Clin. Epidemiol. 58 (2005) 741–742.
[83] K.B. Cohen, H.L. Johnson, K. Verspoor, C. Roeder, L.E. Hunter, The structural and content aspects of abstracts versus bodies of full text journal articles are different, BMC Bioinform. 11 (2010) 492.
[84] O. Dieste, A.G. Padua, Developing search strategies for detecting relevant experiments for systematic reviews, in: First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007), IEEE, 2007, pp. 215–224.
[85] O. Davydova, Text Preprocessing in Python: Steps, Tools, and Examples, 2018.
[86] U. Verma, Text Preprocessing for NLP (Natural Language Processing), Beginners to Master, 2014.
[87] M. Balatsko, Text Preprocessing Steps and Universal Pipeline, 2019.
[88] M. Kana, Representing Text in Natural Language Processing, 2019.
[89] Microsoft, Machine Learning Tasks in ML.NET, 2019.
[90] S. Minaee, 20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics, 2019.
[91] S. Mutuvi, Introduction to Machine Learning Model Evaluation, 2019.
[92] O. Frunza, D. Inkpen, S. Matwin, W. Klement, P. O'Blenis, Exploiting the systematic review protocol for classification of medical abstracts, Artif. Intell. Med. 51 (2011) 17–25.