

Automated Support for Searching and Selecting Evidence in Software Engineering: A Cross-domain Systematic Mapping

Bianca Minetto Napoleão
Université du Québec à Chicoutimi
Chicoutimi, QC, Canada
bianca.minetto-napoleao1@uqac.ca

Fabio Petrillo
Université du Québec à Chicoutimi
Chicoutimi, QC, Canada
fabio@petrillo.com

Sylvain Hallé
Université du Québec à Chicoutimi
Chicoutimi, QC, Canada
shalle@acm.org

Abstract—**Context:** Searching and selecting relevant evidence is crucial to answer research questions from secondary studies in Software Engineering (SE). The activities of search and selection of studies are labour-intensive, time-consuming and demand automation support. **Objective:** Our goal is to identify and summarize the state-of-the-art on automation support for searching and selecting evidence for secondary studies in SE. **Method:** We performed a systematic mapping on existing automating support to search and select evidence for secondary studies in SE, expanding our investigation in a cross-domain study addressing advancements from the medical field. **Results:** Our results show that the SE field has a variety of tools and Text Classification (TC) approaches to automate the search and selection activities. However, medicine has more well-established tools with a larger adoption than SE. Cross-validation and experiment are the most adopted methods to assess TC approaches. Furthermore, recall and precision are the most adopted assessment metrics. **Conclusion:** Automated approaches for searching and selecting studies in SE have not been applied in practice by SE researchers. Integrated and easy-to-use automated approaches addressing consolidated TC techniques can bring relevant advantages on workload and time saving for SE researchers who conduct secondary studies.

Index Terms—Secondary Studies; Systematic Review Automation; Search of Studies; Selection of Studies

I. INTRODUCTION

Evidence-Based Software Engineering (EBSE) has as basis secondary studies which consist in Systematic Literature Reviews (SLRs) and Systematic Mappings (SMs). Both types of secondary studies follow the same process of conduction [1]. Over the last years, the number of conducted secondary studies in Software Engineering (SE) has increased substantially [1], [2]. Despite the high adoption and importance of secondary studies in SE, the conduction of secondary studies still being time-consuming and demanding considerable human effort due to the manual process involved [3].

Two of the most labor-intensive and time-consuming activities in the process of conduction of a secondary study are the search and selection of studies activities [4], [5]. Considering that a secondary study's inputs are primary studies, locating and selecting relevant primary evidence is fundamental to impartially answer the proposed research question(s) [1]. During the search for studies activity, SE researchers usually perform searches in computer science Digital Libraries (DLs)

such as *Scopus*, *IEEE Xplore*, *ACM DL*, *Web of Science*, among others. To realize this activity, they made use of search queries [1]. On the other hand, these available DLs present limitations and not enough mechanisms to support the search activity [6], [7]. Snowballing is another search technique widely applied during the search activity. It consists in looking for relevant studies through references and citations analysis [8]. Looking for citations and references of studies can demand a lot of human effort. Regarding the selection activity, the most prominent issue is the large amount of primary studies to be read and analyzed [9], mainly with the rapid increase of primary research publication [10], [11].

There are several initiatives [3] and available tools [12] for automating or semi-automating activities of the SLR process. However, automation of the SLR activities is missing [13], [14]. For example, the study conducted by Al-Zubidy et al. [14] which prioritizes value-added requirements for SLR tool infrastructure, highlighted the need for automation for the search execution and study selection activities. However, it is not clear what are the existing automation approaches explored by researchers to support the activities of search and selection of studies for secondary studies in SE. To the best of our knowledge, there is not a secondary study focused on the automation of the search and selection of studies for the SLR process in SE.

In this study, we present a systematic mapping on automated support for searching and selecting studies for secondary studies in SE. We provide a synthesis of the existing approaches and tools to support the activities of search and selection of studies. Considering the establishment of the application of Text Classification (TC) approaches to support SLR's automation [15]–[19], we focused our investigation on the adoption of TC approaches to present insights for future research on automation of the secondary studies' search and selection activities. Since there is an increasing acceptance of the use of TC in medicine [20]–[23], we expand our SLR search to a cross-domain analysis also mapping available evidence from medicine on automation support to search and select studies for SLRs.

The main contributions of our study include: (i) a full detailed catalog on existing automation support to search and selection activities for secondary studies in SE; (ii) an up-

to-date systematic investigation on the application of TC to automate challenges on the search and selection activities; (iii) a summary of the approaches and metrics used to evaluate the performance of the explored TC approaches (iv) a discussion on emerging findings and implications for future research. Several approaches and tools have been explored and implemented in SE and the medical field. On the one hand, SE explored a wider variety of TC approaches. On the other hand, medicine has more well-established results on practical application of TC approaches to automate the challenges faced during the search and selection activities. Integrated solutions addressing both search and selection activities as well as the automation of another type of search strategy such as the snowballing technique can bring relevant advantages on workload and time saving for SE researchers who conduct secondary studies.

The remainder of this study is organized as follows: Section II details the SLR protocol followed in this study. Section III answers the proposed research questions. Section IV discusses our results, presents the study's limitations and discusses related work. Finally, Section V concludes our study.

II. STUDY DESIGN

In this section, we present the key aspects of the study design.

A. Research Questions

In order to facilitate understanding, we translated our research goal into three Research Questions (RQs):

RQ1: *What are the existing approaches and tools to support the search and selection of studies for secondary studies in SE?*

RQ2: *Which text classification approaches have been explored to automate the search and selection of studies for secondary studies in SE and medicine?*

RQ3: *What methods and metrics are used to assess the performance of the applied text classification approaches?*

B. Search Strategy

The adopted search strategy includes a two-stage search: An automatic search and a snowballing search [8]. Our two-stage search process and its results are illustrated in Figure 1. To perform the automatic search, we developed a search query and we ran a search pilot test as recommended by Kitchenham *et al.* [1]. Our search query is described next.

```
((("systematic review automat*" OR "SLR tool" OR "literature review automat*")) OR
(("text classification" AND "machine learning") AND ("systematic review" OR
"literature review" OR "systematic mapping")))
```

We chose to run our search query on the most renowned DLs in SE [1]: *IEEE Xplore*, *ACM DL*, *Scopus* and *Web of Science*. *Scopus* and *Web of Science* were chosen because they index studies of several international publishers, including

Springer, *Wiley-Blackwell*, *Elsevier*, *IEEE Xplore* and *ACM DL*; although not necessarily the most recent conference proceedings. Therefore, we opted for searching at *IEEE Xplore* and *ACM DL* individually because they are considered the two-key publisher-specific resources which together covers the most important SE and computer science conferences [1]. We executed the search query in three metadata fields: title, abstract and keywords. The search query was adapted to meet specific search criteria (e.g. syntax) of each DL.

Our selection criteria is organized into three Inclusion Criteria (IC) and five Exclusion Criteria (EC):

- **IC1:** The study must present an automation approach or tool applied to support the activities of search and selection of studies; AND
- **IC2:** The study must be within the SE or medicine domain; AND
- **IC3:** The study must present results from automating approach addressed.
- **EC1:** The study is just published as an abstract; OR
- **EC2:** The study is not written in English; OR
- **EC3:** The study is an older version of another study already considered; OR
- **EC4:** The study does not discuss approaches or strategies to automate the search and selection of studies; OR
- **EC5:** The study is not a primary study, such as tutorials, keynotes, editorials, etc.

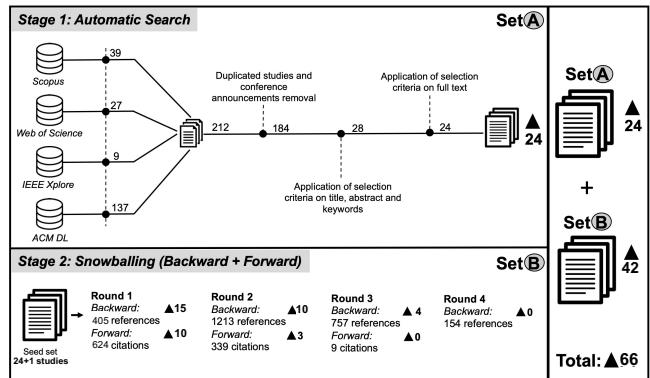


Figure 1. Search strategy process

The adoption of SLRs in SE emerged from the medical field [24]. medicine has been adopting SLRs since long before SE and it presents several advancements regarding SLR process automation. We opted to consider in our analysis studies that address search and selection automation for SLRs from medicine in order to map potential TC techniques employed in the medicine domain that could be explored in the SE context. In order to ensure the inclusion of relevant studies from medicine, in addition to selecting medicine related studies returned by the selected DLs, we added a well-known SLR on TC from the medicine [23] to our Snowballing “seed set”, as preformed in [25].

As illustrated in Figure 1 – Stage 1, a total of 212 items were returned from the automated search execution. Then, we

removed all duplicated studies and conference announcements, totaling 184 studies. Next, we read the papers' title, abstract and keywords and applied the IC and EC criteria on these fields which reduced our number to 28 candidate studies. Finally, the selection criteria were applied considering the reading of each study full text, resulting in a set of 24 included studies from this stage. This step was performed by the first author, and revised by the second author (100% of agreement).

The starting point of the snowballing technique is to define a “seed set” of relevant studies [8]. We considered as “seed set” the 24 included studies from the automated search strategy and the SLR from the medical field [23] (24+1 studies). Next, we performed snowballing forward and backward considering the citations and references’ list of the included studies, respectively. The studies’ citations were extracted with support of search engines, such as *Google Scholar*. In each snowballing iteration, we applied IC and EC criteria first on title, abstract and keywords, and next on full text. We performed four backward snowballing iterations and three forward snowballing iterations, stopping their execution when no more relevant study was detected. The results from each snowballing iteration can be observed in Figure 1 – Stage 2. As final result from the snowballing technique, 42 new studies were added to our set of included studies.

In total, 66 studies were included (Stage 1: 24 studies + Stage 2: 42 studies). We believe that the discrepancy between the number of included studies in stages 1 and 2 is due to the non-standardization of the SE terminology, which reflected in the construction of the search string [8], even after the search pilot test and the calibration of the string through the studies included in [26]. Thus, the adoption of Snowballing proved to be essential for the inclusion of relevant evidence. From the 66 included studies, coincidentally 33 are from the SE domain and 33 from the medicine domain. The final list of included studies is available at: <https://bit.ly/3jFJJAW>.

C. Data Extraction and Analysis

We created a data extraction form based on our RQs goals. The data extraction form contains all the fields necessary to enable analysis and synthesis from the data extracted to answer the RQs.

In Table I, we summarized the content of our data extraction form as well as the rationality of the extracted content. The data synthesis was performed through a combination of qualitative and quantitative analysis. The data synthesis results are presented as answers to our RQs in Section III.

III. RESULTS

In the following Sections we answer our proposed RQs.

A. RQ1: What are the existing approaches and tools to support the search and selection of studies for secondary studies in SE?

To answer the RQ1, we divided our analysis into two focuses: (i) proposed automation approaches to support the activities of search and selection of secondary studies; and (ii) general and

Table I
SUMMARY OF THE DATA EXTRACTION FORM

Category	Rationale	Addressed RQs
Study metadata	Identification and management of the study to detect the domain and publication data from the study.	RQ1, RQ2, RQ3
Search and selection automation approaches and tools in SE	Identification of automated approaches and tools that fully or partially support the activities of search and selection of studies for secondary studies in SE.	RQ1
Text classification approaches	Identification of approaches and metrics of text classification approaches for searching and selecting studies for secondary studies in SE and medicine.	RQ2, RQ3

specific tools that automate the search and selection activities. As general tools, we considered tools that address automation of several activities of the SLR process including the search and selection activities, and as specific tools, we considered tools that address only specific challenges faced during the performance of the search and selection activities.

(i) Automated approaches: In the SE field, several approaches have been investigated to provide automated support to the activities of search and selection of studies for secondary studies. As can be seen in Figure 2, the majority of automated approaches address the activity of selecting studies (14 studies – [16], [18], [19], [27]–[37], followed by approaches that support the searching for studies (4 studies - [7], [38]–[40]). Only one study combined in an integrated approach an automated solution to support both search and selection activities [17].

Finding 1: *Integrated approaches addressing the activity of search and selection of studies together have not been widely explored by researchers.*

Due to space limitations, in an external document¹ we present a spreadsheet with data details from each approach presented in Figure 2 including a brief description of the approach, evaluation method, corpus considered in the evaluation method, results/conclusions and future work.

Visual Text Mining (VTM) was first introduced in the SE field in 2007 by Malheiros et al. [27] and further explored by Felizardo et al. [16], [41] to aid the selection activity. VTM is also investigated in the context of SLR updates [18], [42] and in the context of the search activity to assist the construction of search strings.

TC techniques addressing the use of Text Mining (TM), Natural Language Processing (NLP) and Machine Learning (ML) are strongly adopted by researchers to automate the selection of studies. The most adopted ML models with promising results involves supervised ML models such as Support Vector Machines (SVM) [17]–[19], [34] and/or active learning [17], [33], [36], [43]. Variations of the Naïve Bayes classifier have been also explored [32], [34], [37] as well as Hybrid Feature Selection Method (HFSM) combined with other

¹<https://bit.ly/3AoiSPw>

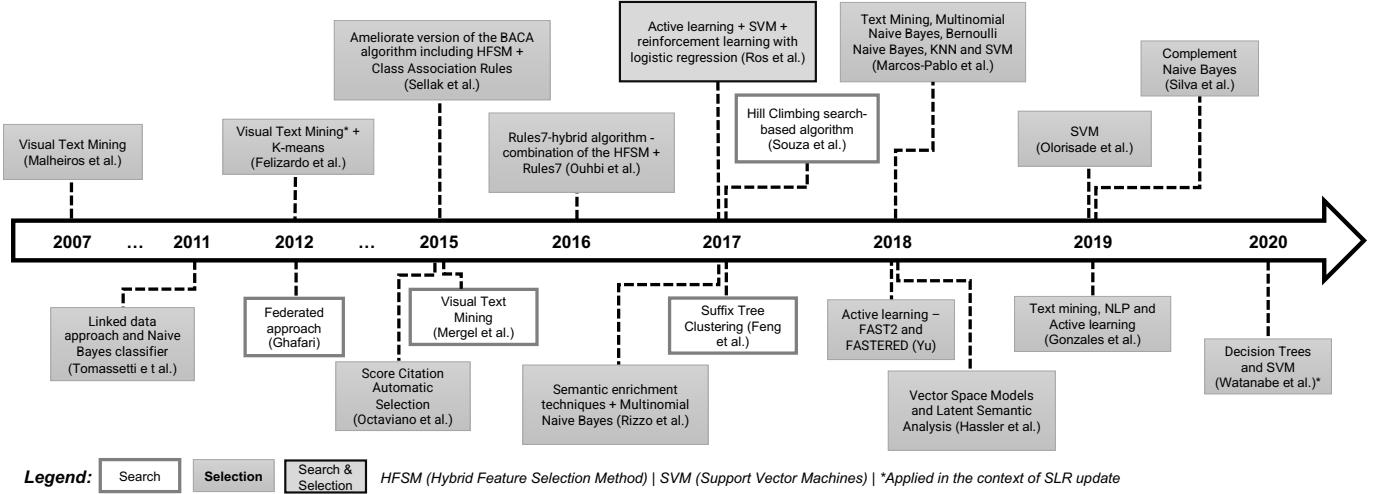


Figure 2. Timeline of existing automated approaches for searching and selecting studies in SE.

algorithms such as the hierarchical low-rank decomposition Blocked Adaptive Cross Approximation (BACA) [30] and the classical Rules7 [31]. Besides VTM, Suffix Tree Clustering and the optimization search algorithm Hill Climbing [39] are employed to automate the search for studies. Unlike other approaches, Ghafari et al. [7] propose a federated search approach to automatically integrate search mechanisms across well-known SE DLs.

Finding 2: *SVM and active learning are two of the most recent adopted models that showed promising results in their application to support the automation of the selection of studies activity in SE.*

(ii) General and Specific tools in SE: In Table II, we detail general and specific SLR tools that present some automation of the activities of search and selection. All presented tools support the activities of search and selection partially, none of them fully automated any of these activities.

Regarding the specific tools that directly address automation of the activities of search and selection (see Table II), we detected two specific tools that provide automation support to SLR search activity [38], [39] and three specific tools to support the SLR study selection activity [27], [33], [50]. Three of the five identified tools use VTM as automation technique [27], [38], [50].

B. RQ2: Which text classification approaches have been explored to automate the search and selection of studies for secondary studies in SE and medicine?

Over past 15 years, TC has gained significant attention in the secondary studies context. As one of the key TM approach (a.k.a document classification), TC can be defined as automatically assigning semantic labels to texts given a set of fixed semantic categories or classes [51]. In fact, the most common TC techniques combine TM approaches and ML models to automatically learn and categorize new data from previously categorised data [21].

We summarize in Table III the TC approaches identified in the selected primary studies categorising them according to application field (SE and medicine), respectively. Overall, the SE field explored more diverse TC approaches. On the other hand, medicine is more consolidated on the exploration of the Naïve Bayes and SVM approaches. According to our selected studies, approaches such as Rocchio, LDA, LMT, and neural network has not been explored by SE researches to address challenges related to search and selection of studies for secondary studies. In the medical field, our results show that approaches and models such as STC, HFSRM, VSM, LSA, reinforcement learning, DT, Rules7, BACA and VTM has not been explored yet.

The TC approaches mentioned in Table III, considered in the studies with different features and evaluation corpus. In some cases, different variants of the algorithms were adopted among the studies (e.g. Complement Naïve Bayes [20], [37], Multinomial Naïve Bayes and Bernoulli Naïve Bayes [34]). Besides, in almost all cases, the source code and the dataset used in the analysis were not made available. This fact prevented us of performing a comparison analysis among the identified approaches.

Two studies [19], [34] considered in their analysis data from both medicine and SE SLRs. For this reason, both studies are mentioned in the two columns (SE studies and medicine studies) in Table III. Several studies, in SE and medicine investigated more than one approach in the their study. For example, Almeida et al. [52] made use of three different classification approaches: Naïve Bayes, LMT and SVM; and in SE; and Hassler et al. [35] adopted two training-by-example classifiers, one based on VSM and a second one based on LSA.

Finding 3: *Despite the greater variety of tools and TC approaches explored in the SE field, the medical field shows a higher systematic evaluation and consolidated practical application of tools and approaches on search and selection activities.*

Table II
GENERAL AND SPECIFIC SE SLR TOOLS ADDRESSING THE SEARCH AND SELECTION ACTIVITIES

Tool	Search & Selection support	Year
General Tools		
SLR-TOOL [44]	Refinement of search using text mining; clustering studies thought similarities among them; exportation of data and references on EndNote, Bibtex and Ris formats.	2010
SLuRp [45]	Execution of search terms on some DLs; Semi-automatic extract and store studies' full text .pdf (if there are appropriate permissions); recording bibliographical data in Bibtext and Ris format; recording of assessment from reviewers as well as managing reviewer's selection and exclusions of studies.	2012
Slrtool [46]	Automatic extraction of the Bibtex data from the located studies and automatic download of full text studies .pdf (subject to permission of the host institutions); definition of the search criteria independent of target resource database; possibility of categorize studies and perform the management of the application inclusion and exclusion.	2014
SESRA [47]	Importation of search results from SE DLs (i.e. IEEE Xplore, IET Digital Library and SpringerLink) or though a Bibtex file; support on the consensus decision on the inclusion or exclusion of one study.	2015
StArt [9]	Support to the main online search databases, including Scopus, IEEE, ACM and Web of Science; automated calculation of an study's score based on keywords occurrences on title, abstract and keywords and number of citations; automatic detection of duplicated and similar studies; semi-automation of the Snowballing technique (under development).	2016
SLR Toolkit [48]	Simple literature filtering; design of a taxonomy; classification of studies; analysis of the classification by generated diagrams.	2018
SLR-Tool [49]	Importation of search results from DLs and evaluation of the quality of the search results; Management of search results by including or excluding each paper.	2020
Specific Tools		
PEx [27]	Projection Explorer (PEx) tool uses VTM to increase study selection efficiency and allow researchers to broaden their search algorithms to create a larger corpus, since the tool quickened the identification of irrelevant studies.	2007
ReViS [50]	ReViS uses VTM to support the selection task in systematic reviews.	2014
SLR.qub [38]	Automated support the researcher by suggesting new terms for the string using VTM algorithms.	2015
SLRPSS [39]	Unified search engine wrapper for the SLR DLs: IEEEXplore, ACM DL, the Web of Science, Science Direct, Scopus, and Google Scholar.	2017
FAST2 [33]	Automated support to studies selection to minimize efforts by using keywords to identify and rank relevant studies.	2018

Table III
TEXT CLASSIFICATION APPROACHES EXPLORED IN SE AND MEDICINE

TC approach	SE studies	medicine stud-ies
Naïve Bayes	[28], [34], [37]	[21], [34], [52]–[56]
Support Vector Machine (SVM)	[17], [17]–[19], [34]	[19], [19]–[21], [34], [52]–[55], [57]–[65]
K-Nearest Neighbor (KNN)	[16], [17], [17]–[19], [34]	[21], [34], [52], [54]
Rocchio	–	[21]
Suffix Tree Clustering (STC)	[39]	–
Active Learning	[17], [33], [36], [39]	[20]
Label spreading	[65]	[66]
Label propagation	[65]	[66], [67]
Hybrid Feature Selection Measure (HFSRM)	[30], [31]	–
Vector Space Models (VSM)	[35]	–
Latent Semantic Analysis (LSA)	[35]	–
Latent Dirichlet allocation (LDA)	–	[57]
Unsupervised K-means	[16]	[68]
Logistic Model Trees (LMT)	–	[52]
Reinforcement Learning	[17]	–
Decision Trees (DT)	[17], [18]	–
Rules7	[31]	–
Blocked Adaptive Cross Approximation (BACA)	[30]	–
Neural Network	–	[59], [67]
Visual Text Mining (VTM)	[16], [27], [38]	–

Our results demonstrates that researchers recognize the need of further exploration and validation of their proposed

approaches. The most observed type of future work in SE and medicine studies is the validation of the study' results with a larger dataset from the same field or from different fields; followed by the parameters' variation of the adopted ML models in order to improve results performance.

Finding 4: *SE researchers should further explore TC approaches (alone or combined) already applied to the SE field as well as approaches applied to medicine and not yet explored in SE; and vice-versa.*

During the conduction of our study, we identified eight different studies that directly reports the practical use and evaluation of well-established SLR selection (screening) tools: Abstrackr [69]–[71], RobotAnalyst [70], [72], DistillerSR [70], [73], ReIRank [74], and SWIFT-Review [75]. In contrast, in the SE field, only two tools StArt [9] and ReVis [50] from our selected studies have reported practical adoption. Some of the possible reasons that we have concluded for the low practical adoption of automated approaches or tools are because the automation solution is: (i) presented only as a prototype; (ii) not available online (e.g. broken access links); (iii) lack in exhaustive validation and documentation; or (iv) not easy to use.

Finding 5: *The vast majority of studies from both fields do not provide a replication package of the implemented approach or a workable link to access the proposed tool or the used corpus dataset.*

C. RQ3: What methods and metrics are used to assess the performance of the applied text classification approaches?

In this section we present the methods and metrics adopted to assess the TC approaches presented in our selected primary studies.

1) Methods to assess the results from TC approaches:

From the 66 selected primary studies, 46 studies presented an assessment form of the proposed TC approach. Considering that the majority of the studies presented results from the application of ML models, the two most adopted methods to assess the results from the applied TC approaches were cross validation followed by experiment.

Cross validation is performed dividing a sample of data in subsets, considering the analysis performed on a unique subset (training set) while other subsets (testing set) are kept to subsequent use to validate the analysis [21]. The most adopted type of cross validation present in our selected studies is N-fold cross validation which consists in divide the dataset into N equally-sized mutually-exclusive “folds” with one fold serving as the test set and the remaining N-1 folds to form the training set. This process is repeated, until each fold be used once as the training set. 10-fold cross validation was the predominate type of cross validation [17], [21], [30], [31], [54], [58], [76] followed by 5-fold cross validation [57], [59], [77]–[79] and 7-fold cross validation [34]. One different type of cross validation called Monte-Carlo cross validation [80] was adopted by Hassler et al. [35] which consists in randomly select a portion of the data as training set and the rest of the data is used as test set, repeating this process several times.

Another highly adopted form of assessment for TC approaches is experiment considering data from published SLRs performed manually [18], [19], [28], [33], [36], [37], [43], [55], [62], [64], [65], [68], [71], [73]–[75], [81]. In these studies, the authors usually have two or more groups of participants to emulate the search or selection process using the proposed automated approach and compare their results against the manually performed search or selection process [16], [27], [38], [39].

2) Adopted performance metrics: Our selected studies used several metrics to describe their results. Table IV describes each adopted metric with its a brief definition and the studies from each field that adopted the respective metric.

The most adopted metrics to evaluate the performance of the automation techniques are recall, precision and F-measure. 21 of 46 SE and medicine studies (45.65%) presented an assessment approach using these metrics. Work Saved over Sampling (WSS), a measure defined by Cohen et al. [59], was adopted in 11 medicine studies and only in one SE study. The Area Under the Curve (AUC), Burden, Yeld and Utility were applied only in medicine studies.

Finding 6: *In the context of search and selection activities for secondary studies adopting TC approaches, cross validation and experiment are the most chosen form of assessment considered. Recall, precision and, consequently F-measure have shown the most significant performance metrics.*

IV. DISCUSSION

Despite the several search and selection approaches presented in our study, the lack of automation is still present in these secondary studies’ activities. Efforts have been applied to reduce the search and selection workload and time-spent, but it still needs to reduce the human effort required to search and select studies for secondary studies in SE.

The majority of the proposed search and selection automated approaches presented some validation. Cross-validation and experiment are the most adopted types of validation (see Section III-C). However, each study validated their proposed approaches considering a limited number of sources (e.g. search only in one DL such as *Scopus* or *IEEEExplore*) and population (e.g. few SLRs studies from different SE and medicine domains). These facts prevent an accurate comparison of efficiency and workload reduction among the proposed approaches. Large-scale and exhaustive validation is needed to support results obtained through preliminary analysis and demonstrate the real applicability and benefits of the proposed approaches in the SE field.

There is a need for a better dissemination and adoption in practice of automated and search and selection tools and approaches in SE. In this way, researchers could be able to obtain a practical validation of their tools and approaches as well as to get valuable feedback by end users, enabling the prioritization of value-added requirements for improvements.

The combination of TM and ML applied to automate or semi-automate the SLR search and selection activities provides cost savings and allows replicability [17]. One known difficulty concerning the use of TC approaches is that most supervised learning approaches used in these studies rely on a dataset for training the ML model [18]. Considering this fact, in the scenario of SLR update which the dataset for training is already known (original SLR selected studies), TC techniques can be promising. The works of Watanabe et al. [18] and Felizardo et al. [50] exemplify in their results the potential of TC techniques applied on the study selection activity during SLR update.

Our results highlighted the adoption of the TC approaches to support secondary studies’ search and selection activities. However, selecting the most appropriate ML algorithm, related methods and text sections (e.g. title, abstract, keywords, references, etc.) are fundamental to bring results with a high recall and precision.

Integrated solutions to automate the search and selection activities for secondary studies using TM and ML approaches is the most suitable combination of approaches since they can bring several benefits such as: researchers do not need to build, calibrate and adapt search strings to meet DLs requirements; the search and selection activity can be automatically executed by the automated tool (once the ML model is trained) enabling time saving, reduction of manual effort and possibly avoiding human-error; researchers can keep track of the whole search and selection decisions through the tool’s records, leading to a more complete and transparent protocol documentation [17].

Furthermore, our results (see Section III-C) show quantitative and qualitative approaches that have been used to demonstrate

Table IV
ASSESSMENT METRICS FOR TEXT CLASSIFICATION APPROACHES [23]

Metric	Definition	SE Studies	MED Studies
Precision	Ratio of correctly identified relevant studies to all of those predicted as relevant.	[18], [28], [30], [31], [34], [35], [37], [66]	[19], [21], [53]–[59], [63], [65], [81]–[83]
Recall (or Sensitivity)	Ratio of correctly predicted relevant studies to all relevant ones.	[18], [28], [30], [31], [33]–[35], [37]	[19], [21], [53]–[59], [63], [65], [66], [81]–[83]
F-Measure	Combines Precision and Recall values. It corresponds to the harmonic mean of Precision and Recall.	[18], [30], [31], [34], [35], [37]	[19], [21], [53]–[59], [63], [65], [66], [81]–[83]
Accuracy	Ratio of included and commonly excluded studies with the combination of included and excluded studies.	[17]	[55], [57], [62]
WSS@95% (Work Saved over Sampling)	The percentage of studies that the reviewers do not have to read because they have been screened out by the classifier considered at 95% recall.	[33], [43]	[19], [57], [63], [65], [72], [75]–[78]
Area Under the Curve (AUC)	Area under the curve obtained by graphing the true positive rate against the false positive rate; 1.0 is a perfect score and 0.5 is equivalent to a random ordering.	–	[20], [57], [60], [61]
Burden	The fraction of the total number of studies that a human must screen.	–	[20], [64], [67], [73]
Yield	The fraction of studies that are identified by a given screening approach.	–	[20], [64], [67]
Utility	It is a weighted sum of Yield and Burden. Here, β is a constant. It represents the relative importance of Yield in comparison to Burden.	–	[20], [64], [67]

the efficiency of the proposed automated solutions. Since recall and precision are the most adopted metrics to analyze the efficiency of automated proposals, we state the adoption of these metrics in all studies addressing secondary studies automation, especially to enable comparison results from different researches. Our observations corroborate with [25] on considering that in the TC adoption scenario, it is fundamental that the authors make available online the data used in their evaluation and the replication package of the study to enable detailed comparisons and study replication.

A. Threats to Validity

In this section, we describe the main threats to validity of our study as well as the adopted mitigation strategies.

Construct validity – Some of the relevant studies related to the topic of this study could have been missed, especially from medicine. To mitigate this limitation, we performed a two-stage search strategy including a pilot search adopting a study control group to calibrate our search string. Our automated search strategy did not cover medical databases. However, besides the medicine studies returned by our selected DLs, we considered in our Snowballing “seed set” a well-known (over 300 citations) and widespread (93 references) SLR on TC for SLRs from the medicine domain. We believe that most of the relevant studies were covered by our search strategy.

Conclusion validity – This threat addresses the results and conclusions presented in this study. To mitigate this threat, we systematically created the data extraction form emerging from the RQs as well as we interactively performed refinements on it during the data extraction execution aiming to reduce potential biases during the extraction process. Lastly, due to the different datasets, approaches and assessment metrics from the studies considered in our analysis as well as the insufficient information for study replication, it was not possible to establish a impartial

performance in the results among the reported tools and TC approaches.

B. Related Work

Felderer and Travassos [26] present in their book a literature survey on strategies used to automated the SLR process in SE. Unlike our study, the authors provided a general summary addressing all SLR process activities. Similarly to this work, in [3] is reported a SM providing an general overview of tools to support the whole SLR process. Unlike both work, our study is focused on a detailed exploration of the activities of search and selection In addition, the summary of the search and selection automation strategies presented in the Felderer and Travassos books’ [26], motivated us to further explore TM and ML applied on the secondary study search and selection context. Lastly, our study addresses evidence from the medical field providing a cross-domain mapping that can contribute with future research on reducing the manual effort in the search and selection activities for SLRs in SE.

Olorisade et al. [25] critically analyzed the use of TM approaches used to support the activity of searching for studies. Similarly to our study, they considered the work of O’Mara-Eves et al. [23] from the medical domain to analyze results from multiple domains. However, the search performed in their work considered only evidence published until February 2014, while our study considers studies published until December 2020. Besides, our work analyze approaches addressing not only the selection activity, but also the search activity.

V. CONCLUSIONS

This study provides a synthesis of the existing approaches to support the search and selection of primary sources for secondary studies.

Our results show that SE researchers explored several TC approaches for searching and selecting evidence for secondary studies. On the other hand, medicine has more well-established

results and practical application of TC approaches to automate the challenges from the search and selection activities. Integrated solutions addressing both search and selection activities as well as the automation of another type of search strategy such as the snowballing technique can bring relevant advantages on workload and time saving for SE researchers who conduct secondary studies.

As future work, we intend to investigate integrated search and selection approaches and apply them on different SLR contexts. Also, we intend to deeply analyze and develop the search and selection approaches from the medical field that are not investigated in SE yet.

REFERENCES

- [1] B. Kitchenham, D. Budgen, and P. Brereton, *Evidence-Based Software Engineering and Systematic Reviews*, ser. Chapman & Hall/CRC Innovations in Software Engineering and Software Development Series. Chapman & Hall/CRC, 2015.
- [2] F. da Silva, A. Santos, S. Soares, A. França, and C. Monteiro, "Six years of systematic literature reviews in software engineering: an extended tertiary study," in *ICSE*. "": IEEE Computer Society, 2010, pp. 1–10.
- [3] C. Marshall and P. Brereton, "Tools to support systematic literature reviews in software engineering: A mapping study," in *ESEM*, 2013, pp. 296–299.
- [4] A. Al-Zubidy and J. Carver, "Identification and prioritization of slr search tool requirements: an slr and a survey," *Empirical Software Engineering*, vol. 24, 02 2019.
- [5] E. Hassler, J. Carver, N. Kraft, and D. Hale, "Outcomes of a community workshop to identify and rank barriers to the systematic literature review process," in *EASE*. ACM, 2014, pp. 1–10.
- [6] S. Imtiaz, M. Bano, N. Ikram, and M. Niazi, "A tertiary study: Experiences of conducting systematic literature reviews in software engineering," in *EASE*. ACM, 2013, pp. 177–182.
- [7] M. Ghafari, M. Saleh, and T. Ebrahimi, "A federated search approach to facilitate systematic literature review in software engineering," *International J. of Software Engineering & Applications*, vol. 3, 04 2012.
- [8] C. Wohlin, "A snowballing procedure for systematic literature studies and a replication," in *EASE*, 2014, pp. 321–330.
- [9] S. Fabbri, C. Silva, E. Hernandes, F. Octaviano, A. Di Thommazo, and A. Belgamo, "Improvements in the start tool to better support the systematic review process," in *EASE*. New York, NY, USA: ACM, 2016.
- [10] K.-J. Stol and B. Fitzgerald, "A holistic overview of software engineering research strategies," in *CESI*. IEEE Press, 2015, p. 47–54.
- [11] L. Zhang, J.-H. Tian, J. Jiang, Y. Liu, M.-Y. Pu, and T. Yue, "Empirical research in software engineering — a literature survey," *J. of Computer Science and Technology*, vol. 33, pp. 876–899, 2018.
- [12] C. Marshall, P. Brereton, and B. Kitchenham, "Tools to support systematic reviews in software engineering: A cross-domain survey using semi-structured interviews," in *EASE*. New York, NY, USA: ACM, 2015, pp. 26:1–26:6.
- [13] C. Marshall, B. Kitchenham, and P. Brereton, "Tool features to support systematic reviews in software engineering – a cross domain study," *e-Informatica Software Eng. Journal*, vol. 12, no. 1, pp. 79–115, 2018.
- [14] A. Al-Zubidy, J. C. Carver, D. P. Hale, and E. E. Hassler, "Vision for slr tooling infrastructure: Prioritizing value-added requirements," *Information and Software Technology*, vol. 91, pp. 72 – 81, 2017.
- [15] D. Cruzes and T. Dybå, "Synthesizing evidence in software engineering research," in *ESEM*, 2010, pp. 1–10.
- [16] K. Felizardo, G. Andery, F. Paulovich, R. Minghim, and J. Maldonado, "A visual analysis approach to validate the selection review of primary studies in systematic reviews," *Information and Software Technology*, vol. 54, no. 10, pp. 1079–1091, 2012.
- [17] R. Ros, E. Bjarnason, and P. Runeson, "A machine learning approach for semi-automated search and selection in literature studies," in *EASE*, ser. EASE'17. New York, NY, USA: ACM, 2017, p. 118–127.
- [18] W. M. Watanabe, K. R. Felizardo, A. Candido, E. F. de Souza, J. ao Ede de Campos Neto, and N. L. Vijaykumar, "Reducing efforts of software engineering systematic literature reviews updates using text classification," *Information and Software Technology*, vol. 128, p. 106395, 2020.
- [19] B. K. Olorisade, P. Brereton, and P. Andras, "The use of bibliography enriched features for automatic citation screening," *J. of Biomedical Informatics*, vol. 94, p. 103202, 2019.
- [20] M. Miwa, J. Thomas, A. O'Mara-Eves, and S. Ananiadou, "Reducing systematic review workload through certainty-based screening," *J. of Biomedical Informatics*, vol. 51, pp. 242–253, 2014.
- [21] J. García Adeva, J. Pikatza Atxa, M. Ubeda Carrillo, and E. Ansuategi Zengotitabengoa, "Automatic text classification to support systematic reviews in medicine," *Expert Systems with Applications*, vol. 41, no. 4, Part 1, pp. 1498 – 1508, 2014.
- [22] T. Bekhuis, E. Tseytin, K. J. Mitchell, and D. Demner-Fushman, "Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence," *PLOS ONE*, vol. 9, pp. 1–10, 01 2014.
- [23] A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou, "Using text mining for study identification in systematic reviews: A systematic review of current approaches," *Systematic Reviews*, vol. 4, 01 2015.
- [24] B. Kitchenham, "Procedures for performing systematic reviews," Software Engineering Group - Department of Computer Science - Keele University and Empirical SE - National ICT Australia Ltd, Joint Technical Report TR/SE-0401 (Keele) - 0400011T.1 (NICTA), 2004.
- [25] B. K. Olorisade, E. de Quincey, P. Brereton, and P. Andras, "A critical analysis of studies that address the use of text mining for citation screening in systematic reviews," in *EASE*. New York, NY, USA: ACM, 2016.
- [26] M. Felderer and G. H. Travassos, *Contemporary Empirical Methods in Software Engineering*. Springer, 2020.
- [27] V. Malheiros, E. Hohn, R. Pinho, M. Mendonca, and J. Maldonado, "A visual text mining approach for systematic reviews," in *ESEM*. ACM, 2007, pp. 245–254.
- [28] F. Tomassetti, G. Rizzo, A. Vetro, L. Ardito, M. Torchiano, and M. Morisio, "Linked data approach for selection process automation in systematic reviews," in *EASE*, 2011, pp. 31–35.
- [29] F. Octaviano, K. Felizardo, J. Maldonado, and S. Fabbri, "Semi-automatic selection of primary studies in systematic literature reviews: is it reasonable?" *Empirical Software Engineering (Dordrecht. Online)*, pp. 1–20, 2014.
- [30] H. Sellak, B. Ouhbi, and B. Frikh, "Using rule-based classifiers in systematic reviews: A semantic class association rules approach," in *iiWAS*. New York, NY, USA: ACM, 2015.
- [31] B. Ouhbi, M. Kamoune, B. Frikh, E. M. Zemmouri, and H. Behja, "A hybrid feature selection rule measure and its application to systematic review," ser. iiWAS '16. New York, NY, USA: ACM, 2016, p. 106–114.
- [32] G. Rizzo, A. Vetro, L. Ardito, M. Torchiano, and R. Troncy, "Semantic enrichment for recommendation of primary studies in a systematic literature review," *Digital Scholarship in the Humanities*, vol. 32, pp. 195–208, 04 2017.
- [33] Z. Yu and T. Menzies, "FAST2: an intelligent assistant for finding relevant papers," *Expert Systems with Applications*, vol. 120, 11 2018.
- [34] S. Marcos-Pablos and F. García-Péñalvo, "Information retrieval methodology for aiding scientific database search," *Soft Computing*, vol. 24, 04 2020.
- [35] E. E. Hassler, D. P. Hale, and J. E. Hale, "A comparison of automated training-by-example selection algorithms for evidence based software engineering," *Information and Software Technology*, vol. 98, pp. 59–73, 2018.
- [36] S. Gonzalez-Toral, R. Freire, R. GualÃ;n, and V. Saquicela, "A ranking-based approach for supporting the initial selection of primary studies in a systematic literature review," 09 2019, pp. 1–10.
- [37] G. Silva, P. Santos Neto, R. Santos Moura, I. C. Aratijo, O. Cury da Costa Castro, and I. Ibiapina, "An approach to support the selection of relevant studies in systematic review and systematic mappings," in *2019 8th Brazilian Conference on Intelligent Systems*, 2019, pp. 824–829.
- [38] G. D. Mergel, M. S. Silveira, and T. S. da Silva, "A method to support search string building in systematic literature reviews through visual text mining," in *SAC*, ser. SAC '15. New York, NY, USA: ACM, 2015, p. 1594–1601.
- [39] L. Feng, Y. Chiam, E. Abdullah, and U. Obaidellah, "Using suffix tree clustering method to support the planning phase of systematic literature review," *Malaysian Journal of Computer Science*, vol. 30, pp. 311–332, 12 2017.
- [40] F. C. Souza, A. Santos, S. Andrade, R. Durelli, V. Durelli, and R. Oliveira, "Automating search strings for secondary studies," *Information Technology - New Generations*, pp. 839–848, 07 2017.

- [41] K. Felizardo, N. Salleh, R. Martins, E. Mendes, S. MacDonell, and J. Maldonado, "Using visual text mining to support the study selection activity in systematic literature reviews," in *ESEM*. ACM, 2011, pp. 1–10.
- [42] K. Felizardo, S. MacDonell, E. Mendes, and J. Maldonado, "A systematic mapping on the use of visual data mining to support the conduct of systematic literature reviews," in *J. of Software*, vol. 7, no. 2. Academy Publisher, 2011, pp. 450–461.
- [43] Z. Yu, N. Kraft, and T. Menzies, "Finding better active learners for faster literature reviews," *Empirical Software Engineering*, vol. 23, 12 2018.
- [44] M. Fernandez-Saez, M. Bocco, and F. Romero, "SLR-tool – a tool for performing systematic literature reviews," in *ICSOFT*, 2010, pp. 157–166.
- [45] D. Bowes, T. Hall, and S. Beecham, "SLuRp: a tool to help large complex systematic literature reviews deliver valid and rigorous results," in *EAST*, 2012, pp. 33–36.
- [46] B. Barn, F. Raimondi, L. Athappian, and T. Clark, "SLRTool: a tool to support collaborative systematic literature reviews," in *ICEIS*, 2014, pp. 440–447.
- [47] J. S. Molléri and F. B. V. Benitti, "Sesra: A web-based automated tool to support the systematic literature review process," in *EASE*. New York, NY, USA: ACM, 2015.
- [48] S. Götz, "Supporting systematic literature reviews in computer science: The systematic literature review toolkit," ser. MODELS '18. New York, NY, USA: ACM, 2018, p. 22–26.
- [49] A. Hinderks, F. J. D. Mayo, J. Thomaschewski, and M. J. Escalona, "An slr-tool: Search process in practice: A tool to conduct and manage systematic literature review (slr)," in *ICSE Companion*. New York, NY, USA: ACM, 2020, p. 81–84.
- [50] K. Felizardo, E. Nakwgawa, S. MacDonell, and J. Maldonado, "A visual analysis approach to update systematic reviews," in *EASE*. "": ACM, 2014, pp. 1–10.
- [51] T. Joachims, "Transductive inference for text classification using support vector machines," in *ICML*, ser. ICML '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, p. 200–209.
- [52] H. Almeida, M.-J. Meurs, L. Kosseim, and A. Tsang, "Data sampling and supervised learning for hiv literature screening," *IEEE Transactions on NanoBioscience*, vol. 15, pp. 1–1, 05 2016.
- [53] Y. Aphinyanaphongs, I. Tsamardinos, A. Statnikov, D. Hardin, and C. Aliferis, "Text categorization models for high-quality article retrieval in internal medicine," *J. of the American Medical Informatics Association : JAMIA*, vol. 12, pp. 207–16, 01 2005.
- [54] T. Bekhuis and D. Demner-Fushman, "Screening nonrandomized studies for medical systematic reviews: A comparative study of classifiers," *Artif. Intell. Med.*, vol. 55, no. 3, p. 197–207, Jul. 2012.
- [55] E. Popoff, M. Besada, J. Jansen, S. Cope, and S. Kanter, "Aligning text mining and machine learning algorithms with best practices for study selection in systematic literature reviews," *Systematic Reviews*, vol. 9, 12 2020.
- [56] O. Frunza, D. Inkpen, S. Matwin, W. Klement, and P. O'Brien, "Exploiting the systematic review protocol for classification of medical abstracts," *Artif. Intell. Med.*, vol. 51, no. 1, pp. 17–25, 2011.
- [57] A. Bannach-Brown, P. Przybyla, J. Thomas, A. Rice, S. Ananiadou, J. Liao, and M. Macleod, "Machine learning algorithms for systematic review: Reducing workload in a preclinical review of animal studies and reducing human screening error," *Systematic Reviews*, vol. 8, 01 2019.
- [58] T. Bekhuis and D. Demner-Fushman, "Towards automating the initial screening phase of a systematic review," *Studies in health technology and informatics*, vol. 160, pp. 146–50, 01 2010.
- [59] S. Götz, "An effective general purpose approach for automated biomedical document classification," in *AMIA Annual Symposium proceeding*. American Medical Informatics Association, 2006, pp. 161–5.
- [60] A. Cohen, K. Ambert, and M. McDonagh, "Cross-topic learning for work prioritization in systematic review creation and update," *J. of the American Medical Informatics Association : JAMIA*, vol. 16, pp. 690–704, 07 2009.
- [61] ——, "A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review," *AMIA Annual Symposium proceedings*, vol. 2010, pp. 121–5, 11 2010.
- [62] S. Kim and J. Choi, "Improving the performance of text categorization models used for the selection of high quality articles," *Healthcare informatics research*, vol. 18, pp. 18–28, 03 2012.
- [63] P. Timsina, J. Liu, and O. El-Gayar, "Advanced analytics for the automation of medical systematic reviews," *Inf Systems Frontiers*, vol. 18, 08 2015.
- [64] B. Wallace, T. Trikalinos, J. Lau, C. Brodley, and C. Schmid, "Semi-automated screening of biomedical citations for systematic reviews," *BMC Bioinformatics*, vol. 11, p. 55, 01 2010.
- [65] P. Timsina, J. Liu, O. El-Gayar, and Y. Shang, "Using semi-supervised learning for the creation of medical systematic review: An exploratory analysis," in *HICSS*, 2016, pp. 1195–1203.
- [66] J. Liu, P. Timsina, and O. El-Gayar, "A comparative analysis of semi-supervised learning: The case of article selection for medical systematic reviews," *Inf. Systems Frontiers*, vol. 20, no. 2, p. 195–207, Apr. 2018.
- [67] G. Kontonatsios, A. J. Brockmeier, P. Przybyla, J. McNaught, T. Mu, J. Y. Goulermas, and S. Ananiadou, "A semi-supervised approach using label propagation to support citation screening," *J. of Biomedical Informatics*, vol. 72, pp. 67–76, 2017.
- [68] Z. Xiong, T. Liu, G. Tse, M. Gong, P. Gladding, B. Smaill, M. Stiles, A. Gillis, and J. Zhao, "A machine learning aided systematic review and meta-analysis of the relative risk of atrial fibrillation in patients with diabetes mellitus," *Frontiers in Physiology*, vol. 9, 07 2018.
- [69] A. Gates, M. Gates, M. Sebastianski, S. Guitard, S. Elliott, and L. Hartling, "The semi-automation of title and abstract screening: A retrospective exploration of ways to leverage abstrackr's relevance predictions in systematic and rapid reviews," *BMC Medical Research Methodology*, vol. 20, 06 2020.
- [70] A. Gates, S. Guitard, J. Pillay, S. Elliott, M. Dyson, A. Newton, and L. Hartling, "Performance and usability of machine learning for screening in systematic reviews: A comparative evaluation of three tools," *Systematic Reviews*, vol. 8, p. 278, 11 2019.
- [71] J. Rathbone, T. Hoffmann, and P. Glasziou, "Faster title and abstract screening? evaluating abstrackr, a semi-automated online screening program for systematic reviewers," *Systematic reviews*, vol. 4, p. 80, 06 2015.
- [72] P. Przybyla, A. Brockmeier, G. Kontonatsios, M.-A. Le Pogam, J. McNaught, E. Elm, K. Nolan, and S. Ananiadou, "Prioritising references for systematic reviews with robotanalyst: A user study," *Research Synthesis Methods*, vol. 9, pp. 470–488, 07 2018.
- [73] C. Hamel, K. Thavorn, G. Wells, and B. Hutton, "An evaluation of distillersrâs machine learning-based prioritization tool for title/abstract screening â impact on reviewer-relevant outcomes," *BMC Medical Research Methodology*, vol. 20, 10 2020.
- [74] T. K. Saha, M. Ouzzani, H. M. Hammady, A. K. Elmagarmid, W. Diffl, and M. A. Hasan, "A large scale study of svm based methods for abstract screening in systematic reviews," 2018.
- [75] B. Howard, J. Phillips, K. Miller, A. Tandon, D. Mav, M. Shah, S. Holmgren, K. Pelch, V. Walker, A. Rooney, M. Macleod, R. Shah, and K. Thayer, "Swift-review: A text-mining workbench for systematic review," *Systematic Reviews*, vol. 5, 05 2016.
- [76] G. Kontonatsios, S. Spencer, P. Matthew, and I. Korkontzelos, "Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews," *Expert Systems with Applications: X*, vol. 6, p. 100030, 2020.
- [77] S. Matwin, A. Kouznetsov, D. Inkpen, O. Frunza, and P. O'Brien, "A new algorithm for reducing the workload of experts in performing systematic reviews," *J. of the American Medical Informatics Association : JAMIA*, vol. 17, pp. 446–53, 07 2010.
- [78] A. Cohen, K. Ambert, and M. McDonagh, "Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the wss@95 measure," *J. of the American Medical Informatics Association : JAMIA*, vol. 18, pp. 104–05, 2011.
- [79] B. K. Olorisade, P. Brereton, and P. Andras, "Reproducibility of studies on text mining for citation screening in systematic reviews: Evaluation and checklist," *J. of Biomedical Informatics*, vol. 73, pp. 1–13, 2017.
- [80] R. R. Picard and R. D. Cook, "Cross-validation of regression models," *J. of the American Statistical Association*, vol. 79, no. 387, pp. 575–583, 1984.
- [81] G. Tsafnat, P. Glasziou, G. Karystianis, and E. Coiera, "Automated screening of research studies for systematic reviews using study characteristics," *Systematic Reviews*, vol. 7, 04 2018.
- [82] S. Ananiadou, B. Rea, N. Okazaki, R. Procter, and J. Thomas, "Supporting systematic reviews using text mining," *Social Science Computer Review*, vol. 27, pp. 509–523, 10 2009.
- [83] O. Frunza, D. Inkpen, and S. Matwin, "Building systematic reviews using automatic text classification techniques," vol. 2, 01 2010, pp. 303–311.