
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 automation 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](image-url)

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
we applied IC and EC criteria first on 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 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 document 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 Malheiro 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

Table I

<table>
<thead>
<tr>
<th>Category</th>
<th>Rationale</th>
<th>Addressed RQs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study meta-data</td>
<td>Identification and management of the study to detect the domain and publication data from the study.</td>
<td>RQ1, RQ2, RQ3</td>
</tr>
<tr>
<td>Search and selection automation approaches and tools in SE</td>
<td>Identification of automated approaches and tools that fully or partially support the activities of search and selection of studies for secondary studies in SE.</td>
<td>RQ1</td>
</tr>
<tr>
<td>Text classification approaches</td>
<td>Identification of approaches and metrics of text classification approaches for searching and selecting studies for secondary studies in SE and medicine.</td>
<td>RQ2, RQ3</td>
</tr>
</tbody>
</table>

1 https://bit.ly/3AoiSPw
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 uses 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.
Our results demonstrate that researchers recognize the need for 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’s 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], RelRank [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 into 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], [34], [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.
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 an 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


