POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

Development of a text-analytics based framework to support automated clinical literature research and study classification

TESI MAGISTRALE IN BIOMEDICAL ENGINEERING – TECHNOLOGIES FOR ELECTRONICS

Author: Giorgia Mancini
Advisor: Prof. Enrico Gianluca Caiani
Co-advisor: Prof.ssa Alessia Paglialonga
Academic Year: 2020-2021

1. Introduction

Nowadays, technological innovation is proceeding at a fast pace, especially in the biomedical field. The need to automatically retrieve and aggregate medical information from the Web is increasing, and for this reason, several platforms have been developed to guarantee the possibility of speeding up some essential processes for conducting complete and efficient clinical literature research. However, the automation of such processes still presents numerous challenges; in fact, many already developed tools still work independently and cannot be combined to create a solution that could include all the necessary steps.

Accordingly, the aim of this project was to provide a tool to be used to retrieve and analyze information from published studies throughout two of the most important databases used in literature research: PubMed and Google Scholar. Specifically, the developed solution included the automated classification, based on text analysis, of the retrieved records according to the type of study (Systematic Review and Meta-Analysis - SRMA, Randomized Clinical Trial - RCT, or Other) to possibly evaluate the level of clinical evidence for the queried topic.

2. Materials and Methods

Figure 1 schematizes the different steps involved in the process. All phases were automated and implemented using *Python.* Using the developed Web interface, the user can enter a query string that will be searched in the two chosen search engines, by which all the corresponding scientific articles will be downloaded in a local database. The implemented classification algorithm allowed categorization of the articles according to the type of study, by comparing the titles and

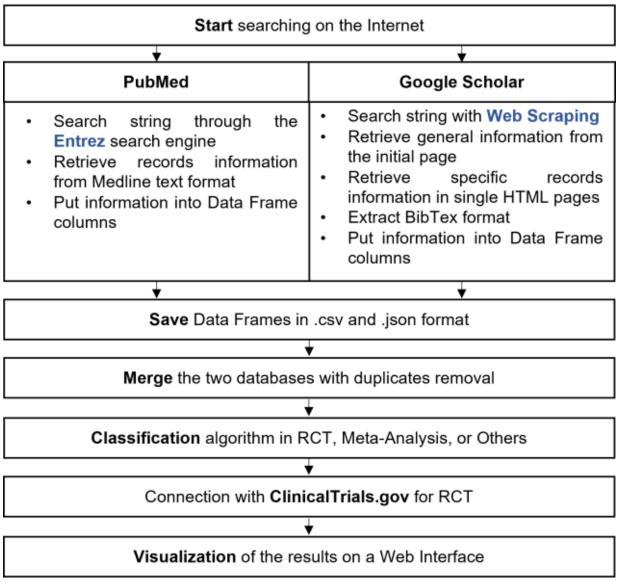

Start searching on the Internet
PubMed
Search string through the **Entrez** search engine
Retrieve records information from Medline text format
Put information into Data Frame columns
Google Scholar
Search string with **Web Scraping**
Retrieve general information from the initial page
Retrieve specific records information in single HTML pages
Extract BibTex format
Put information into Data Frame columns
Save Data Frames in .csv and .json format
Merge the two databases with duplicates removal
Classification algorithm in RCT, Meta-Analysis, or Others
Connection with **ClinicalTrials.gov** for RCT
Visualization of the results on a Web Interface

Figure 1 - General workflow of the thesis

abstracts with manually created dictionaries. At the end of this operation, data were presented to the user through the Web Interface in an intuitive and aggregated way.

2.1. Search in PubMed

The automatic research implemented for the PubMed website was performed thanks to a powerful search engine called Entrez and its Programming Utilities (E-Utilities), used to connect with the primary interface and retrieve information directly through programming. The ESearch Utility was used to find the results for a query string, inserted as an input parameter in the corresponding function, which returns the records PMIDs. Such PMIDs were used as input for the EFetch Utility function, used to obtain the Medline text format for each article (a simple text characterized by different fields representing the record information). Title, abstract, publication type, journal title, International Standard Serial Number (ISSN) code, publication date, Secondary Source, and Digital Object Identifier (DOI) were extracted from each record, according to specific criteria: the title and the abstract were used for classifying the article according to the type of study, the publication date was used to keep track of the final database update, the DOI was used as a comparator between records to delete eventual duplicates, and the Secondary Source is a code that characterizes officially registered RCTs, used to create a connection with the *ClinicalTrials.gov* website, thus providing the trial registration number on such website (NCT code).

2.2. Search in Google Scholar

The searching process implemented to access Google Scholar was performed through Web Scraping techniques, which are methods for extracting information from the Internet by sending HTTP (HyperText Transfer Protocol) requests to websites and receiving a response from which retrieve data. In this case, it was necessary to create a URL (Uniform Resource Locator) composed of different parts, corresponding to a specific searching parameter. The URL was set to start the searching from the first Google Scholar page, to sort the records from the most recent published (if no specific date ranges were chosen), to exclude patents, and to include only scientific articles.

Information like title, publication year, authors, and link, were retrieved from the Google Scholar starting page, which presents the list of the resulting articles, while other information like abstract and DOI were extracted from each single article page, by searching the corresponding tag of the HyperText Markup Language (HTML) page format.

More detailed data, such as publication date and journal information, were extracted after downloading the BibTex article format, a standardized format used for storing bibliographic data, which contains different information about the record.

2.3. Total Database

The PubMed and Google Scholar databases were created separately, from searching the information from scratch if the query string was not used before, or to update the database if it already existed. For this reason, the created files were named with the query string inserted by the user, plus the date of the database creation, extracted from a *logging* file (a text file that saved the date of each code execution). At the end of the searching process, PubMed and Google Scholar's results databases were merged to obtain a unique one, after removing duplicate records.

2.4. Classification

The classification process was implemented to label papers into RCT, SRMA or Others through the following phases.

Creation of dictionaries: A total of 200 SRMA and 200 RCTs were manually selected, classified, and then read to identify and understand the structure and the lexicon used. The most common terms and idioms were extrapolated to create a dictionary of words.

Grouping words with the Levenshtein Distance: During the classification process, titles and abstracts were compared with the dictionary words, but to avoid the repetition of very similar expressions during such comparison, the Levenshtein distance [1] was used to compute the similarity between words, to group them, and use a single statement for the comparison, thanks to the use of Regular Expressions (Regex), sequences of characters used to match combinations of strings.

Score computation and classification: For each paper, a score was computed according to the number of the dictionary's occurrences in the abstract and title (given by single words or by formed Regex), multiplying the result by 100. This score was compared to a threshold value set for SRMA (T_{SRMA}). If the score was above the threshold, the record was classified as SRMA and removed from the total dataset; otherwise, it was maintained. The process was then repeated for the papers to be still classified but using the RCT dictionary. Then, the computed score was compared with the threshold value for RCT (T_{RCT}) and classified as such if above the threshold. To find the best values for T_{SRMA} and T_{RCT} 600 records were manually classified (200 SRMA, RCTs, and 200 Other) and the algorithm previously described was run iteratively by comparing the records' scores with different values (*0.0, 15.0, 30.0, 45.0, 60.0, 75.0, 100.0)*, to first determine the best T_{SRMA} and then the best value for $T_{RCT.}$

After having set the best threshold values in the training step, two types of validation were performed: in the first one additional manually selected 100 RCT, 100 SRMA, and 100 OSs were used. In the second one, 200 RCTs whose label was predetermined by an online database [2], were studied together with 200 SRMA and 200 OSs.

3. Results

3.1. Databases Results

Some databases were created to prove the efficiency of the algorithm developed for the searching process. In any case, the missing values in the created databases were less than 10% of the total length, demonstrating the validity of the implemented Web Scraping techniques used for retrieving information.

3.2. Classification Results

The training phase, validation phase, and test phase were performed to reach to the final algorithm version with the correct parameters, and to compare the results with the classification made by PubMed staff through the PT tag (Publication Type tag).

During the training phase, the algorithm was iterated in a range of values to find the best T_{SRMA} and T_{RCT}. For each value, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) resulting from the classification of the 600 selected records were computed to construct the Confusion Matrices, both for SRMA and RCT, and the Sensitivity and False Positive Rate (FPR) were computed for constructing the ROC curves. For the classification of SRMA, the ROC curve reached a value of Area Under Curve (AUC) of 0.995 (Figure 3); 20 records over 200 did not have the right PubMed classification, but with $T_{SRMA} = 30$ they were all correctly classified, and the Sensitivity was equal to 1. For the RCTs identification, the threshold was set to 15 after iterating the algorithm over two different ranges of values (the first one was the same of SRMA classification, and the second one with a smaller step to perform a more precise analysis). The ROC curve, in this case, presented an AUC of 0.952 and the Sensitivity value reached with $T_{RCT} = 15$ was 0.97 (Figure 4). With this threshold, the 18 studies over 200 identified with a wrong PT tag were all correctly classified.

In the first validation all SRMA were correctly classified, with only 2 FP and no FN, thus achieving a total Accuracy of 0.993, and a Precision of 0.98. In comparison to the PubMed classification, 35 records over 100 with a wrong label were correctly identified by the proposed approach. Among the remaining 198 records, 99 were classified as RCTs (using $T_{RCT} = 15$),

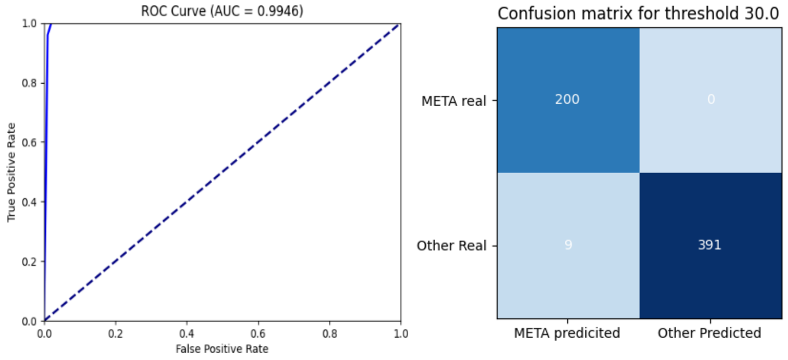


Figure 3 - ROC curve for SRMA (left) and Confusion Matrix for $T_{SRMA} = 30$ *(right)*

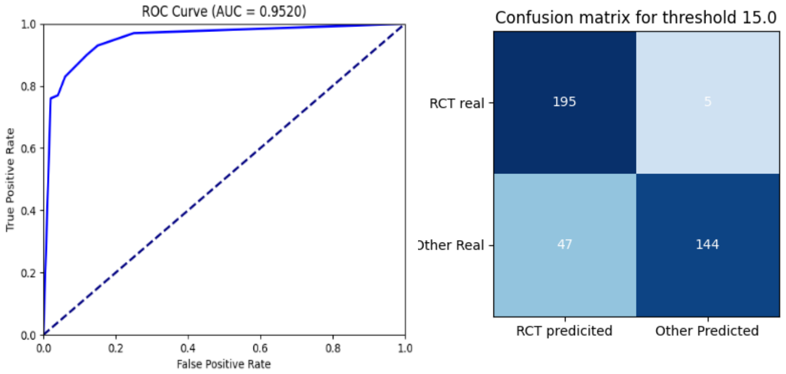


Figure 4 - ROC curve for RCTs (left) and Confusion Matrix for $T_{RCT} = 15$ *(right)*

with 6 FP and no FN, thus achieving a total Accuracy of 0.97, and a Precision of 0.943. 49 papers identified with a wrong PT tag were all correctly labeled by the implemented algorithm. In the second validation, the reached values of Accuracy and Precision were, respectively, 0.983 and 0.952.

In the test phase, 100 records of some of the created databases were randomly extracted and classified, manually screening afterwards their content to verify the correctness of the automated classification. In all the analyzed cases, the Accuracy was higher than 90%.

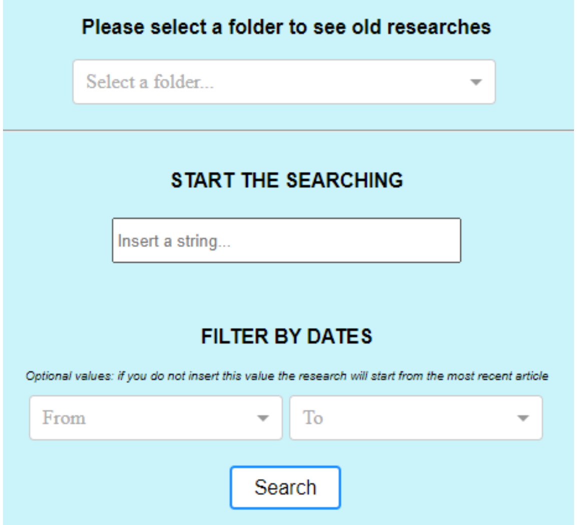
3.3. Web Interface Presentation

The Web Interface (implemented with the *dash* library in *Python* to integrate all the processes already described) was divided into two Tabs: the first one to implement the research on PubMed and Google Scholar, while the second one to display the retrieved information.

The first Tab (Figure 5) was organized as follows: at the top, a dropdown menu can be selected to see the already existing databases, relevant to previous queries. An input box is present to insert the new query. The section "Filter by dates" allows filtering the research in a specific range of publication years; if the user does not make any selection, the research will start backward from the most recent published article. By clicking the button "Search", the research is performed automatically both in PubMed and Google Scholar, and, once completed, the number of results obtained for both websites are shown. In the second Tab, first, only two elements are present: a dropdown menu with the list of all the existing databases and a button "Update list", to update such list if new files are added in the local folder. After clicking an item, the remaining part of the page displays general information about the selected database, information on journals, information on papers' publication years and a preview of the created database.

Years' information is presented through a timeline chart (Figure 6), and a range slider was inserted to filter the plot within the time desired period. At the end of the page, some records' information (title, abstract, links, publication date, journal, secondary source) were visualized in form of a table, with a checklist that the user can select to choose only particular types of studies (Figure 7).

For records classified as RCTs, the 'Secondary Source' column of the table can contain the NCT code or a set of uppercase letters, extracted from the record Title in case the NCT code was not retrieved, to potentially obtain the acronym of the corresponding RCT. By clicking on the not null cells of this column, the corresponding trial link in *ClinicalTrials.gov* will be opened. The visualized information can be downloaded locally after filtering, through the corresponding buttons.



Please select a folder to see old researches

START THE SEARCHING

FILTER BY DATES

Optional values: if you do not insert this value the research will start from the most recent article

Figure 5 - Organization of first Tab, with a zoom on the 'From' dropdown menu

4. Discussion and Conclusion

This project aimed to build the basis for a framework able to unify the necessary phases for conducting clinical literature research, from the automated retrieval of the information from websites to the organization of the results, up to the final visualization for users, including the classification of the reported studies into SRMA, RCT, or Other. PubMed and Google Scholar were chosen for retrieving information because PubMed is one of the most important online resources providing clinical data and a simple search engine, called Entrez, is provided to extract information from the Internet through code. Google Scholar was chosen because it has free access, and it is considered a valid search engine for academic research. The advantage of the implemented process for database creation is the possibility to have structured information inside a file, while other existent platforms allow only the download of entire text files for offline use or the analysis of the record directly on the interface [3].

The classification was implemented without using Machine Learning (ML) techniques, as opposite to other studies, but analyzing the titles and abstracts lexicon through strings comparison. The total workflow was designed to first classify the SRMA, removing them from the total dataset after classification, and then classify the RCTs, where the not identified records will be considered as Other. As regards the SRMA classification, setting $T_{SRMA} = 30$ gave the maximum number of TP and TN, while the choice of T_{RCT} was more challenging, with the value of 15 representing the best compromise between a large number of TP and minimizing the number of FN.

Anyway, both in validation and test phases, the Accuracy reached values higher than 90%, both for SRMA and RCT, and it was demonstrated that the implemented process achieved better results than the actual PubMed classification.

The choice of basing the classification process on strings comparison was made to speed up the categorization of papers still obtaining good results: methods based on ML techniques achieved very high results but using different evaluation metrics. Thomas et al. [4] reached a *recall* of 93.8%. Bulla et al [5] used Natural Language Processing (NLP) techniques obtaining an accuracy of the classifier of 76%. Besides these optimal results, existing studies concentrated only on the development of a classification process, focusing exclusively on the categorization of articles in RCT or non-RCT [6]. The classification process proposed in this work is also integrated into the final interface, to connect all the various phases of the project and present a complete workflow to the user.

The Web Interface was created to provide an intuitive way of using the entire framework. It is easy to use and well-organized and allows visualizing a preview of the newly created database, and to directly connect to one of the most important websites for retrieving information about RCTs, that is

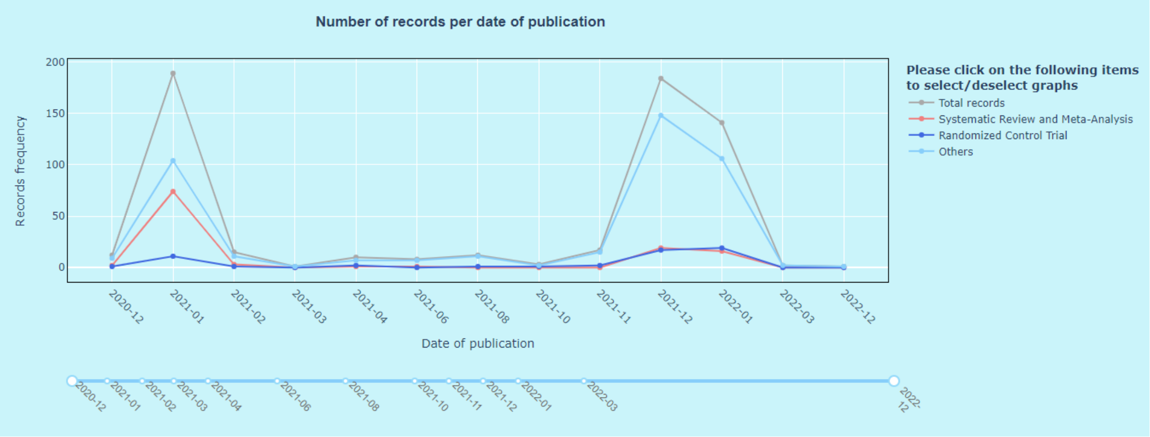
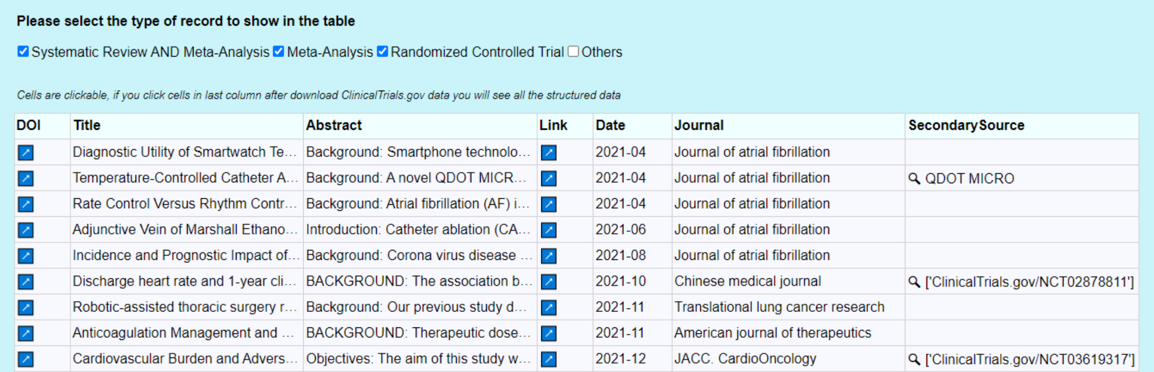


Figure 6 - Section 'Information on Years' of Tab2: information taken from database created with the string "atrial fibrillation"



Please select the type of record to show in the table

☑ Systematic Review AND Meta-Analysis ☑ Meta-Analysis ☑ Randomized Controlled Trial □ Others

Cells are clickable, if you click cells in last column after download ClinicalTrials.gov data you will see all the structured data

DOI	Title	Abstract	Link	Date	Journal	SecondarySource
↗	Diagnostic Utility of Smartwatch Te...	Background: Smartphone technolo...	↗	2021-04	Journal of atrial fibrillation	
↗	Temperature-Controlled Catheter A...	Background: A novel QDOT MICR...	↗	2021-04	Journal of atrial fibrillation	🔍 QDOT MICRO
↗	Rate Control Versus Rhythm Contr...	Background: Atrial fibrillation (AF) i...	↗	2021-04	Journal of atrial fibrillation	
↗	Adjunctive Vein of Marshall Ethano...	Introduction: Catheter ablation (CA...	↗	2021-06	Journal of atrial fibrillation	
↗	Incidence and Prognostic Impact of...	Background: Corona virus disease ...	↗	2021-08	Journal of atrial fibrillation	
↗	Discharge heart rate and 1-year cli...	BACKGROUND: The association b...	↗	2021-10	Chinese medical journal	🔍 ['ClinicalTrials.gov/NCT02878811']
↗	Robotic-assisted thoracic surgery r...	Background: Our previous study d...	↗	2021-11	Translational lung cancer research	
↗	Anticoagulation Management and ...	BACKGROUND: Therapeutic dose...	↗	2021-11	American journal of therapeutics	
↗	Cardiovascular Burden and Advers...	Objectives: The aim of this study w...	↗	2021-12	JACC. CardioOncology	🔍 ['ClinicalTrials.gov/NCT03619317']

Figure 7 - Section 'DataTable' of Tab2: information taken from database created with the string "atrial fibrillation"

ClinicalTrials.gov, which represents a plus compared to other developed platforms.

However, the online resources used to search and download articles from the Internet were only two, while in a normal process of clinical literature research several different archives are queried, thus possibly resulting in a non-completeness of the obtained results. Another limitation is the computational time required to extract all the information from the Internet, and the limited number of records considered in the various validations of the classification algorithm. In any case, the final implementation represents certainly a good starting point for the development of a more powerful platform that could be very useful for healthcare professionals and researchers, created in a modular way so that any improvement could be simply added to the framework or substituting an already existing process.

References

[1] https://en.wikipedia.org/wiki/Levenshtein_distance#:~:text=Informally%2C%20the%20Levenshtein%20distance%20between,considered%20this%20distance%20in%201965

[2] https://github.com/Franck-ernoncourt/pubmed-rct

[3] Soto, A. J., Przybyła, P., & Ananiadou, S. (2019). Thalia: Semantic search engine for biomedical abstracts. *Bioinformatics*, *35*(10), 1799–1801.

[4] Thomas, J., McDonald, S., Noel-Storr, A., Shemilt, I., Elliott, J., Mavergames, C., & Marshall, I. J. (2021). Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. *Journal of Clinical Epidemiology*, *133*, 140–151.

[5] Bulla, L., Gangemi, A., Golinelli, D., Mongiovì, M., Nuzzolese, A. G., Rucci, P., Sanmarchi, F., Catania, R., & Bologna, U. (n.d.). *Toward AI-assisted Systematic Literature Reviews for Evidence-Based Medicine. I.*

[6] Marshall, I. J., Noel-Storr, A., Kuiper, J., Thomas, J., & Wallace, B. C. (2018). Machine learning for identifying Randomized Controlled Trials: An evaluation and practitioner's guide. *Research Synthesis Methods*, *9*(4), 602–614.