Master of Science – Applied Analytics and Data Science

Web content mining and information extraction to review Covid-19 rapid diagnostic test evaluations

Dègninou Yehadji

Technological University Dublin
Ireland
B00108474[at]mytudublin.ie

Abstract

In the context of information deluge from research output related to COVID-19 pandemic, some organizations have setup databases of scientific article collections related to the topic. Yet, it is hard for academics and health professionals to summarize or keep up to date with the information. It is been demonstrated that advanced computational approaches such as web mining and natural language processing can be used to review scientific literature. This project is an effort to use a set of web mining, information extraction and text analytics strategies to extract and analyze abstracts from the WHO Global literature on coronavirus disease. The purpose was to generate a structured database from article abstracts on COVID-19 rapid diagnostic test evaluations, identify top journals in which such studies are published, identify, and visualize top keywords used in article titles and summaries, and summarize diagnostic test performances.

Web mining and information extraction was performed using Python with various libraries. The relevant tags identified on the webpage were scraped using beautifulsoup and regular expression operations were performed to extract test performances (sensitivities and specificities). The scraped article titles and abstracts were then preprocessed, analyzed, and visualized using the natural language toolkit, the rapid automatic keyword extraction (RAKE) algorithm, and the WordCloud library.

As result, a structured dataset was generated with 11 columns and 73 records. It is found that most of the papers were Preprints and the analysis of performances showed that sensitivities varied between 58% and 99%, and specificities varied between 35% and 99%. As keywords, "test", "rapid", and "antibody" appeared in the top 10s for both titles and abstracts.

This project confirmed the possibility to used advanced computing methods to mine and extract key information from scientific literature. However, machine retrieved information, particularly in-text metrics such as sensitivity and specificity need to be check manually to validate them.

Keywords: web mining, information extraction, literature review, Covid-19, rapid diagnostic test, diagnostic test evaluation, diagnostic test performance, sensitivity, specificity.

1. Introduction

Medical literature is the scientific literature encompassing journals articles and texts in books dedicated to the field of medicine. The state of the art, opinions, discoveries, and research output related to the diagnosis, prognosis and treatment of medical conditions are being documented since thousands of years in the medical literature. Scientific literature is experiencing a phenomenon known as information overload. It is estimated that the number of published scientific papers has increased by 8–9% each year over the past several decades. In the medical field, more than 1 million papers are published each year, representing about two papers per minute. The medical literature is archived in databases such as PubMed, is a free search engine accessing a database of references and abstracts on life sciences and biomedical topics. Other medical literature databases such as Google Scholar, Cochrane Library, Scopus, Science Direct, and Ovid are available, but PubMed is the most popular. The U.S. National Library of Medicine (NIH, NLM), the institutional curator of PubMed database contains more than 30 million citations and abstracts of biomedical literature (U.S. National Library of Medicine, 2020).

In the context of the COVID-19 pandemic, scientific research is focused on the disease and is productive more than ever. Millions of dollars are being invested in COVID research and new philanthropic investments are flowing (Pai, 2020). Teixeira da Silva et al. (2020) analyzed data from Clarivate Analytics' Web of Science, and Elsevier's Scopus and found about 23,634 unique documents published between January 1 and June 30, 2020. Keeping up to date with the medical information is an arduous task for medical academics and professional, particularly in such a context of pandemic with fast and high-volume research. To track research output related to the pandemic, several organizations have setup

COVID-19 collection databases. For example, the 2019 Novel Coronavirus Research Compendium (NCRC), the Cochrane Library COVID-19 Resources, CORD-19: COVID-19 Open Research Dataset, EBSCO COVID-19 Portal, iSearch COVID-19 Portfolio, the LitCovid, and WHO Global literature on coronavirus disease are databases that collect COVID-19 related scientific publications (Hardi, 2020).

In this context of information overload, artificial intelligence (AI) is seen as a tool that can help harness the deluge of COVID-19 research output. For example, the White House, challenged data scientists to develop tools to analyse the pandemic data set to help researchers answer 10 high-priority research questions identified by the U.S. National Academy of Sciences and the World Health Organization. One of the results of this data mining challenge is an AI-powered literature review. Using algorithms, researchers collected data points of interest from a subset of papers in the collection grouped in categories and created a web page for each topic that displays the results (Brainard, 2020). Such AI-powered literature reviews are based on combination of technics from web mining, natural language processing (NLP) and information extraction (IE). Web mining is the process designated to automatically discover and extract useful information from Web documents. Information extraction (IE) is the process of extracting useful structured information from the unstructured data such as text, image, audio, and video. NLP is a field of AI intended to enabling human–computer interaction using natural languages. Human talking to machine, machine capturing audio, audio to text conversion, data to audio conversion, machine responding to the human by playing audio file, and processing text data are some examples of humans-machine interactions using NLP (Kumar and Gosul, 2011; Gudivada and Arbabifard, 2018; Adnan and Akbar, 2019).

As part of the effort to utilized advanced computational tools to extract useful information from scientific literature, this project is intended to combine web content mining and information extraction approaches to review COVID-19 rapid diagnostic test evaluation studies. Specifically, the objective is to: i) generate a structured database from retrieved article abstracts, ii) identify top journals in which such studies are published; iii) identify and visualize top keywords used in article titles and summaries; and iv) summarize diagnostic test performances (sensitivity and specificity).

2. Methods

2.1. Data and overall approach

In the context of high productivity of COVID-19 related research and information deluge, some organizations have setup literature databases to monitor scientific publications related to the pandemic. One of these databases is the WHO Global literature on coronavirus disease. It is a collection of international multilingual scientific findings and knowledge on COVID-19. The database is updated daily

(Monday through Friday) from searches of bibliographic databases, hand searching, and the addition of other expert-referred scientific articles (Hardi, 2020). To conduct this project, a search was performed on the WHO Global literature on coronavirus disease database to extract and mine relevant information from abstracts of publications related to the COVID-19 rapid diagnostic tests.

The web mining and information extraction were performed with Python (Rossum, Drake and Van Rossum, 2010) in a Jupyter notebook from the Anaconda software distribution with a set of various libraries (Anaconda, Inc., 2012). The Requests: HTTP for Humans™ library was used to send http requests allowing query of strings to the URL (Reitz, 2011), the Beautifulsoup4 library to scrap the webpage content (Richardson, 2004), and the re module for regular expression operations from the Python standard library to extract test performances (sensitivities and specificities) from abstracts (Python Software Foundation, 2017).

2.2. Search strategy

The search for articles related to COVID-19 rapid diagnostic test performance evaluation was performed on WHO Global literature on coronavirus disease [https://search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov]. Because the search was performed on a collection of articles related to the pandemic only, there was no need to include COVID-19 keywords in the search equation.

The keywords used to identify rapid diagnostic tests were "rapid test" or "rapid diagnostic test", and "sensitivity" and "specificity" were added to retrieve studies reporting diagnostic performances. The final equation input in the search bar was (("rapid test" OR "rapid diagnostic test") AND sensitivity AND specificity) which generated the following specific URL: *https://search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/?output=site&lang=en&from=0&sort=&format=summary&count=100&fb=&page=1&skfp=&index=tw&q=%28%22rapid+test%22+OR+%22rapid+diagnostic+test%22%29+AND+sensitivity+AND+specificity*

2.3. Web mining

The webpage generated by the search is a series of article abstracts as presented in Figure 1. Python was used with a combination of modules to scrap items over the page. After getting the webpage with `request`, the content was scraped with *beautifulsoup*. All the items of interest in each article summary were scraped depending on their nodes and elements (Table 1).

Table 1. List of items with their nodes and elements scraped on the webpage

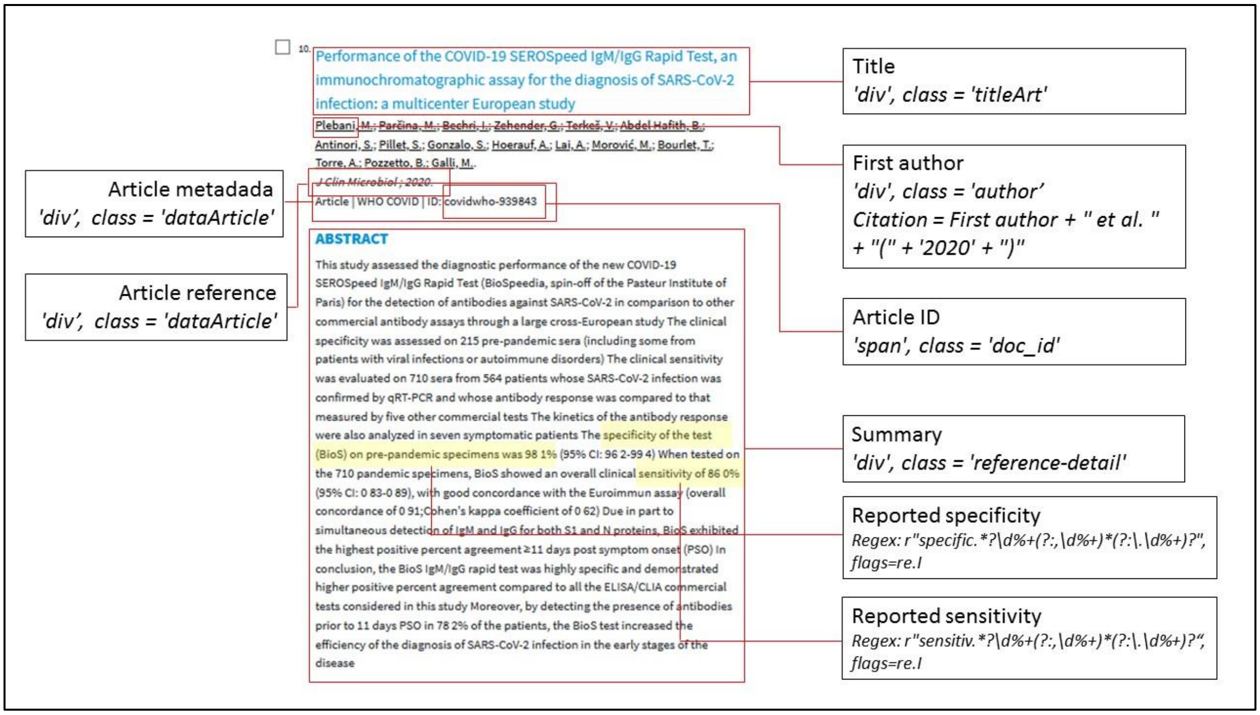
Item	**Node and element**
Article ID	'span', class = 'doc_id'
Article metadata	'div', class = 'dataArticle'
Article container	'div', class = 'textArt'
Reference	'div', class = 'reference'
Authors list	'div', class = 'author'
Article title	'div', class = 'titleArt'
Article summary	'div', class = 'reference-detail'

2.4. Information extraction

An information extract method was applied to extract reported of diagnostic test performances, specifically, sensitivity and specificity. Regular expression was used to extract all sentences containing "sensitiv" and "specific" followed by digits in % format. The regular expression equations were set as follows: `(r"sensitiv.*?\d%+(?:,\d%+)*(?:\.\d%+)?", flags=re.I)` for sensitivity and `(r"specific.*?\d%+(?:,\d%+)*(?:\.\d%+)?", flags=re.I)` for specificity. These equations scraped all reported sensitivities and specificities in % format as lists. The first reported performances were kept without the % sign and all other figures were ignored. Finally, these first performances were filtered and only those between 25 and 100 were kept as those outside this range seemed unrealistic.

2.5. Structured dataset

All items scraped from the URL using web mining and information extraction were passed to list placeholders. These lists were then used to create a dictionary that was then converted into a data frame using Pandas library (McKinney, 2008). The Pandas data frame generated was saved in a Microsoft Excel (Microsoft Corporation, 2016) sheet for further analysis.



□ 10. Performance of the COVID-19 SEROSpeed IgM/IgG Rapid Test, an immunochromatographic assay for the diagnosis of SARS-CoV-2 infection: a multicenter European study

Plebani, M.; Parčina, M.; Bechri, I.; Zehender, G.; Terkeš, V.; Abdel Hafith, B.; Antinori, S.; Pillet, S.; Gonzalo, S.; Hoerauf, A.; Lai, A.; Morović, M.; Bourlet, T.; Torre, A.; Pozzetto, B.; Galli, M..

J Clin Microbiol ; 2020.

Article | WHO COVID | ID: covidwho-939843

ABSTRACT

This study assessed the diagnostic performance of the new COVID-19 SEROSpeed IgM/IgG Rapid Test (BioSpeedia, spin-off of the Pasteur Institute of Paris) for the detection of antibodies against SARS-CoV-2 in comparison to other commercial antibody assays through a large cross-European study The clinical specificity was assessed on 215 pre-pandemic sera (including some from patients with viral infections or autoimmune disorders) The clinical sensitivity was evaluated on 710 sera from 564 patients whose SARS-CoV-2 infection was confirmed by qRT-PCR and whose antibody response was compared to that measured by five other commercial tests The kinetics of the antibody response were also analyzed in seven symptomatic patients The specificity of the test (BioS) on pre-pandemic specimens was 98 1% (95% CI: 96 2-99 4) When tested on the 710 pandemic specimens, BioS showed an overall clinical sensitivity of 86 0% (95% CI: 0 83-0 89), with good concordance with the Euroimmun assay (overall concordance of 0 91;Cohen's kappa coefficient of 0 62) Due in part to simultaneous detection of IgM and IgG for both S1 and N proteins, BioS exhibited the highest positive percent agreement ≥11 days post symptom onset (PSO) In conclusion, the BioS IgM/IgG rapid test was highly specific and demonstrated higher positive percent agreement compared to all the ELISA/CLIA commercial tests considered in this study Moreover, by detecting the presence of antibodies prior to 11 days PSO in 78 2% of the patients, the BioS test increased the efficiency of the diagnosis of SARS-CoV-2 infection in the early stages of the disease

Article metadada
'div', class = 'dataArticle'

Article reference
'div', class = 'dataArticle'

Title
'div', class = 'titleArt'

First author
'div', class = 'author'
Citation = First author + " et al. " + "(" + '2020' + ")"

Article ID
'span', class = 'doc_id'

Summary
'div', class = 'reference-detail'

Reported specificity
Regex: r"specific.?\d%+(?:,\d%+)*(?:\.\d%+)?", flags=re.I*

Reported sensitivity
Regex: r"sensitiv.?\d%+(?:,\d%+)*(?:\.\d%+)?“, flags=re.I*

Figure 1. Anatomy of an abstract with scraped items and their tags

2.6. Data analysis

The structured dataset saved in Excel was loaded in a pandas data frame for further analysis in Python (Rossum, Drake and Van Rossum, 2010). Text preprocessing and analysis was performed on the titles and abstracts with the natural language toolkit (NLTK) before determining term frequencies (Bird, 2011). The steps presented in Figure 2 were followed for text preprocessing. The *count()* function was applied on the preprocessed tokens to determine term frequencies in titles and abstracts, and the preprocessed string was passed to the WordCloud module (Mueller, 2015) to draw word clouds, and the *value_counts()* function applied to the "Journal title" column to obtain frequencies of journal titles, which allowed identifying top journals in which such studies are published.

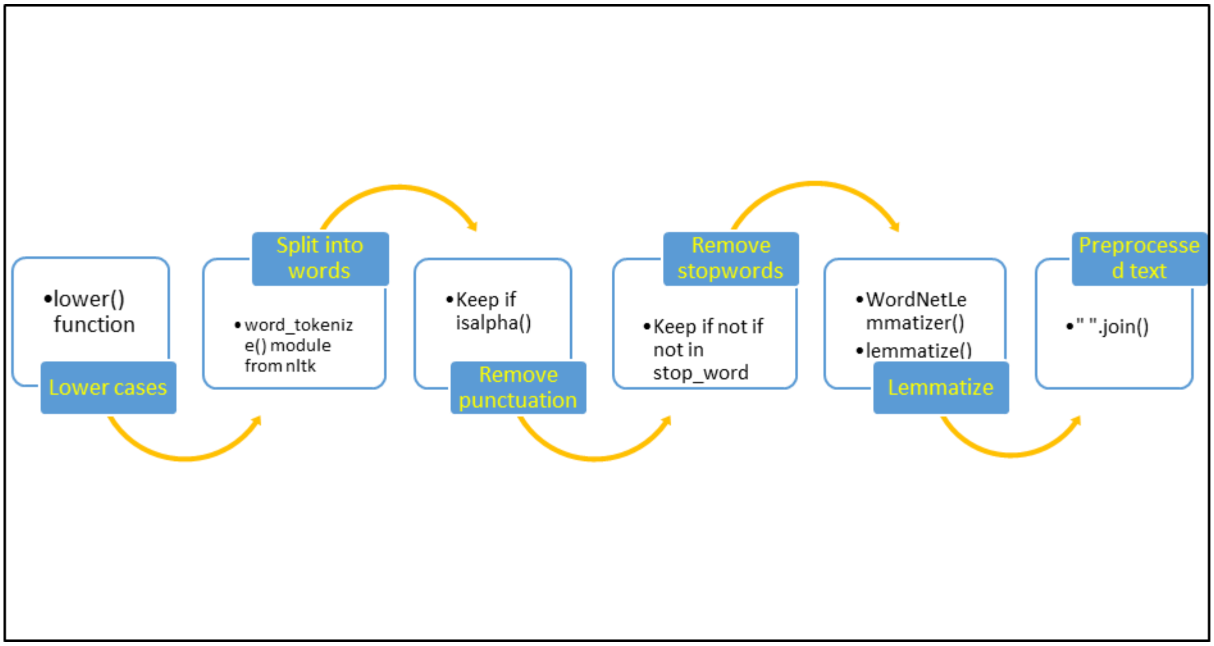


Figure 2. Text pre-processing steps performed on titles and abstracts

The top keywords were also determined for article titles and abstract using the Python module for implementing the rapid automatic keyword extraction algorithm (RAKE) (Pinho and von Feilitzsch, 2015). This did not require preprocessing as Rake directly handle strings and filter stop words as provided. A Rake object was generated using the smart stop list edited by adding the Chinese characters found in one of the abstracts (Salton, 1971). The Rake object was then run over the title and abstract strings with lowered characters and the keyword tuples were sorted from largest to lowest score.

The summary of diagnostic test performances (sensitivity and specificity) was performed using the *describe()* function to generate descriptive statistics, and the Seaborn package was used to draw box plots to visualize the distribution (Waskom, 2013). A bar chart was also drawn in Microsoft Excel to visualize the performances per citation.

3. Results and discussion

The web content mining and information extraction approach adopted helped scraping relevant information for articles summaries related to COVID-19 rapid diagnostic tests performance evaluations available on the WHO Global literature on coronavirus disease database.

The structured dataset was produced with the following columns: ID, Metadata, Citation, Reference, Journal title, Title, Url, Raw sensitivity extract, Raw specificity extract, Sensitivity (Clean), Specificity

(Clean), Summary. A total of 73 articles were retried as of December 17, 2020 at 08:20 PM UTC[1], and consequently, the dataset contains 73 rows. For sensitivity and specificity only 13 and 17 records meeting the scraping criteria were found (Table 2).

Table 2. Description of the dataset generated

#	Column	Valid record	Data type
0	ID	73	object
1	Metadata	73	object
2	Citation	73	object
3	Reference	73	object
4	Journal title	73	object
5	Title	73	object
6	Url	73	object
7	Raw sensitivity extract	73	object
8	Raw specificity extract	73	object
9	Sensitivity (Clean)	13	float64
10	Specificity (Clean)	17	float64
11	Summary	73	object

The analysis of publications showed that most of them are preprints (45%) and J Clin Virol (11%), Clin Microbiol Infect (4%), J Clin Microbiol (3%), and Am J Clin Pathol (3%) are the top journals which published such studies (Table 3).

[1] The website is subject to daily update and may return different result

Table 3. Top journals which published retrieved studies

	Frequency	
Journal	**n**	**%**
Preprint	33	45
J Clin Virol	8	11
Clin Microbiol Infect	3	4
J Clin Microbiol	2	3
Am J Clin Pathol	2	3
Indian J Med Res	1	1
Journal of Medical Virology	1	1
Int J Infect Dis	1	1
Complex & Intelligent Systems	1	1
Clin Transl Med	1	1
Other journals	20	27
Total	**73**	**100**

The analysis of performances showed that sensitivities varied between 58% and 99%, and specificities varied between 35% and 99%. The highest performances were found in Valter et al. (2020), 张稳健 et al. (2020), Kathrine et al. (2020), Ong et al. (2020), Percevent et al. (2020), Choe et al. (2020), Herroelen et al. (2020), and Pauline et al. (2020) abstracts (Figure 3).

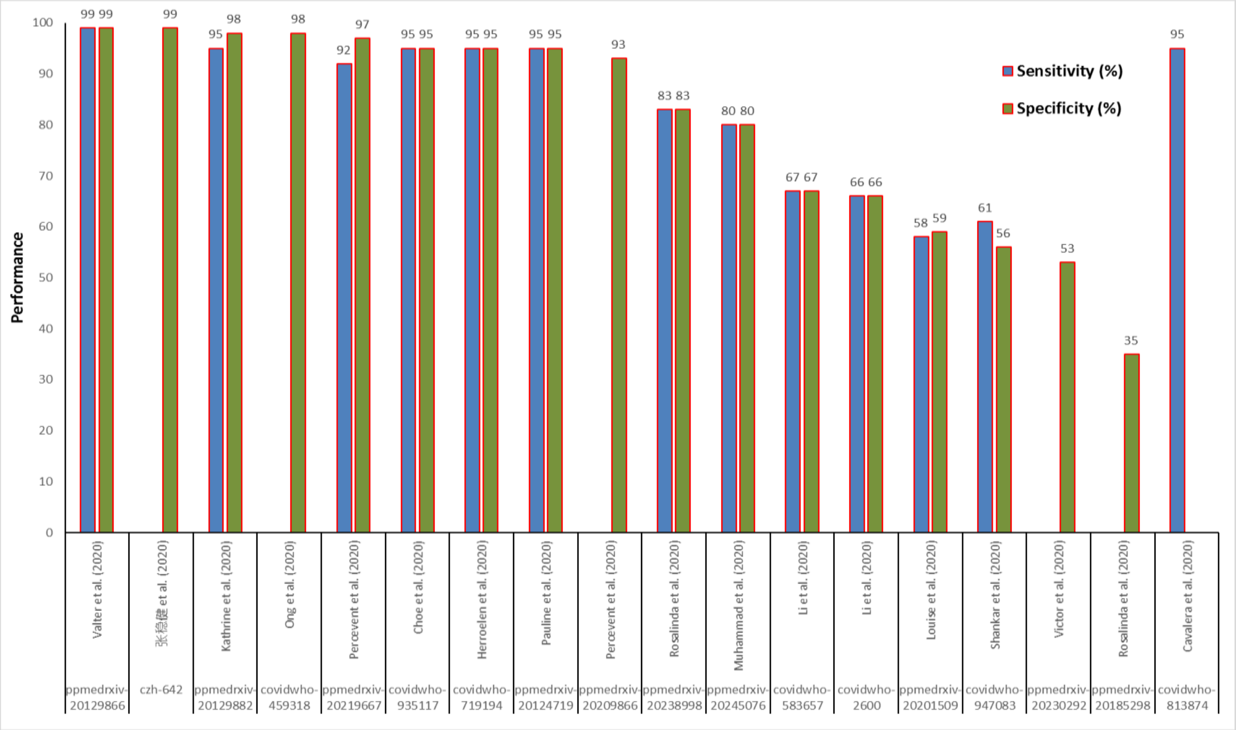


Figure 3. Bar chart of performances per abstract

Also, the average sensitivity was 83% (Sd = 15) for sensitivity and 80% (Sd = 20) for specificity, with medians of 92% and 93% respectively (Figure 4).

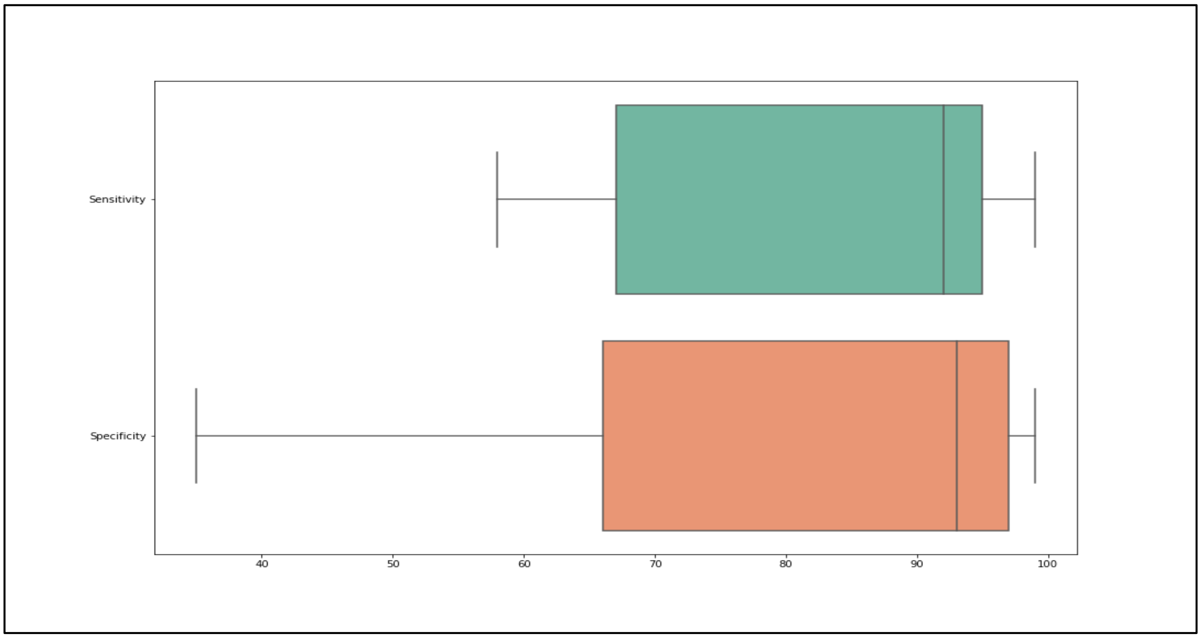


Figure 4. Box plots of diagnostic test performances

The term frequency analysis showed that "test", "rapid", "evaluation", "antibody", and "infection" are the top 5 terms in article titles and "test", "rapid", "sensitivity", "specificity", and "patient" are the top 5s in abstracts. Also, "test", "rapid", and "antibody" are present in the top 10s for both titles and abstracts (Table 4).

Table 4. Term frequencies in titles and abstracts with punctuations and stop words removed

Title			Abstract		
Word	Frequency		Word	Frequency	
	n	%		n	%
Top ten			**Top ten**		
test	58	8	test	356	3
rapid	53	8	rapid	191	2
evaluation	22	3	sensitivity	187	2
antibody	18	3	specificity	141	1
infection	17	2	patient	137	1
diagnosis	16	2	result	125	1
performance	14	2	assay	116	1
diagnostic	13	2	sample	107	1
antigen	13	2	positive	105	1
detection	12	2	antibody	103	1
Others			**Other**		
Other words	466	66	Other words	8680	85
Total	**702**	**100**	**Total**	**10248**	**100**

🟨 Term in top ten for both titles and abstracts

This result of term frequencies is also confirmed by word clouds where "test", "rapid", "evaluation", "antibody", and "infection" are more visible for titles and "test", "rapid", "sensitivity", "specificity", and "patient" are more visible for abstracts (Figure 5).

Table 5. Top phrases in titles and abstracts stop words removed

Title		Abstracts	
Phrase	Score	Phrase	Score
ultra-rapid real-time rt-pcr method	38	colloidal gold-labeled mouse-antihuman lgm/lgg antibody	48
multi-target lateral flow immunoassay enabling	32	well-established high-throughput bench-top solutions	39
antigen-detecting point-of-care diagnostic test	31	real-time reverse transcription-polymerase chain reaction	35
primary healthcare centers real-life validation	30	real-time reverse-transcriptase polymerase chain reaction	34
roche/sd biosensor rapid antigen test	30	combining ultra-rapid real-time rt-pcr	31

In terms of phrases, "ultra-rapid real-time rt-pcr method", "multi-target lateral flow immunoassay enabling", and "antigen-detecting point-of-care diagnostic test" are found to be the top 3 phrases in titles, and "colloidal gold-labeled mouse-antihuman lgm/lgg antibody", "well-established high-throughput bench-top solutions", and "real-time reverse transcription-polymerase chain reaction" are the top 3 ones in abstracts (Table 5).

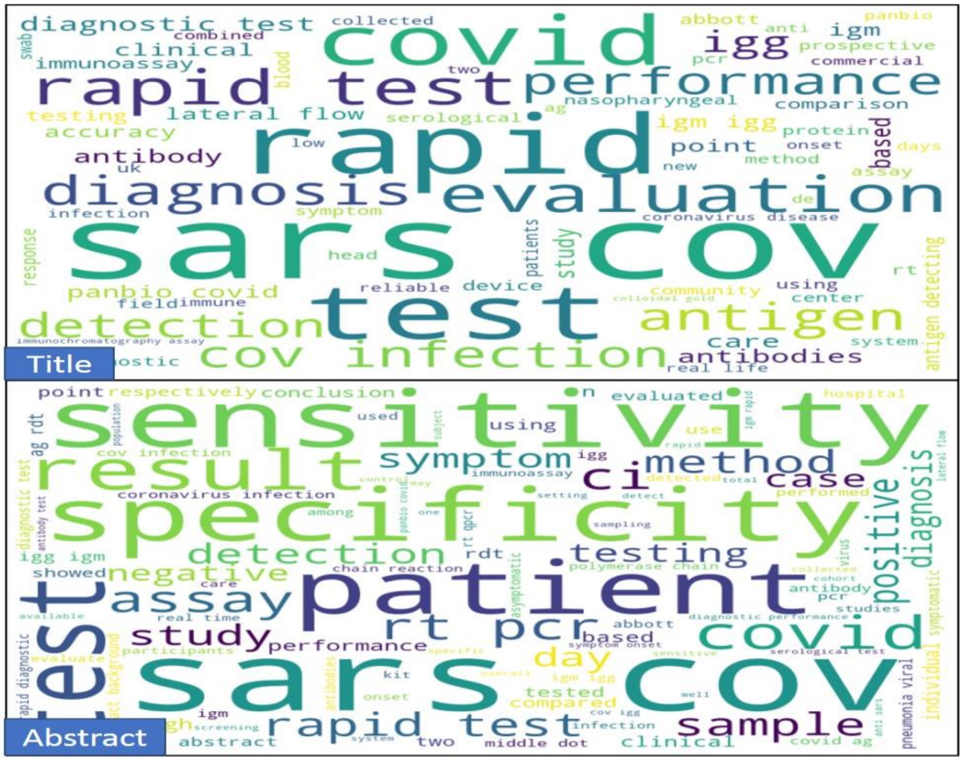


Figure 5. Word clouds of titles and abstracts with stop words removed

The methods adopted in this project was able to scrape relevant information from the WHO Global literature on coronavirus disease website and generate a structured dataset for analysis. More importantly, it was able to extract diagnostic test performances from the abstracts. It was then possible to review studies related to rapid diagnostic test performance evaluations by analyzing the journals in which such studies are published, by extracting the most common terms and key phrases in titles and abstracts, and by summarizing performance metrics (sensitivity and specificity).

Although the approach used in this project was able to help review COVID-19 rapid diagnostic tests, it is far from meeting the quality of a systematic review. The studies retrieved were not manually reviewed to exclude those not meeting the criteria. The purpose was to review the rapid diagnostic test performance evaluation, but some studies evaluating the performance of other diagnostic methods such as molecular biology (e.g. reverse transcription polymerase chain reaction), deep learning methods, and even other reviews or systematic reviews.

Only the first digits in % format following "sensitivity" and "specificity" were scraped, which consequently ignore other figures reported. So, studies reporting performances for more than one diagnostic test may not be fully reviewed. The regular expressions were set as if the performances were always reported after "sensitivity" and "specificity" terms, which may be different in some cases. Also, the regular expression was set as if the performances were all reported in English percentage format (e.g. 85.96%), which may miss or partially scrape figures reported in Latin languages (e.g. 85,96%) or other formats imposed by editors (e.g. 85*96%, 85˙96% or 85·96%).

Another limitation of this project is that the search was performed on a single website. Some relevant articles may not be available on the WHO Global literature on coronavirus disease. So, it would be recommended to search over other COVID-19 literature databases to ensure comprehensiveness of the search.

4. Conclusion

In this project, a set of web mining and information extraction techniques were used to scrape relevant information on the WHO Global literature on coronavirus disease website containing abstracts on research articles. A search was performed on the website to retrieve abstracts relevant to COVID-19 rapid diagnostic test evaluations. The Id, metadata, citation, reference, journal title, title, url, reported sensitivity and specificity, and summary were scraped and saved in a structured dataset. The dataset was then analyzed to summarize COVID-19 rapid diagnostic test performances, and to determine and visualize keywords and phrases in titles and summaries. Nevertheless, the project does not meet the quality criteria of a systematic review. One of the quality improvements would be performing a manual check of the diagnostic performances retrieved.

5. Supplementary information

Webpage in html: Covid rdt litterature webpage.thml

Python code in Jupyter notebook: Covid rdt summaries_Web mining_Code v3.3.ipynb

Dataset and descriptive statistics: Covid rdt litterature data.xlsx

6. References

Adnan, K. and Akbar, R. (2019) 'An analytical study of information extraction from unstructured and multidimensional big data', *Journal of Big Data*, 6(1), p. 91. doi: 10.1186/s40537-019-0254-8.

Anaconda, Inc. (2012) *Anaconda Software Distribution*. (Anaconda Documentation). Available at: https://docs.anaconda.com.

Bird, S. (2011) *NLTK: Natural Language Toolkit*. Available at: http://nltk.org (Accessed: 18 December 2020).

Brainard, J. (2020) 'Scientists are drowning in COVID-19 papers. Can new tools keep them afloat?', *Science*. Available at: https://www.sciencemag.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat (Accessed: 13 December 2020).

Gudivada, V. N. and Arbabifard, K. (2018) 'Chapter 3 - Open-Source Libraries, Application Frameworks, and Workflow Systems for NLP', in Gudivada, V. N. and Rao, C. R. (eds) *Handbook of Statistics*. Elsevier (Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications), pp. 31–50. doi: 10.1016/bs.host.2018.07.007.

Hardi, A. (2020) 'COVID-19 Resources: Databases & COVID-19 Search Strategies'. Becker Medical Library. Available at: https://beckerguides.wustl.edu/covid19/databasesearches (Accessed: 13 December 2020).

Kumar, G. D. and Gosul, M. (2011) 'Web mining research and future directions', in *Advances in Network Security and Applications*. *International Conference on Network Security and Applications*, Springer, Berlin, Heidelberg, pp. 489–496. doi: 10.1007/978-3-642-22540-6_47.

McKinney, W. (2008) *Pandas: Powerful data structures for data analysis, time series, and statistics*. Available at: https://pandas.pydata.org (Accessed: 15 December 2020).

Microsoft Corporation (2016) *Microsoft Excel*. Redmond, Washington, U.S.

Mueller, A. (2015) *Wordcloud: A little word cloud generator*. Available at: https://github.com/amueller/word_cloud (Accessed: 18 December 2020).

Pai, M. (2020) '"Covidisation" of academic research: opportunities and risks', *Nature Research Microbiology Community*. Available at: http://naturemicrobiologycommunity.nature.com/users/20892-

madhukar-pai/posts/65638-covidisation-of-academic-research-opportunities-and-risks (Accessed: 13 December 2020).

Pinho, T. and von Feilitzsch, F. (2015) *Python-rake: A python module implementing the Rapid Automatic Keyword Extraction algorithm.* Available at: https://github.com/fabianvf/python-rake (Accessed: 17 December 2020).

Python Software Foundation (2017) *The Python Standard Library*. Available at: https://python.readthedocs.io/en/latest/library/index.html (Accessed: 18 December 2020).

Reitz, K. (2011) *Requests: Python HTTP for Humans*^{TM}. Available at: https://requests.readthedocs.io (Accessed: 15 December 2020).

Richardson, L. (2004) *Beautifulsoup4: Screen-scraping library*. Available at: http://www.crummy.com/software/BeautifulSoup/bs4/ (Accessed: 15 December 2020).

Rossum, G. van, Drake, F. L. and Van Rossum, G. (2010) *The Python language reference*. Release 3.0.1 [Repr.]. Hampton, NH: Python Software Foundation (Python documentation manual, Guido van Rossum; Fred L. Drake [ed.] ; Pt. 2).

Salton, G. (1971) *The SMART retrieval system: Experiments in automatic document processing*. USA: Prentice-Hall, Inc.

Teixeira da Silva, J. A., Tsigaris, P. and Erfanmanesh, M. (2020) ‘Publishing volumes in major databases related to Covid-19’, *Scientometrics*. doi: 10.1007/s11192-020-03675-3.

U.S. National Library of Medicine (2020) *PubMed Overview, PubMed*. Available at: https://pubmed.ncbi.nlm.nih.gov/about/ (Accessed: 12 December 2020).

Waskom, M. (2013) *Seaborn: seaborn: statistical data visualization*. Available at: https://seaborn.pydata.org (Accessed: 18 December 2020).