EXPLORATORY ANALYSIS OF SCHOLARLY PUBLICATIONS ON ARTIFICIAL INTELLIGENCE (AI) IN COLONOSCOPY USING LITSTUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Due to the enormous quantity of scholarly papers, it can be difficult to provide comprehensive answers to general enquiries regarding the research area, such as, who are the key authors, what are the key issues or insights about the research topic. On this basis, we use Litstudy, a python library to investigate and gain insights into the scholarly publications of artificial intelligence (AI) for colonoscopy. ”AI” and ”colonoscopy” keywords were used as search results for bibliographic data, extracted, and merged as CSV files. 5872 bibliographic data were downloaded, with 3865 from IEEE Xplore and 2007 from Springer. Scopus yielded 5083 citation papers, while 789 unavailable citations were discarded. The non-negative matrix factorization (NMF) model was used to generate topic clusters, with a threshold of 0.85. The results showed that ”Patients” had the highest number of occurrences in four topics: 2, 3, 7, and 10, which we further visualised using topic clouds. Surprisingly, despite using the ”Artificial Intelligence” keyword to query the IEEE, Springer, and Scopus databases, topic 5 with AI was the vaguest word in the topic clouds. Topic 10 words cluster on colon cancer rehabilitation in colonoscopy revealed a weak relationship among the topic clusters. The project enables the selection of specific scientific articles, the examination and visualisation of their scholarly contribution using natural language processing (NLP), bibliographic network analysis, and, most importantly, revealed insights into word clusters in AI for colonoscopy publications.

1 INTRODUCTION

Litstudy Heldens et al. (2022) [1], for the selection of scholarly scopus articles, visualisation with bibliographic network analysis that show connections between publications and their authors, graph statistics on document metadata and automated topic discovery using natural language processing.

2 METHODOLOGY

2.1 ABOUT THE DATASETS

“AI” AND “Colonoscopy”[2] keywords were used as search results for primary metadata which includes the title, author, publication date, DOI (Digital Object Identifier) and the secondary data generated were topic clusters from abstract and bibliographic cocitation networks from many references/citations extracted. 3865 papers from IEEE Xplore and 2007 papers from Springer making 5872 papers in CSV format. Out of which 5083 citation papers were found on Scopus and 789 unavailable citations were discarded.

[1] Litstudy is a Python module that makes it possible to respond to queries using simple scripts

[2] Colonoscopy is the standard method of removing polyps in Colon cancer patientsRex et al. (2015)

2.2 FOLLOWED METHODS

We used the litstudy libraryHeldens et al. (2022) in this project, installed through the Python Package Index (PyPi) Bommarito & Bommarito (2019). It is based on several well-known tools from the Python data science ecosystem, including Pandas McKinney et al. (2011); McKinney (2012) and NumPyOliphant et al. (2006). Furthermore, we used non-negative matrix factorization (NMF) [4], an unsupervised model, to detect the topics and their semantic relationship using a threshold of 0.85. The five key steps are as follows: Explore and download scientific document metadata from a variety of sources, including IEEE, Springer's multiple research databases. The bibliography data from several sources was merged into a comma-separated variable (CSV) format; Sort, choose, remove duplicates from, and annotate document collections by comparing them with the Scopus bibliography data; Calculate and display broad statistics about the documents' metadata (e.g., statistics per year, per author, per journal, etc.); Create, map, and evaluate different bibliographic networks that illustrate the connections between the use of AI for colonoscopy publications and their authors, and topic discovery automatically using natural language processing (NLP).

3 RESULTS AND DISCUSSION

According to Fig.1, patients was highly prevalent in four topics: 2, 3, 7, and 10, which we further validated using cocitation graphical network as seen in Fig.2. Surprisingly, despite using the same keyword to search the database, topic 5 with AI was the vaguest word in topic clouds.

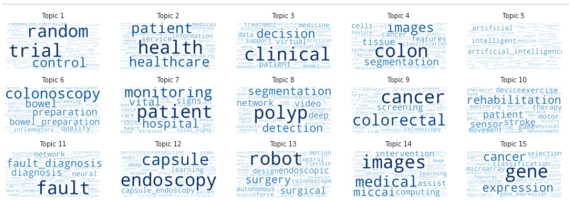

Figure 1: Top Word Clusters

Topic 10 in Fig.1 on cancer rehabilitation[3] revealed a weak relationship among the topic clusters, which requires additional investigation with the foreknowledge that the subject has a significant impact on colon cancer recovery.

In retrospective, co-citation analysis provides a forward-looking assessment of document similarity between AI and Colonoscopy. The number of citations a paper receives in the future is determined by the evolution of an academic field, so co-citation frequencies may change in the future.

4 CONCLUSION

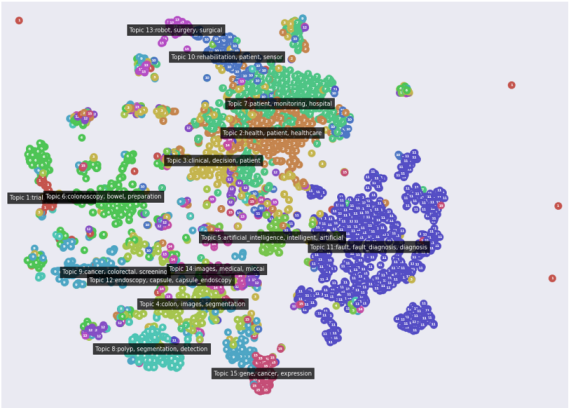

Figure 2: Top Words Histogram

In this paper, we used Litstudy, a Python tool, to explore past work on AI in colonoscopy. This enables us to examine the relevant information from scholarly articles using natural language processing, topic matrix model, and cocitation network analysis regarding past work on the use of artificial intelligence for colonoscopy. The study further clarified that, despite the clamour about artificial intelligence's exploration in cancer diagnosis and treatment, it has been underexplored in colonoscopy and colon cancer in general and some topics or subject need special attention than others.

[3]Rehabilitation is one of the procedures used to follow up on cancer recovery. van Weert et al. (2005)

ACKNOWLEDGEMENTS

Special thanks to Elsevier Customer support and UTAS Librarian for helping out with Scopus API error during coding phase.

URM STATEMENT

We acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

Ethan Bommarito and Michael Bommarito. An empirical analysis of the python package index (pypi). *arXiv preprint arXiv:1907.11073*, 2019.

Stijn Heldens, Alessio Sclocco, Henk Dreuning, Ben van Werkhoven, Pieter Hijma, Jason Maassen, and Rob V van Nieuwpoort. litstudy: A python package for literature reviews. *SoftwareX*, 20: 101207, 2022.

Wes McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.

Wes McKinney et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9, 2011.

Travis E Oliphant et al. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

Douglas K Rex, Philip S Schoenfeld, Jonathan Cohen, Irving M Pike, Douglas G Adler, M Brian Fennerty, John G Lieb, Walter G Park, Maged K Rizk, Mandeep S Sawhney, et al. Quality indicators for colonoscopy. *Gastrointestinal endoscopy*, 81(1):31–53, 2015.

Ellen van Weert, Josette Hoekstra-Weebers, Brigit Grol, Renée Otter, Hans J Arendzen, Klaas Postema, Robert Sanderman, and Cees van der Schans. A multidimensional cancer rehabilitation program for cancer survivors: effectiveness on health-related quality of life. *Journal of psychosomatic research*, 58(6):485–496, 2005.

A APPENDIX

A.1 CODES

The Notebook and the data would be added after the paper must have been accepted.

A.2 OTHER FIGURES FROM RESULTS

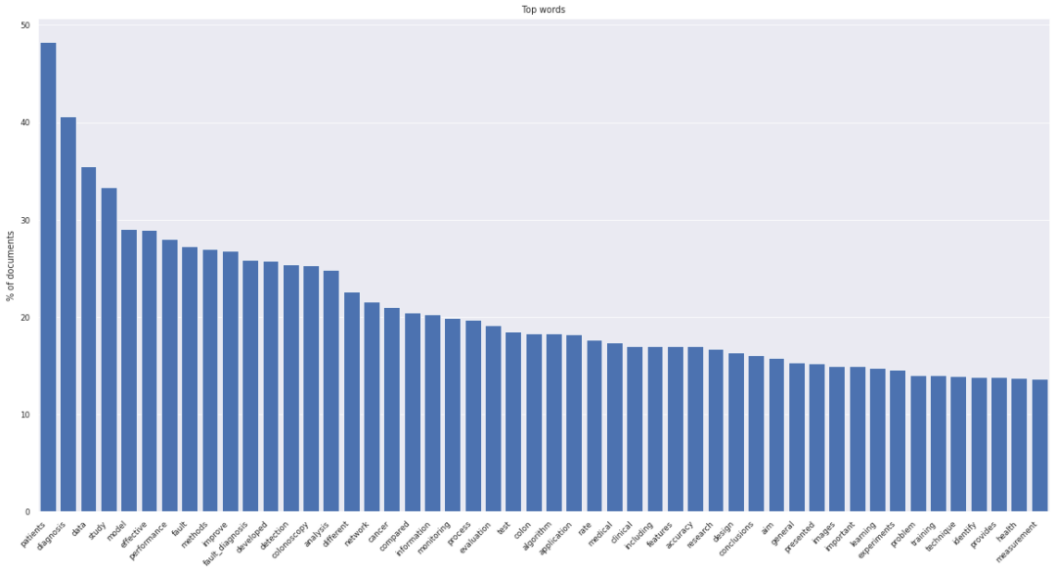

Figure 3: Top words Histogram

```
Topic 1: ['trial', 'random', 'control', 'randomized_controlled', 'adenoma']
Topic 2: ['health', 'patient', 'healthcare', 'service', 'information']
Topic 3: ['clinical', 'decision', 'patient', 'virtual', 'data']
Topic 4: ['colon', 'images', 'segmentation', 'tissue', 'features']
Topic 5: ['artificial_intelligence', 'intelligent', 'artificial', 'medicine', 'application']
Topic 6: ['colonoscopy', 'bowel', 'preparation', 'bowel_preparation', 'quality']
Topic 7: ['patient', 'monitoring', 'hospital', 'vital', 'signs']
Topic 8: ['polyp', 'segmentation', 'detection', 'video', 'network']
Topic 9: ['cancer', 'colorectal', 'screening', 'colonoscopy', 'risk']
Topic 10: ['rehabilitation', 'patient', 'sensor', 'stroke', 'exercise']
Topic 11: ['fault', 'fault_diagnosis', 'diagnosis', 'network', 'neural']
Topic 12: ['endoscopy', 'capsule', 'capsule_endoscopy', 'learning', 'deep']
Topic 13: ['robot', 'surgery', 'surgical', 'endoscopic', 'design']
Topic 14: ['images', 'medical', 'miccai', 'intervention', 'computing']
Topic 15: ['gene', 'cancer', 'expression', 'classification', 'microarray']
```

Figure 4: Topwords with NNMF model

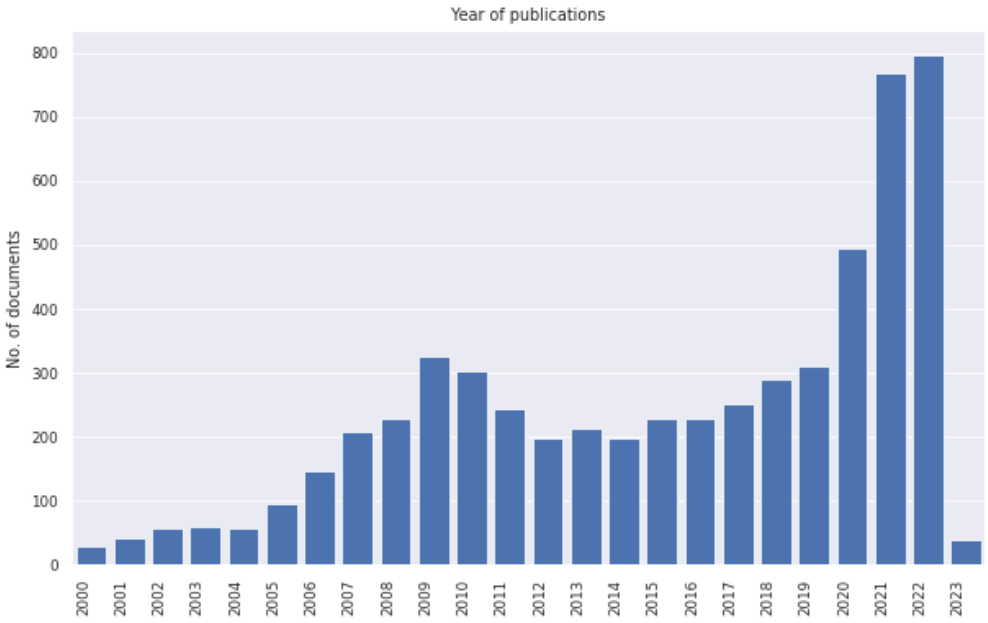

Figure 5: Year of Publication Histogram