Exploratory Analysis of Scholarly Publications on Artificial Intelligence (AI) in Colonoscopy using Litstudy

Mary Adewunmi & Kwang Chien Yee
CHM, School of Medicine,
University of Tasmania(UTAS),
Hobart, TAS7000, Australia.
{mary,adewunmi}{kwang,yee}@utas.edu.au

Dr Ming Chao Wong
School of ICT,
University of Tasmania(UTAS),
Sandybay, TAS7005, Australia.
{ming,wong}@utas.edu.au

Abstract

Due to the large number of scholarly papers on AI and colonoscopy and the short research period, it can be difficult to answer general questions about the research area, such as who the key authors are and what the key issues or insights are. We use Litstudy, a Python library, to study colonoscopy AI research. ”AI” and ”colonoscopy” keywords were used as search results. 3865 IEEE Xplore and 2007 Springer bibliographies were downloaded. Scopus found 5083 citation papers, excluding 789 unavailable citations. Topic clusters were created using the NMF model with a 0.85 threshold. Topic clouds showed that ”Patient” occurred most in four topics: 2, 3, 7, and 10. Despite querying IEEE, Springer, and Scopus databases with the ”Artificial Intelligence” keyword, subject 5 with AI has the lowest topic phrase weight in topic clouds. Topic 10 words cluster on colon cancer rehabilitation in colonoscopy showed weak topic clusters. The project selects scientific articles, analyses and visualises their scholarly contribution using natural language processing (NLP), bibliographic network analysis, and, most importantly, reveals word clusters in AI for colonoscopy publications.

1 Introduction

AI systems aid colonoscopyXu et al. (2023)Mori et al. (2021). 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.

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

[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)

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) an unsupervised model for topic detection and semantic relationship using a threshold of 0.85 [4] for the simplicity and easier interpretability of its results. The five key steps are as follows: Explored and downloaded 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; Sorted, chose, removed duplicates from, and annotate document collections by comparing them with the Scopus bibliography data; Calculated and displayed broad statistics about the documents' metadata (e.g., statistics per year, per author, per journal, etc.); Created, mapped, and evaluated 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 graph as seen in Fig.2. which was similarly used by Song et al. (2023) to study relationship among authors in scientometrics. Despite utilising the same keyword to search the database, subject 5 with AI has the lowest topic phrase weight.

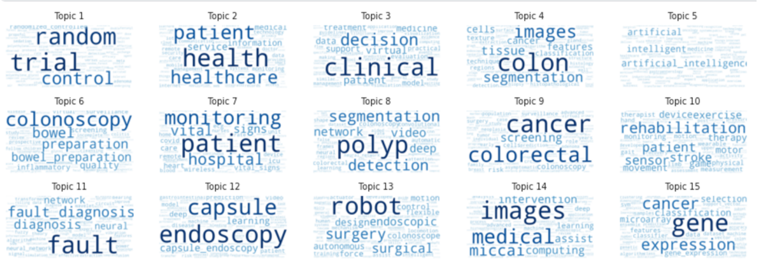

Figure 1: Top Word Clusters

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.

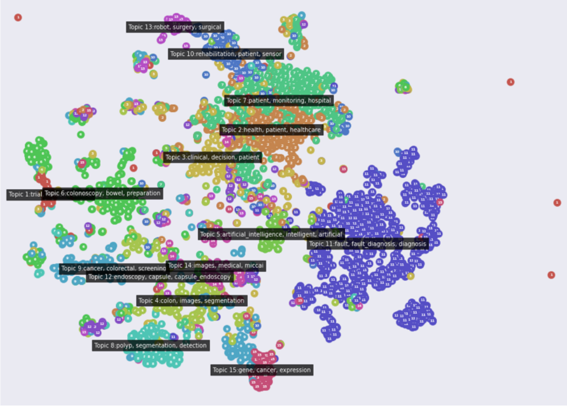

Figure 2: Cocitation Bibliography Network Graph

In retrospect, co-citation analysis evaluates AI and Colonoscopy document similarities. Future co-citation frequencies fluctuate with academic field evolution.

4 Conclusion

We examined AI in colonoscopy research using Litstudy, a Python application. This lets us apply natural language processing, topic matrix model, and cocitation network analysis to analyse research publications on colonoscopy and artificial intelligence. The study also revealed that artificial intelligence has been understudied in colonoscopy and colon cancer in general, and some topics or subjects require more 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, Paul Jordan Biyo, and UTAS librarian, Nathaniel Enright, for helping out with sorting Scopus API errors during the coding phase.

URM Statement

Author MA 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.

Yuichi Mori, Helmut Neumann, Masashi Misawa, Shin-ei Kudo, and Michael Bretthauer. Artificial intelligence in colonoscopy-now on the market. what's next? *Journal of Gastroenterology and Hepatology*, 36(1):7–11, 2021.

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.

Yanhui Song, Lixin Lei, Lijuan Wu, and Shiji Chen. Studying domain structure: a comparative analysis of bibliographic coupling analysis and co-citation analysis considering all authors. *Online Information Review*, 47(1):123–137, 2023.

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.

Hong Xu, Raymond SY Tang, Thomas YT Lam, Guijun Zhao, James YW Lau, Yunpeng Liu, Qi Wu, Long Rong, Weiran Xu, Xue Li, et al. Artificial intelligence–assisted colonoscopy for colorectal cancer screening: a multicenter randomized controlled trial. *Clinical Gastroenterology and Hepatology*, 21(2):337–346, 2023.

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

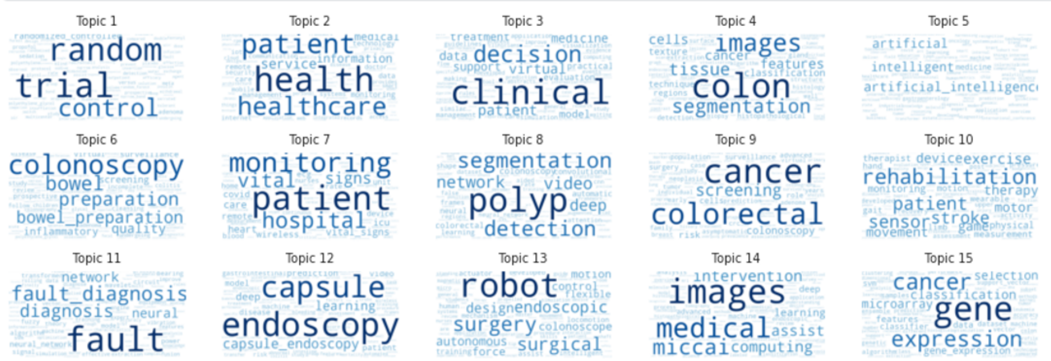

Figure 3: Top Word Clusters

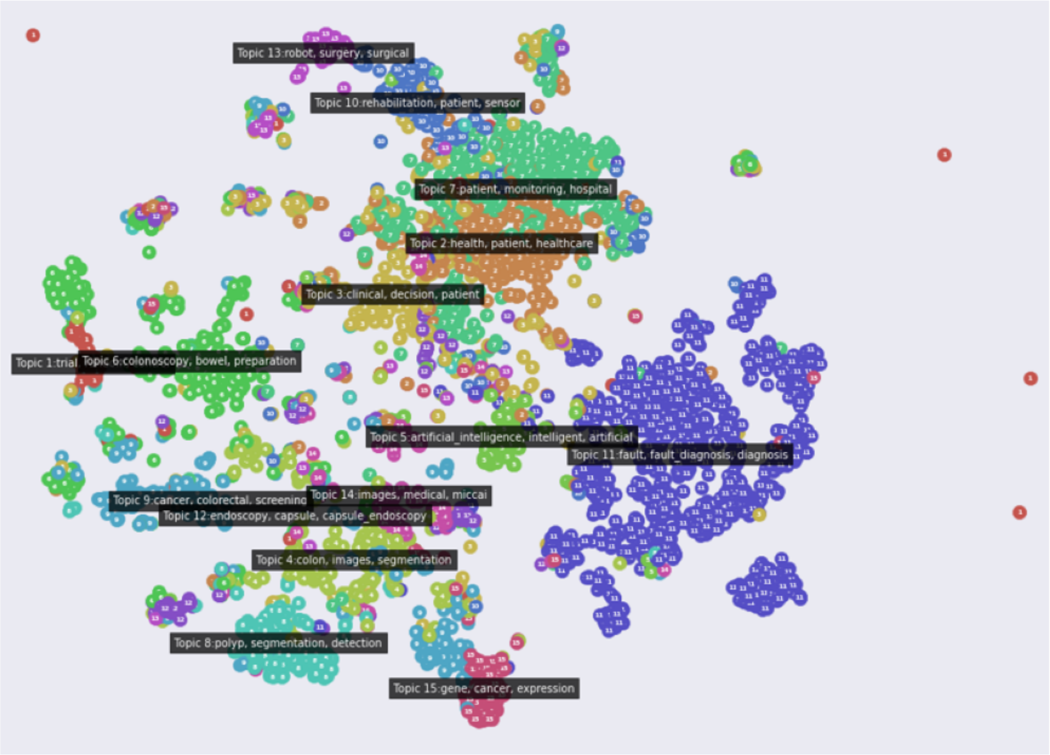

Figure 4: Cocitation Bibliography Network Graph

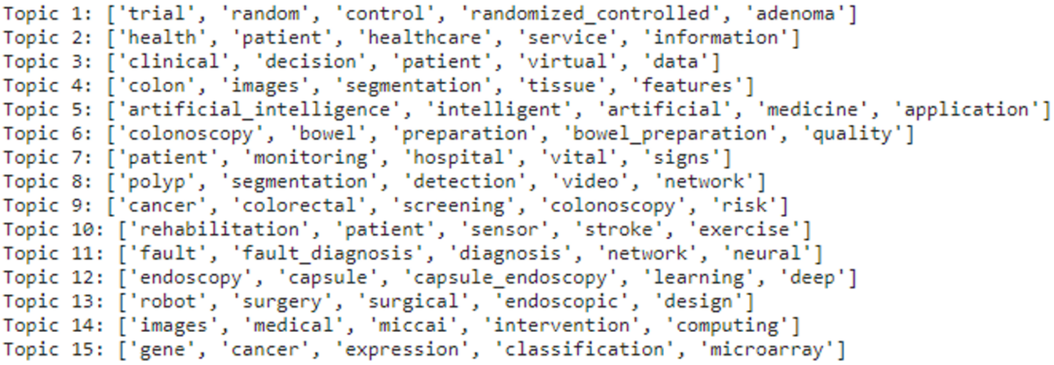
```
Topic 1: ['trial', 'random', 'control', 'randomized_controlled', 'adenoma']
Topic 2: ['health', 'patient', 'healthcare', 'service', 'information']
Topic 3: ['clinical', 'decision', 'patient', 'virtual', 'data']
Topic 4: ['colon', 'images', 'segmentation', 'tissue', 'features']
Topic 5: ['artificial_intelligence', '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 5: Topwords with NNMF model

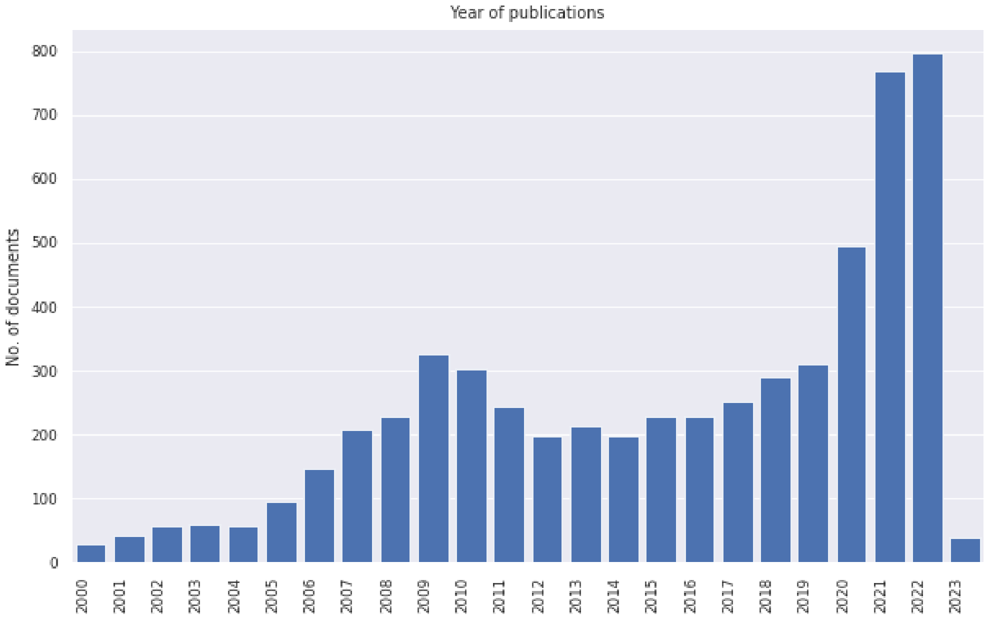

Figure 6: Year of Publication Histogram

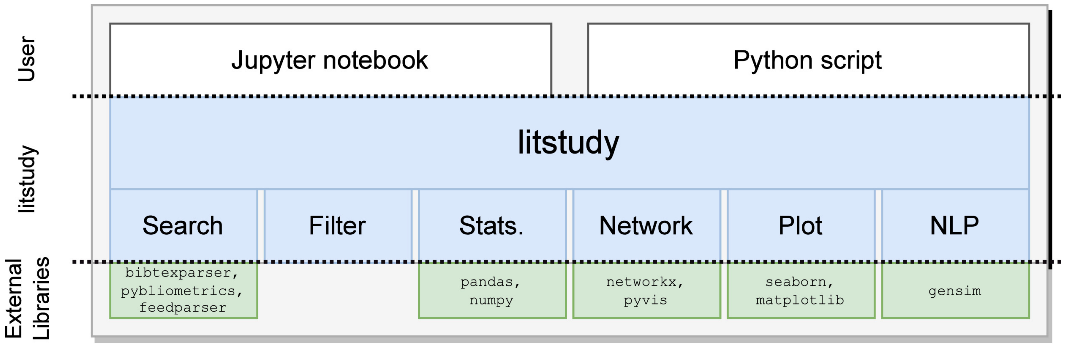

Figure 7: Litstudy conceptual frameworkHeldens et al. (2022)