Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240566

Unveiling Medical Insights: Advanced Topic Extraction from Scientific Articles

Ehsan BITARAF[a], Maryam JAFARPOUR[b,c,1], Sina SHOOL[a] and Reza SABOORI AMLESHI[a]
[a] *Rajaie Cardiovascular Medical and Research Center, Iran University of Medical Sciences, Tehran, Iran*
[b] *Department of Algorithms and Computation, School of Engineering Science, College of Engineering, University of Tehran, Tehran, Iran*
[c] *Center for Medical Data Science, Medical University of Vienna, Vienna, Austria*
ORCiD ID: Ehsan Bitaraf http://orcid.org/0000-0002-6588-7349, Maryam Jafarpour https://orcid.org/0000-0001-7266-5018, Sina Shool https://orcid.org/0000-0002-0280-3187, Reza Saboori Amleshi https://orcid.org/0000-0002-0299-5027

Abstract. In the ever-evolving landscape of medical research and healthcare, the abundance of scientific articles presents both a treasure trove of knowledge and a daunting challenge. Researchers, clinicians, and data scientists grapple with vast amounts of unstructured information, seeking to extract meaningful insights that can drive advancements in the biomedical domain including, research trends, patient care, drug discovery, and disease understanding. This paper utilizes the topic extraction algorithms on Breast Cancer Research to shed light on the current trends and the path to follow in this field. We utilized TextRank and Large Language Models (LLM) using the TripleA tool to extract topics in the field, analyzing and comparing the results.

Keywords. Topic Extraction, Natural Language Processing, Large Language Models, Text Rank, Breast Cancer, Bibliometric Analysis

1. Introduction

Natural Language Processing (NLP) is a subfield of AI that provides algorithms for text analysis. Large language models (LLMs) are AI models trained on vast text data using NLP methods. Combining NLP and LLMs enables effective analysis of unstructured text like scientific articles. This integrated approach identifies key medical concepts and extracts actionable insights. NLP and LLMs together transform raw text into valuable knowledge for healthcare and research [1–3].

This paper explores advanced topic extraction from scientific articles. It builds on our previous work developing a tool for accessing, storing, and analyzing papers using a graph/network approach [4]. The focus is on navigating the complexities of medical research, aiming to identify relationships between symptoms, medications, and diseases. The goal is to provide actionable insights to healthcare professionals.

[1] Corresponding Author: Maryam Jafarpour; E-mail: m.jafarpour@ut.ac.ir, maryam.jafarpour@meduniwien.ac.at.

Topic extraction is a popular NLP task that has received extensive research interest. A machine-learning technique that organizes and understands large text data collections. It automatically assigns topics or themes as "tags" or categories to individual texts [5]. Topic extraction is also known as topic analysis or topic detection. Its uses include document clustering, trend analysis, bibliometric analysis, and more.

There are several approaches for topic extraction, including probabilistic models, meta-heuristic algorithms, n-gram based methods, and unsupervised techniques. Some examples are LDA, MVO, TextRank, TopicRank, PositionRank, and Large Language Models (LLMs) [6–12]. This paper utilized TextRank and LLMs as topic extraction methods. A comparison of the results from these two methods is provided.

2. Methods

We have used TextRank and LLM to extract topics. We extracted paper abstracts using PubMed API as provided by the “Triple A”. The complete source code, results and documentations are available at GitHub[2]. To extract topics we have taken the following eight steps:

1. **Initial Configuration**: Initial configurations and machine setup actions, such as setting environment variables, were performed in this step.
2. **Article Abstract Extraction**: To retrieve relevant papers with minimum quality content, we used the search strategy keywords: ("Breast Cancer"[Title]) AND (Therapy[Title]).
3. **Article Metadata Preparation**: In this step, "Triple A" operators were used to process paper metadata and content at different states, including extracting keywords and MeSH terms from the metadata.
4. **Extract Topic Using TextRank**: Analyzed the title and abstract of the articles using TextRank Algorithm based on the top 10 keyword ranks per each paper, i.e. top 10 highest ranked words were selected as main topics.
5. **Extract Topic Using LLM**: The titles and abstracts were analyzed using an LLM based on the top 10 identified topics. The open-source Mistral model[3] [13], was used for topic identification, with a template specifying parameters like temperature, penalties, and token limits.
6. **Export Dataset**: Exported the cleaned data to be publicly available at figshare [14] for further research, making the process and analyses reusable.
7. **Co-occurrence Graph Construction**: The extracted topics and keywords were used to construct three co-occurrence graphs, which were exported as graphml files for further analysis.
8. **Visualization**: VOSviewer [15] was used to visualize the cleaned co-occurrence graphs. Some nodes were discarded as they forced bias towards obvious breast cancer-related keywords/terms, suppressing visualization of other nodes like "patient," "treatment," "breast cancer," etc.

[2] https://github.com/mjafarpour87/medical-insights/tree/main

[3] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

3. Results

This paper proposes a pipeline to extract relevant topics in the breast cancer domain. 9,691 paper abstracts were extracted. Seven steps were followed to perform topic extraction using different approaches, resulting in graphs of various topics related to symptoms, treatments, procedures, etc. Comparing the large keyword/term/topic co-occurrence graphs from unstructured biomedical literature is challenging. However, network metrics facilitated by networkx[16] enable analysis and comparison of these co-occurrence networks, as shown in Table 1.

Table 1. A comparison of the constructed topic/keyword co-occurrence networks

Topic extraction method	# Nodes	# Edges	Average Degree	Density	Average Clustering Coefficient	Degree Assortativity Coefficient	# Components
LLM	45,806	357,482	7.8	0.0003	0.905	-0.069	514
Keyword	15,555	337,659	21.7	0.0027	0.869	-0.151	69
TextRank	41,185	288,024	6.99	0.0003	0.89	-0.076	86

Table 1 describes the complete and raw co-occurrence graphs. However, only excerpts of the graphs with highest degree nodes were visualized in Figures 1 to 3, discarding biased obvious nodes due to the limitations of the processing platform.

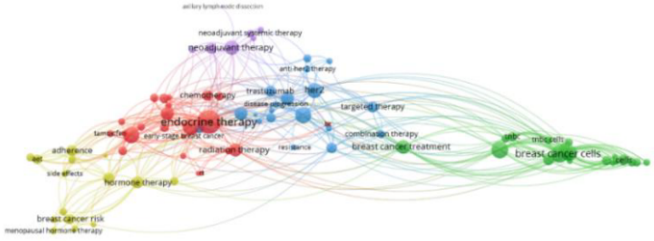

Figure 1. Topic Co-occurrence network using TextRank Algorithm

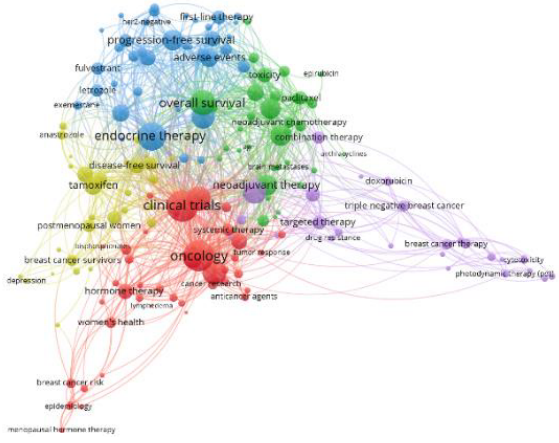

Figure 2. Topic Co-occurrence network using LLM

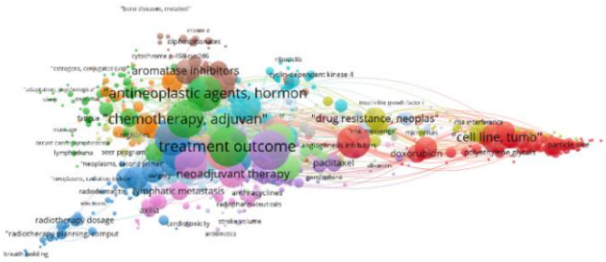

Figure 3. Keyword co-occurrence network

According to Table 1, a comparison of the performance of TextRank and LLM for topic extraction shows that LLM was capable of extracting more topics with more relations leading to a higher average degree and higher clustering coefficient which means that the nodes tend to have more relations and they tend to have existent neighbors, respectively. For the latter expression, a high clustering coefficient suggests that there is a community structure, where nodes within a community have strong ties. This is shown by the number of components, as well. Comparing these networks to the keyword co-occurrence network indicates that although the keywords network has a higher average degree, it has a lower clustering coefficient tending to have fewer components. This may be due to using more specific keywords by the authors as the metadata of the papers to make them accessible.

4. Discussion

Extracting meaningful information from scientific articles presents several challenges, especially for the advanced task of topic extraction. These challenges include [17]:

- Lack of domain-specific labeled data for accurate topic modeling.
- Key insights buried in lengthy background, experimental details and citations.
- Linguistic diversity for the same concepts requiring robust NLP techniques.
- Variations in writing styles and languages from diverse author backgrounds.
- Context-dependent scientific terms with multiple meanings requiring disambiguating based on context.

To overcome these issues as much as possible, providing a reproducible pipeline for information extraction, one can utilize several tools and packages. In this regard, we developed and upgraded TripleA [4] to analyze scientific papers. Being equipped with Graph Structure and processing capabilities, makes this tool a great potential for information extraction, as well. On the other hand, we selected TextRank and LLM primarily due to their robustness to noisy input, making them suitable for real-world scenarios. The graph-based nature of TextRank allows for transparent interpretation of results. These algorithms do not require domain-specific training, making them widely applicable and they handle large documents efficiently without sacrificing quality.

5. Conclusions

In this paper, we developed a pipeline based on Triple A, to extract topics from scientific papers. We tested this pipeline in the Breast Cancer field. We extracted 9,691 papers and

analyzed the topics using TextRank and LLM. An analysis of the co-occurrence networks of topics extracted by TextRank, LLM, and articles' keywords, showed that LLM extracted more topics that tend to construct concept clusters and TextRank stays in the second place in comparison to that. An analysis of the degree distribution of these networks will reveal a better understanding of how the topics are connected and how such a network behaves. It seems that if Clinical Trial articles are used, the extracted topics in treatments and factors affecting the disease will form a suitable knowledge graph that will be used for clinical decision support systems.

References

[1] Sezgin E, Hussain S-A, Rust S, Huang Y. Extracting Medical Information From Free-Text and Unstructured Patient-Generated Health Data Using Natural Language Processing Methods: Feasibility Study With Real-world Data. JMIR Form Res 2023;7:e43014. https://doi.org/10.2196/43014.

[2] Hahn U, Oleynik M. Medical Information Extraction in the Age of Deep Learning. Yearb Med Inform 2020;29:208–20. https://doi.org/10.1055/s-0040-1702001.

[3] Nasar Z, Jaffry SW, Malik MK. Information extraction from scientific articles: a survey. Scientometrics 2018;117:1931–90. https://doi.org/10.1007/s11192-018-2921-5.

[4] Jafarpour M, Bitaraf E, Moeini A, Nahvijou A. Triple A (AAA): a Tool to Analyze Scientific Literature Metadata with Complex Network Parameters. 2023 9th Int. Conf. Web Res. ICWR, 2023, p. 342–5. https://doi.org/10.1109/ICWR57742.2023.10139229.

[5] Diamantini C, Lo Giudice P, Potena D, Storti E, Ursino D. An Approach to Extracting Topic-guided Views from the Sources of a Data Lake. Inf Syst Front 2021;23:243–62. https://doi.org/10.1007/s10796-020-10010-x.

[6] Abasi AK, Khader AT, Al-Betar MA, Naim S, Alyasseri ZAA, Makhadmeh SN. An ensemble topic extraction approach based on optimization clusters using hybrid multi-verse optimizer for scientific publications. J Ambient Intell Humaniz Comput 2021;12:2765–801. https://doi.org/10.1007/s12652-020-02439-4.

[7] Parambath SAP. Topic Extraction and Bundling of Related Scientific Articles 2015. https://doi.org/10.48550/arXiv.1212.5423.

[8] Mihalcea R, Tarau P. TextRank: Bringing Order into Text. In: Lin D, Wu D, editors. Proc. 2004 Conf. Empir. Methods Nat. Lang. Process., Barcelona, Spain: Association for Computational Linguistics; 2004, p. 404–11.

[9] Zhu L, Huang M, Chen M, Wang W. An N-gram based approach to auto-extracting topics from research articles 2021. https://doi.org/10.48550/arXiv.2110.11879.

[10] Bougouin A, Boudin F. TopicRank: Topic ranking for automatic keyphrase extraction 2014;55:45–69.

[11] Florescu C, Caragea C. PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. Proc. 55th Annu. Meet. Assoc. Comput. Linguist. Vol. 1 Long Pap., Vancouver, Canada: Association for Computational Linguistics; 2017, p. 1105–15. https://doi.org/10.18653/v1/P17-1102.

[12] Dagdelen J, Dunn A, Lee S, Walker N, Rosen AS, Ceder G, et al. Structured information extraction from scientific text with large language models. Nat Commun 2024;15:1418. https://doi.org/10.1038/s41467-024-45563-x.

[13] Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas D de las, et al. Mistral 7B 2023. https://doi.org/10.48550/arXiv.2310.06825.

[14] Bitaraf E. Topic Extraction Dataset 2024:14235412 Bytes. https://doi.org/10.6084/M9.FIGSHARE.25533532.V1.

[15] van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010;84:523–38. https://doi.org/10.1007/s11192-009-0146-3.

[16] Hagberg A, Swart PJ, Schult DA. Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States); 2008.

[17] Hong Z, Ward L, Chard K, Blaiszik B, Foster I. Challenges and Advances in Information Extraction from Scientific Literature: a Review. JOM 2021;73:3383–400. https://doi.org/10.1007/s11837-021-04902-9.