Alryalat SA, Qasem A, Albdour K, Rawashdeh B. Assessment of Topics Published in Leading Medical Journals Using Natural Language Processing. *High Yield Medical Reviews*. Published online June 1, 2023. doi:10.59707/hymrHMDO2739

Online first

Assessment of Topics Published in Leading Medical Journals Using Natural Language Processing

Saif Aldeen Alryalat, MD[1], Ahmad Qasem[2], Karam Albdour[3], Badi Rawashdeh[4]

[1] Ophthalmology, University of Jordan, [2] Computer Science, University of Central Missouri, [3] School of Medicine, University of Jordan, [4] Division of Transplant Surgery, Medical College of Wisconsin

Keywords: Natural language processing (NLP), Topic modeling, Machine learning, Topics, Journals, Trends

https://doi.org/10.59707/hymrHMDO2739

High Yield Medical Reviews

Introduction

Topic detection can be used to identify trends in literature, providing valuable insight into the direction of the field. We developed a natural language processing (NLP) based method to identify topics from given abstracts and assessed the main topics of published articles by top medical journals in the last three years.

Methods

This study utilized a two-part methodology to extract and classify original articles published by four non-specialized medical journals; Lancet, New England Journal of Medicine, Journal of the American Medical Association, and British Medical Journal. The first part employed bibliometric data collection to search for original articles published between 2020 and 2022. The second part used an NLP approach based on the BERTopic model to classify the articles included into separate topics.

Results

The model was able to classify 1,540 articles out of the included 2,081 (79.42%) into 39 different topics in 11 fields. COVID-19-related and cancer treatment-related articles constituted approximately 25% and 7% of all published papers during 2020-2022 respectively. The study found that each of the included general medical journal tended to focus on certain topics more than others.

Conclusion

We identified a new methodology that can identify topics discussed in medical literature from abstracts as an input. We also demonstrated the potential of this methodology for analyzing trends in medical literature more efficiently and effectively. This study's methodology can be replicated on a larger scale with more papers, more journals, and over a longer period, highlighting the importance of further research using NLP models.

OVERVIEW

Medical literature has been rapidly expanding throughout the last few decades, where the number of yearly publications has an exponential growth since early 2000.[1] Processing and understanding what is being published and analyzing the topics that are considered "hot topics" is now more difficult to be done manually. However, there has been a parallel development in automatic detection techniques and natural language processing (NLP). NLP is a machine learning technique involving a set of methods and computer-aided algorithms designed to detect patterns in textual data.[2] With the exponential growth of medical data, NLP has the potential to establish itself as fundamental in every aspect of the healthcare industry due to the necessity of more efficient and accurate methods of analyzing and utilizing this information. NLP algorithms can process and understand vast amounts of unstructured medical text data, such as electronic health records, medical notes, and research papers.[3] It has the potential to be used in automating routine tasks, which reduces errors and streamlines the workflow, improving patient outcomes, lowering costs, and increasing the efficiency of the healthcare industry. Topic detection, in the context of NLP, can be defined as an algorithm that automatically identifies topics based on the content of a scientific article.[4] Topic detection algorithms analyze large amounts of text data and identify the underlying themes or topics present in the data. Topic detection can jumpstart the process of creating specific, topic-based databases by eliminating the time-consuming manual labor that has previously stifled its progress,[4] enabling physicians to remain informed about the latest developments in their respective fields. In addition, topic detection can also be used to identify trends in literature, providing valuable insight into the direction of the field, which may be useful to editorial boards in assembling a qualified pool of editorial members or reviewers by identifying topics that are receiving the most attention and investment in the field or guiding researchers to areas with gaps in literature where their efforts may be impactful. The purpose of this research article was to develop an NLP-based method for extracting topics from abstracts. Using such a novel model, we assessed the main topics of published articles by top medical journals in the last three years. Such an assessment will provide valuable insight to researchers and funders about current research trends and guide future work in these fields. We will also provide the details of this novel and robust methodology that can be used by other researchers from different fields and subfields of medicine.

METHODS

The current study composed of two main parts, a bibliometric method to extract original articles published by top journals, followed by a NLP method to classify topics discussed by extracted articles.

BIBLIOMETRIC DATA COLLECTION

We chose the top non-specialized medical journals in the 2021 Journal Citation Report®, with an impact factor (IF) above 50. These journals included Lancet (IF 202.731), New England Journal of Medicine (IF 176.082), Journal of the American Medical Association (IF 157.375), and British Medical Journal (IF 96.216). We did not include journals that publish reviews (i.e., Nature Reviews Disease Primers). All included journals were categorized under the “Medicine, General & Internal – SCIE” category.

We used Web of Science to extract the data on the 7th of January 2023, where we searched for all articles published in the aforementioned journals in the years 2020, 2021, or 2022. We restricted the search for original articles. We excluded non-original articles that were included even after applying the search query restriction (e.g., clinical cases published in NEJM). For each article, we included details about its journal, year, authors, affiliations, and funding body. In addition, we obtained the number of citations received according to the Web of Science database up to the date of search. We also included each article’s abstract, which will be further used as input to the NLP model to classify the articles. We manually excluded article types that did not have an abstract and did not report original data, including JAMA performance improvement, analysis in BMJ, updates on systematic review in BMJ, and special reports in NEJM without original data.

NATURAL LANGUAGE PROCESSING CLASSIFICATION

Transformers are a class of deep learning architectures that were introduced by Vaswani et al in 2017. They use self-attention mechanisms to attend to different parts of the input during training and inference which makes it particularly strong in NLP. BERT (Bidirectional Encoder Representations from Transformers) is one of the well-known transformer models for its benchmark performance in a variety of NLP tasks.[5]

To find the most prevalent themes in our dataset, we utilized BERTopic, a topic modeling approach based on the BERT language model. Modern topic modeling techniques like BERTopic use language models that have already been trained to capture the semantic meaning of words and sentences. Topic representations in BERTopic are generated in three steps, each document is first transformed using a trained language model into its embedding representation, second, the dimensionality of the generated embeddings is then decreased prior to clustering to improve the clustering procedure, and finally, the topic representations are retrieved from the document clusters using a customized class-based form of TF-IDF (term frequency-inverse document frequency).[5]

To cluster the documents in our dataset into different topics we used the default settings for the open-source BERTopic model, which includes using the Cosine Similarity measure to calculate the similarity between documents. The values for cosine similarity range between -1 and 1, with 1 denoting that the documents are identical, 0 denotes that they are orthogonal, and -1 for being diametrically opposed, the values correspond to the angle between the vectors. Consequently, BERTopic generates a similarity matrix to express the degree of similarity between every pair of texts in the dataset[6] (figure 2).

DATA PROCESSING

The model provided 39 different topics, with each having five keywords to describe it. Three physicians were consulted to provide a descriptive title for each topic, and the consensus description from the three physicians was adopted. Discrepancy between physicians mandated collaborative discussions to end up with an agreed upon description. After that, we categorized topics into main specialties.

STATISTICAL ANALYSIS

We used IBM SPSS Statistics for Windows, version 26.0 (IBM Corp., Armonk, N.Y., USA). in our analysis. We used mean (± standard deviation) to describe continuous variables. We used count (frequency) to describe other nominal variables. We compared the number of articles published by each journal on each topic using chi-square test. One-way ANOVA for different bibliometric metrics (i.e., citations, references, authors, and pages) between the four journals, and we reported the results using mean and standard deviation. We adopted a p-value of 0.05 as a significant threshold.

RESULTS

The initial list of articles from included journals comprised a total of 2,081 articles. 144 articles were excluded according to the described criteria, leaving a total of 1,937 in-

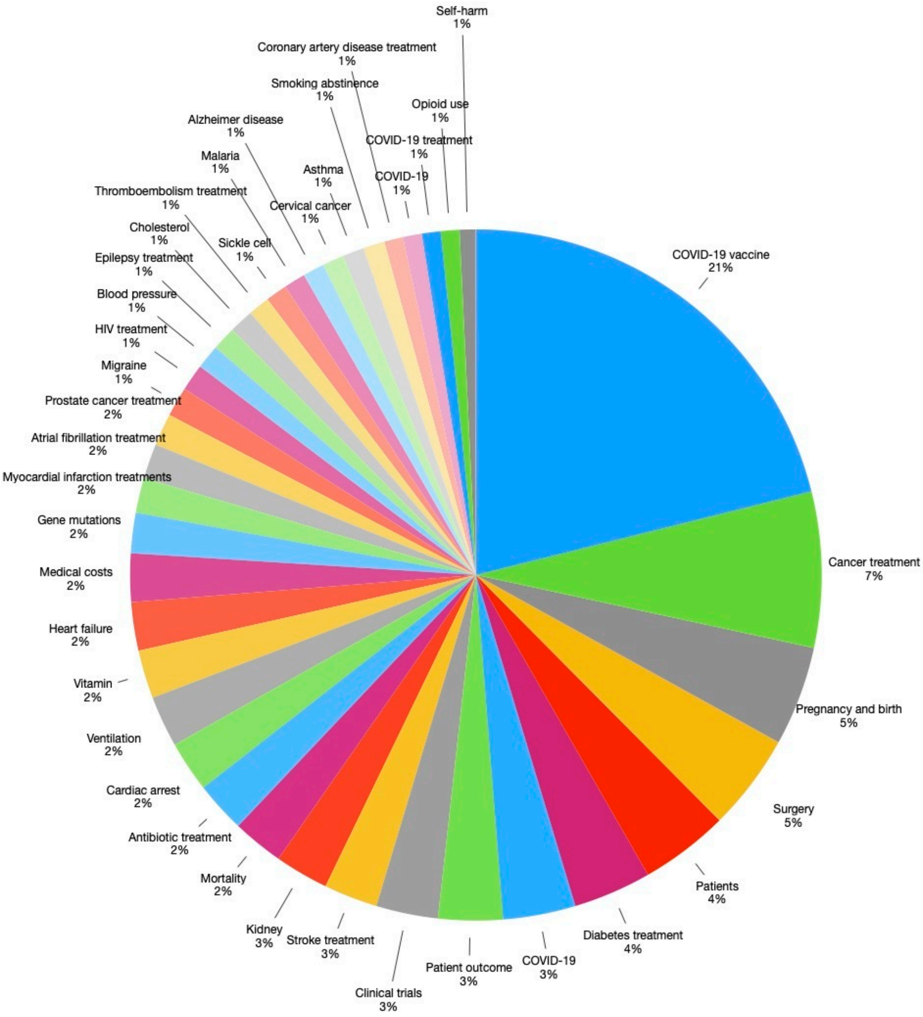

Figure 1. The 39 topics suggested by the natural language processing classification model for the articles published in top medical journals between 2020-2022

cluded articles after manual exclusion. The majority were for NEJM (705, 36.4%), followed by Lancet (448, 23.1%), JAMA (419, 21.6%), and BMJ (367, 18.9%).

TOPIC CLASSIFICATION MODEL

The model was able to categorize 1,540 of the publications (79.42%) into 39 different topics. The most common topics published in top medical journals were involving the COVID-19 vaccine, cancer treatments, pregnancy, and birth. Figure 1 shows the 39 topics published by top medical journals and the proportion of publications in each topic.

Despite separating the topics into 39 distinct ones, some of these topics were related and were discussed in the context of each other. The model generated a similarity matrix that shows the degree of similarity between topics. A high similarity index (i.e., above 80%) was found between COVID-19 vaccine and vitamin (85.5%); prostate cancer and survival (83.4%); COVID-19 vaccine and antibiotic (85%); COVID-19 vaccine and COVID-19 (87.7%); COVID-19 and cardiac arrest (86.6%); surgery and cardiac arrest (86.2%); cancer treatment and kidney (82%); COVID-19 and mortality (82.4%).

According to field categorization, 11 different fields resulted. Figure 3 shows the proportion of articles contributed by each journal in each field. The difference in topic distribution was significantly different between journals (p< 0.001).

BIBLIOMETRIC RESULTS

A significant difference between top journals in respect to:

- The number of citations received (p< 0.001), with Lancet receiving the highest mean (265.36 ±1224.04) and BMJ receiving the lowest mean (64.34 ±202.31) citations.
- The number of references in each article (p<0.001), with BMJ articles having the highest number of references (42.6 ±23.45) and NEJM having the lowest mean (29.39 ±9.84) number of references.
- The number of authors in each article (p< 0.001), with Lancet having the highest mean number of authors

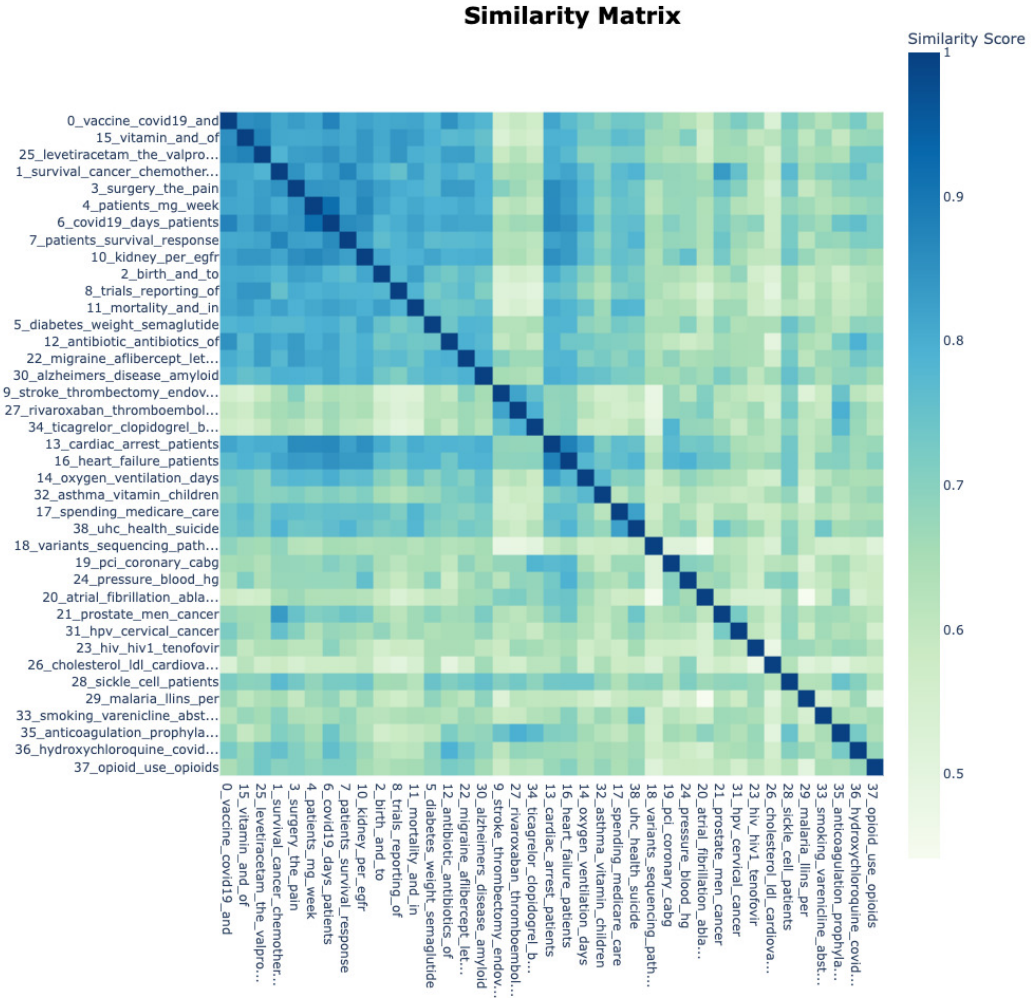

Figure 2. Similarity matrix of the separate topics

(61.03 ±212.93) and BMJ having the lowest mean number of authors (14.38 ±11.21).

- The number of pages ($p < 0.001$), with Lancet having the highest mean number of pages (11.96 ±4.65) and JAMA having the lowest mean number of pages (10.39 ±1.77).

Table 1 details the mean number of citations, references, authors, and pages for articles published in top medical journals.

DISCUSSION

We developed a novel method based on the newly developed NLP model BERTopic.[5] This method can be used to analyze the topics discussed by a large number of articles, where the model can have articles' abstract as the input, and it will provide a number of valuable outputs, including the most common topics discussed, their relation to each other, and the similarity between them. We applied this novel method to articles published by general medical journals in the last three years. We found that almost 80% of published articles were related to 39 distinct topics in 11 fields. We also found that each journal focused on certain topics more than others, rather than being general. For example, JAMA published the highest percentage of pulmonology-related articles compared to other journals, while NEJM published the highest number of nephrology and genetics-related articles. We also performed a bibliometric analysis for the included articles, showing significant differences between journals regarding the number of citations, and number of authors per article, along with several references and pages in each article. We believe such methodology can be applied to other specialties and journals, helping researchers and policymakers understand the fields that are currently well-published and the topics that are now being researched.

Approximately 25% of all the papers published in these 4 journals during 2020-2022 were related to COVID-19 and its treatment. With the preceding global COVID-19 pandemic, this was not surprising. This rapidly increasing momentum that COVID-19-related articles had in literature has been described as one of the biggest explosions of scientific literature ever, with >450000 articles published within a year of the emergence of the disease.[7] Despite the large percentage of papers being sorted into COVID-19,

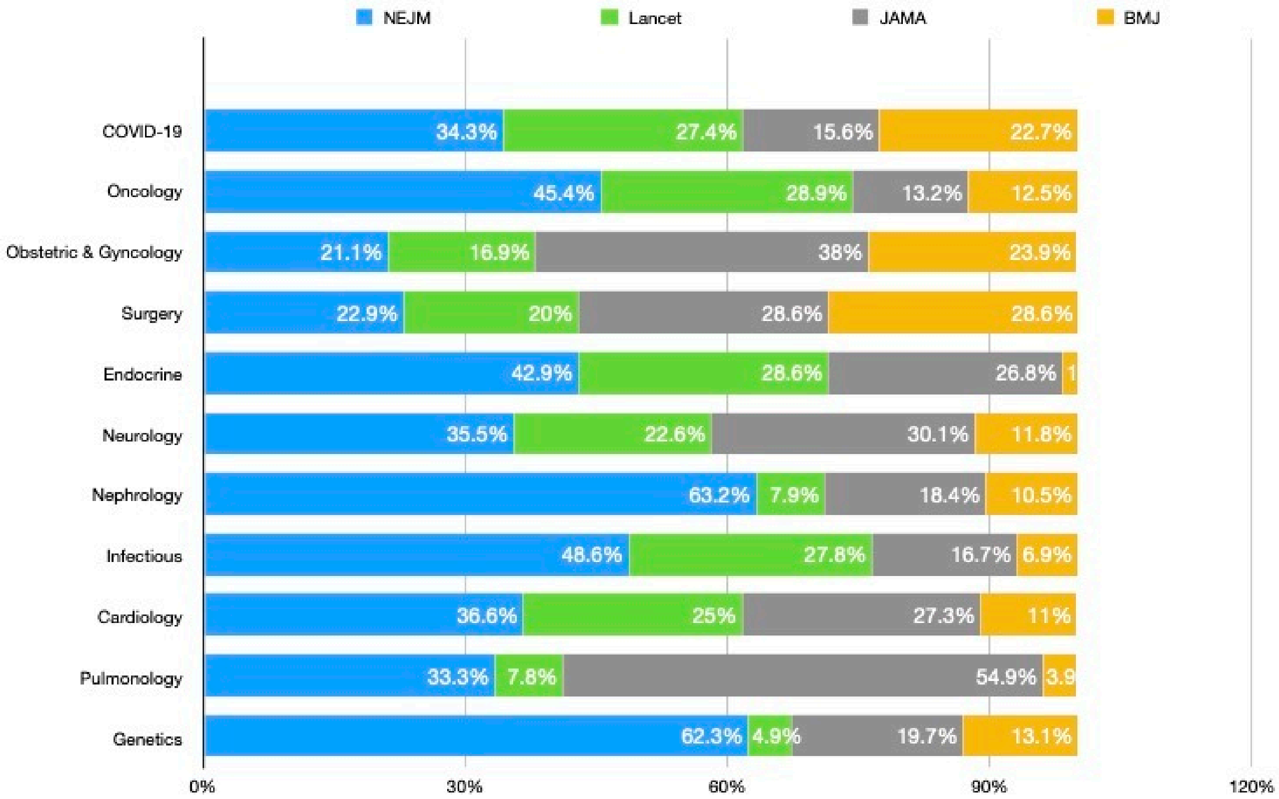

Figure 3. Distribution of papers published by each journal among different fields.

some topics, such as Vitamins, showed a very high degree of similarity with COVID-19, which could mean that the proportion of studies regarding COVID-19 and its treatment is underestimated as these studies could be describing the same topic. The topics published continue to follow an expected trend, with more common entities having more papers published on them. Cancer is the second most common cause of death worldwide (WHO), and 7% of all papers published in these journals have been about the treatment of different cancers. Our model was able to identify these topics among 1500 papers, this can potentially be replicated on a larger scale with more papers, more journals, and over a longer period of time. The ability of the model to do so with minimal manual labor means that identifying trends in literature becomes a much easier and more efficient task. Trends in different topics may enable researchers to gain a deeper understanding of how their field is evolving and potentially identify what factors may be driving these changes.

Each journal contributed a different number of papers to each field, with some journals publishing more papers in some fields and fewer in others. The NEJM contributed 62.3% of all genetics papers included in this study, while the Lancet contributed 4.9%. Since the NEJM included the largest number of studies, it can be expected to contribute a larger percentage to most topics than other journals, however, the difference in the percentages contributed by Lancet and NEJM cannot be explained solely by differences in the number of papers published. The plot shows that different journals tend to publish a different proportion of papers in certain fields when compared to other journals, and therefore when researchers attempt to publish in journals, selecting journals that have published in their desired fields may be worth considering.

Even though topic modeling with conventional methods is modestly discussed in the literature[8-11] and has produced encouraging results, it has not yet revolutionized research methods and conduct. As far as the authors are concerned, there are no previous studies using BERTopic to model topics in medical literature, the most common uses for BERTopic have been limited to analyzing different topics across social media and other media platforms.[12-14] BERTopic and traditional methods (LDA for example) are two different approaches to topic modeling. While LDA has been the traditional method of topic modeling for many years, BERT is a more recent approach and one of the goals of our study was to assess how effective it is in the context of medical literature. One of BERT's distinguishing features is its ability to use contextual embeddings and bidirectional processing, both of which provide BERTopic with a more comprehensive understanding of the data and allow it to better capture the underlying topics,[13,15,16] whereas traditional methods utilize a statistical approach and only consider the frequency of words within a document to draw conclusions. BERTopic and its word embeddings can be trained on massive amounts of text data, allowing them to capture patterns in the data that traditional methods cannot. The above-mentioned features are particularly useful

Table 1. Bibliometric results of the papers included.

		N	Mean	Std. Deviation	Sig.
References in each article	NEJM	705	29.39	9.842	<0.001
	Lancet	448	37.96	23.303	
	JAMA	419	30.86	8.826	
	BMJ	367	42.60	23.450	
	Total	1939	34.19	17.566	
Citation according to Web of Science	NEJM	705	243.93	935.012	<0.001
	Lancet	448	265.36	1224.041	
	JAMA	419	94.26	349.420	
	BMJ	367	64.34	202.307	
	Total	1939	182.55	839.350	
Citations in all databases	NEJM	705	252.08	999.733	<0.001
	Lancet	448	279.07	1314.041	
	JAMA	419	96.59	355.416	
	BMJ	367	66.19	209.300	
	Total	1939	189.53	897.148	
Number of pages	NEJM	705	10.93	1.960	<0.001
	Lancet	448	11.96	4.651	
	JAMA	419	10.39	1.769	
	BMJ	367	11.08	3.440	
	Total	1939	11.08	3.096	
Number of authors	NEJM	705	24.56	20.775	<0.001
	Lancet	448	61.03	212.927	
	JAMA	419	20.60	18.984	
	BMJ	367	14.38	11.207	
	Total	1939	30.20	104.948	

in medical research, where the context of words is critical, and the meaning of certain words differs from their use in other fields.

Despite the novelty of the methodology, the study has several limitations that need to be considered before using it. The model could not classify around 20% of the inputted abstracts. However, such a proportion of unclassified abstracts will not affect the overall trend. Even though most classification stages have been automated, it is still necessary to manually classify the extracted topics because the model will provide researchers with five descriptive words for each topic, and they must then give each word cluster a topic name. Finally, while we extracted data from the Web of Science database, one of the well-curated literature databases,[17] the trends extracted were based on three years of publishing history in the top four medical journals. Future studies might need to consider a larger number of journals over an extended period beyond the COVID-19 publishing pandemic.[18]

CONCLUSION

We demonstrate a novel method of using the BERTopic NLP model to analyze topics discussed in large numbers of articles. The study shows that this methodology can identify the most common topics discussed in medical literature while demonstrating the effectiveness of this approach in identifying and extracting relevant information from a large amount of data. This methodology has the potential to analyze trends in medical literature more efficiently and effectively, potentially enabling physicians and researchers to gain a deeper understanding of how their field is evolving. Furthermore, this study's methodology can be replicated on a larger scale with more papers, more journals, over a longer period, and over different fields. This study also highlights the importance of further research using BERTopic and other NLP models, especially in fields such as the medical field where it is becoming difficult to overcome the challenges of dealing with large volumes of complex data, as NLP has the potential to revolutionize research methods and conduct.

CORRESPONDING AUTHOR

Saif Aldeen AlRyalat: The University of Jordan, 11942 Amman, Jordan. Mobile: +962-798914594. Email: s.alryalat@ju.edu.jo , saifryalat@yahoo.com

COMPETING INTERESTS

All authors declare that they have no competing interests.

Submitted: March 03, 2023 AST, Accepted: April 16, 2023 AST

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at http://creativecommons.org/licenses/by/4.0 and legal code at http://creativecommons.org/licenses/by/4.0/legalcode for more information.

REFERENCES

1. Abdalla SM, Solomon H, Trinquart L, Galea S. What is considered as global health scholarship? A meta-knowledge analysis of global health journals and definitions. *BMJ Glob Health*. 2020;5(10):e002884. doi:10.1136/bmjgh-2020-002884

2. Scaccia JP, Scott VC. 5335 days of Implementation Science: using natural language processing to examine publication trends and topics. *Implement Sci*. 2021;16(1):47. doi:10.1186/s13012-021-01120-4

3. Jung KY, Kim T, Jung J, et al. The Effectiveness of Near-Field Communication Integrated with a Mobile Electronic Medical Record System: Emergency Department Simulation Study. *JMIR Mhealth Uhealth*. 2018;6(9):e11187. doi:10.2196/11187

4. Lee M, Wang W, Yu H. Exploring supervised and unsupervised methods to detect topics in biomedical text. *BMC Bioinformatics*. 2006;7(1):140. doi:10.1186/1471-2105-7-140

5. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. *Advances in Neural Information Processing Systems*. 2017;30:5998-6008. https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

6. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://arxiv.org/abs/2203.05794

7. Al-Habsi J, Al-Hatmi A, Al-Saadi T. The Top 100 Most Cited Neurosurgical Articles on COVID-19: A Bibliometric Analysis. *World Neurosurg*. 2023;170:22-27.e21. doi:10.1016/j.wneu.2022.11.133

8. Berrang-Ford L, Sietsma AJ, Callaghan M, et al. Systematic mapping of global research on climate and health: a machine learning review. *Lancet Planet Health*. 2021;5(8):e514-e525. doi:10.1016/s2542-5196(21)00179-0

9. Song Y, Ni Z, Li Y, et al. Exploring the landscape, hot topics, and trends of bariatric metabolic surgery with machine learning and bibliometric analysis. *Ther Adv Gastrointest Endosc*. 2022;15(26317745221111944):263177452211119. doi:10.1177/26317745221111944

10. Sing DC, Metz LN, Dudli S. Machine Learning-Based Classification of 38 Years of Spine-Related Literature Into 100 Research Topics. *Spine*. 2017;42(11):863-870. doi:10.1097/brs.0000000000002079

11. Danilov GV, Shifrin MA, Kotik KV, et al. Artificial Intelligence in Neurosurgery: a Systematic Review Using Topic Modeling. Part I: Major Research Areas. *Sovrem Tekhnologii Med*. 2020;12(5):106-112. doi:10.17691/stm2020.12.5.12

12. Ng QX, Yau CE, Lim YL, Wong LKT, Liew TM. Public sentiment on the global outbreak of monkeypox: an unsupervised machine learning analysis of 352,182 twitter posts. *Public Health*. 2022;213:1-4. doi:10.1016/j.puhe.2022.09.008

13. Baird A, Xia Y, Cheng Y. Consumer perceptions of telehealth for mental health or substance abuse: a Twitter-based topic modeling analysis. *JAMIA Open*. 2022;5(2):ooac028. doi:10.1093/jamiaopen/ooac028

14. Zankadi H, Idrissi A, Daoudi N, Hilal I. Identifying learners’ topical interests from social media content to enrich their course preferences in MOOCs using topic modeling and NLP techniques. *Educ Inf Technol (Dordr)*. Published online November 4, 2022:1-18. doi:10.1007/s10639-022-11373-1

15. Dieng AB, Ruiz FJR, Blei DM. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*. 2020;8:439-453. doi:10.1162/tacl_a_00325

16. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org. https://arxiv.org/abs/1810.04805

17. AlRyalat SAS, Malkawi LW, Momani SM. Comparing Bibliometric Analysis Using PubMed, Scopus, and Web of Science Databases. *J Vis Exp*. 2019;(152). doi:10.3791/58494

18. Palayew A, Norgaard O, Safreed-Harmon K, Andersen TH, Rasmussen LN, Lazarus JV. Pandemic publishing poses a new COVID-19 challenge. *Nat Hum Behav*. 2020;4(7):666-669. doi:10.1038/s41562-020-0911-0