Towards Advancing Creativity in Libraries Using Natural Language Processing for Data Curation: A Bibliometric Analysis

Pooja Rana and Pramod Kumar Singh

Natural language processing has significantly become one of the most advancing topics in the field of artificial intelligence. Knowing its capabilities of using semantics to extract relevant information from large volumes of data, its applications for data curation are worth exploring. NLP has an increasingly signifying role in different fields; therefore this study aims to examine the research output of the NLP for data curation. For this purpose, a bibliometric study on NLP for data curation is presented and research publications are retrieved from Scopus Database. The network analysis is done using the VOSviewer software. The results have shown an increase in the publication from the year 2014-2021 with an average of 29-30 publications each year. The USA is found to be dominating the field among other countries. Among the most prolific authors,Ananiadou, Vijay-Shanker, and Hirschman lead the list, accounting for 7.63% of the total output. The five main thematic areas identified are natural language processing, human, information processing, information retrieval, and data mining. The bibliometric analysis of NLP for data curation, uncovering the present status, paves the way for future research opportunities in libraries.This study will allow us to explore and understand research developments systematically and optimize libraries by employing NLP models for data curation for effective transmission of data for the end users.

Introduction

Data curation is the process of integrating, creating, organizing, discovering, and maintaining data sets so that they can be accessed by those who need them. Data curation is an effective way for an organization to extract real service value from its data (Thirumurugathan, 2020). In simple words, we can say that it is the process of ensuring that data is findable and usable now and in the future. Data curation ensures the value of data for a long time and fixes problems, may it be fixing issues with the data of unexpected values, mysterious variables, and data containing too many details.

Data collection needs to be presented in a contextual and meaningful way, although this involves significant challenges and flexible and highly scalable infrastructure to curate with large and heterogeneous data sets. Proper metadata, data recording, providing unique identifiers for data, resolving issues through citations, knowing policies, dealing with selection, its usage, and level of service are some of the challenges on the surface. According to Witt (2022), these issues are in principle with library science and librarians do have the skills to deal with such issues of data curation. This new field of data curation specifically addresses the need of stewarding and preserving digital research data.

Librarians assist users in the efficient use of information. However, the amalgamation of computational techniques in handling data sets in libraries is one such theoretically motivated concept. Natural language processing (NLP hereafter) is a computational technique for automatic analysis and helps the computer communicate with the human in their language. Its capabilities are farsighted. NLP makes it possible for computers to hear, read and interpret human language, and reflect human sentiments. Chen et al. in year 2018 presented a paper stating its rich research achievements and application in medical science. The voluminous amount of data is hidden in free text in unstructured form. NLP is crucial for transforming it into structured information. Therefore it is extremely useful in the process of data curation in libraries as it helps data science in extracting insights from complex data structures.

NLP-empowered research is accumulating at a fast pace and draws the attention of the research field. Thus, this paper attempts to find out the intensive research output of applying natural language processing techniques as an aid to data curation using SLR and bibliometric analysis. Bibliometric is a powerful technique that includes evaluating leading scientific research and doing statistical analysis for quantitative assessment of the research output. This paper addresses the questions and presents the results of research output on integrating natural language processing into the curation workflow in different countries and different fields by surveying the papers published on Scopus using VOSviewer software.

2. Review of Literature:

Brackle (2011) discussed the emerging role of librarians in data curation by conducting a case study of agriculture data. In this paper, he presented ways in which agricultural librarians can transform their strategies to deal with data to meet the emerging needs of scientists and users for advanced research in the field and agricultural purpose. He conducted this study at Purdue university libraries to explore their experience with evolving roles in the scientific environment which include knowledge, training, experimentation, and scientists involved in data curation. He also discussed the barriers and challenges associated with data curation. Another such consideration has been given by Weber, Palmer, & Chao (2012) in which they highlighted the current and future trends in data curation research and education. They offer an overview of data curation research and education by evaluating current trends in education, practice, skills, and workforce development. The recommendations emerged out of discussions among more than 50 leading experts from government agencies, data centers, publishing houses, libraries, and Information Science. Based on that, they presented a future map for both basic and applied research in the development of data curation. In the context of data curation, the theory by Bishop, Grubesic, & Prasertong (2013) used a combined method to deal with data in geographical information libraries. In this paper, they introduced geospatial data services in libraries and discussed the current challenges of using GeoWeb by information professionals. They found that this technology has altered the role of data curation and information provision in geographic information libraries. In the end, they concluded with an overview of the importance of incorporating GeoWeb for future data curation training. This provides a dynamic and holistic approach to dealing with data. When addressed in

the context of a bibliometric study on natural language processing Chen et al. (2017) conducted a bibliometric analysis of natural language processing in medical research. They collected NLP-empowered medical research from PubMed during the period 2007-2016. Based on the analysis, they revealed that the average annual growth rate of NLP-empowered medical research is 18.39%. The USA has the highest number of publications. This study is useful to understand the research development systematically. Another important theoretical contribution in the field is provided by Vijaykumar & Sheshadri (2019), in which they explained the compressive applications of artificial intelligence in Academic libraries. They expressed the application of artificial intelligence such as artificial intelligence, expert system, image processing, natural language processing, speech recognition, robotics, etc. They further explained the possible areas of library science where artificial intelligence can be applied to enhance the quality of services and thereby depicting its potential impact on libraries. Also, Cheng et al. (2020) analyzed the combined method for NLP libraries for analyzing software documents. In this paper, they discussed numerous toolkits on the selection of NLP library selection and their pitfalls. The study emphasized utilizing the strengths of different NLP libraries using the combined method to obtain accurate results. The combination consists of two steps, i.e. document level selection of primary library and sentence level overwriting. This resulted in effective results in terms of accuracy.

The above literature review is evidencing the fact that no studies have been conducted pertinent to the use of natural language processing for data curation concerning libraries. The literature review reveals that the concept is in phase from a decade ago yet its application for data curation is in its infancy. Based on the content analysis of the articles published in the recent year, we identify the gap and opportunities for NLP research as an aid for data curation concerning libraries. NLP can be used for different business purposes which may include data analytics, User Interface (UI) optimization, and value proposition. Today the market is flooded with different natural language processing tools. In a nutshell, NLP, a popular branch of AI does semantic analysis, yields crucial insight, and improves overall performance. Its application for data curation in libraries is endless which includes email classification and filtering, sentiment analysis, language translation, structuring queries, etc. Thus it becomes even more complacent to conduct a bibliometric study on such a topic.

3. Methodology

This study is based on a bibliometric analysis of natural language processing for data curation concerning libraries. The data is collected from the Scopus database, which is the most reliable. It is a comprehensive database of the world's output in the field of technology, science, and social science and covers articles from various other disciplinary as well as interdisciplinary subjects. The query for the term "natural language processing for data curation and libraries" yielded 327 results. The search operations used are phrase and Boolean operators. The time limit set for the retrieval of data is from 2003- 30 June 2022. The bibliometric data is analyzed using VOSviewer software, which made it possible to present the data graphically through network visualizations. The graphical representation of the data is necessary to better understand what has

been researched in the field of NLP and data curation. Bibliometric indicators are used as a mechanism to analyze and interpret data collected to understand the mapping of main trends in the area. The main objectives followed in this approach include analysis of:

a) Annual trend in publication, Source-wise publication, Funding and sponsors analysis, and Authorship;

b) Co-authorship by countries, Bibliographic coupling by authors, Author-citation networks analysis, and Co-occurrence of keywords;

c) Research opportunities for libraries on NLP for data curation.

4. Data Analysis:

4.1. Distribution trend of published articles

The first article related to natural language processing identified in Scopus is from 2003 with 2 articles. The literature is growing tremendously after the year 2013 from 10 articles in the year 2013 to 27 articles in the year 2014 which is a percentage increase of 170%. The publication during 2014-2021 ranges from 27-36 articles which means an average of 29 articles is published each year. The literature on NLP and data curation is the maximum in the year 2021 with 36 articles. However, the year 2022 only shows 11 articles as the data is taken till 30th June 2022. More articles are expected to be published in the year 2022. Figure 1 illustrates the annual trends in the publication of natural language processing and data curation concerning libraries.

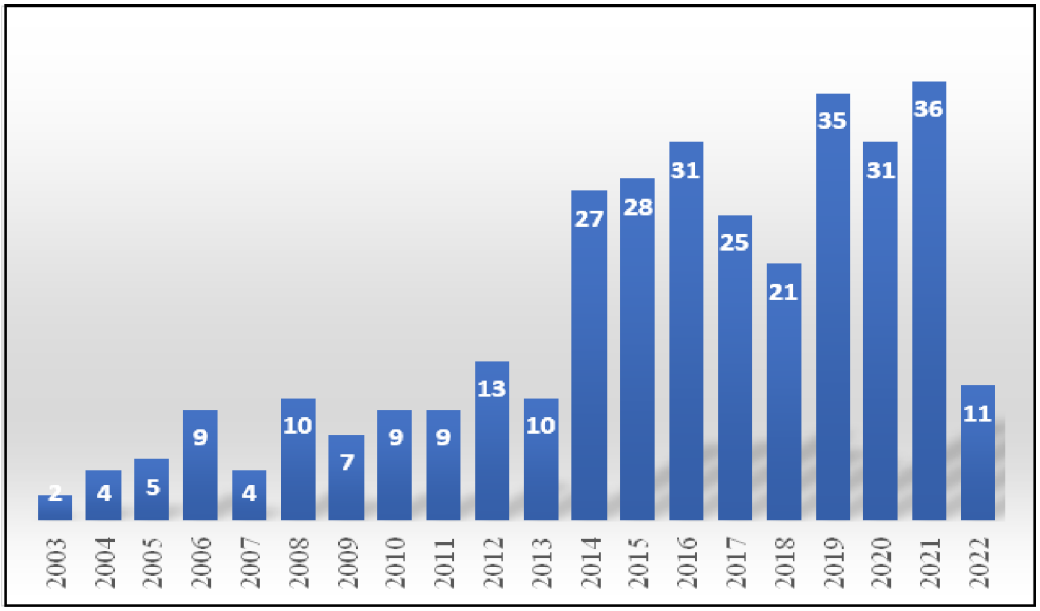

Figure 1: Annual output on NLP for data curation during 2003-2022

4.2. Source-wise publication (NLP for data curation)

Analysing the source of documents in which a total of 327 papers are published, we note that out of the different sources in the field identified in Scopus (shown in Table 1), BMC Bioinformatics (n=27, where n denotes the number of publications) and Journal of Biomedical Informatics (n=27) are found to be most prominent in the field of NLP and data curation publications, followed by Studies in Health Technology and Informatics (n=12), Databases (n=12) and Bioinformatics (n=10). Out of the top 20 sources identified, a major share of publications is related to the field of health and medical science.

Table 1: shows the list of top sources in the publication output on NLP for data curation during 2003-2022

Source Title	No. of documents
BMC Bioinformatics	27
Journal of Biomedical Informatics	27
Studies in Health Technology and Informatics	12
Database	12
Bioinformatics	10
Lecture notes in computer science including subseries Lecture Notes in Artificial Intelligence And Lecture Notes in Bioinformatics	10
Ceur Workshop Proceedings	9
Database: The Journal of Biological Databases and Curation	9
Plos One	7
AMIA Annual Symposium Proceedings AMIA Symposium	6
Journal of the American Medical Informatics Association	6
Pacific Symposium on Biocomputing	6
ACM International Conference Proceeding	5
Communications In Computer And Information Science	5
JCO Clinical Cancer Informatics	5
BMC Medical Informatics and Decision	4
Briefings in Bioinformatics	4
Nucleic Acids Research	4
Journal of Biomedical Semantics	3
Scientific Reports	3

As seen in Table 1, we clearly analyzed that these 20 journals alone published 51.37% of the total publication output. Table 1 is an indicator that the most prominent sources in the publishing of articles on natural

language processing and data curation come from the field of medical science. Although the research topic is very transversal and can be published in different fields, only a few journals and source types have published articles on the topic from different fields.

4.3. Co-authorship network by country

Iis a professional network of researchers among various countries; it is an attempt to show how fragmented the research community is and the extent of collaborative authors in the networks (Kumar, 2015). The cluster here in figure 2, shows countries, lines connecting the points indicate links of the co-authorship by countries and the distance between them shows the strength of cooperative relationships between countries. Thus the map indicated the cooperative relationship between different countries in NLP and Data curation research.

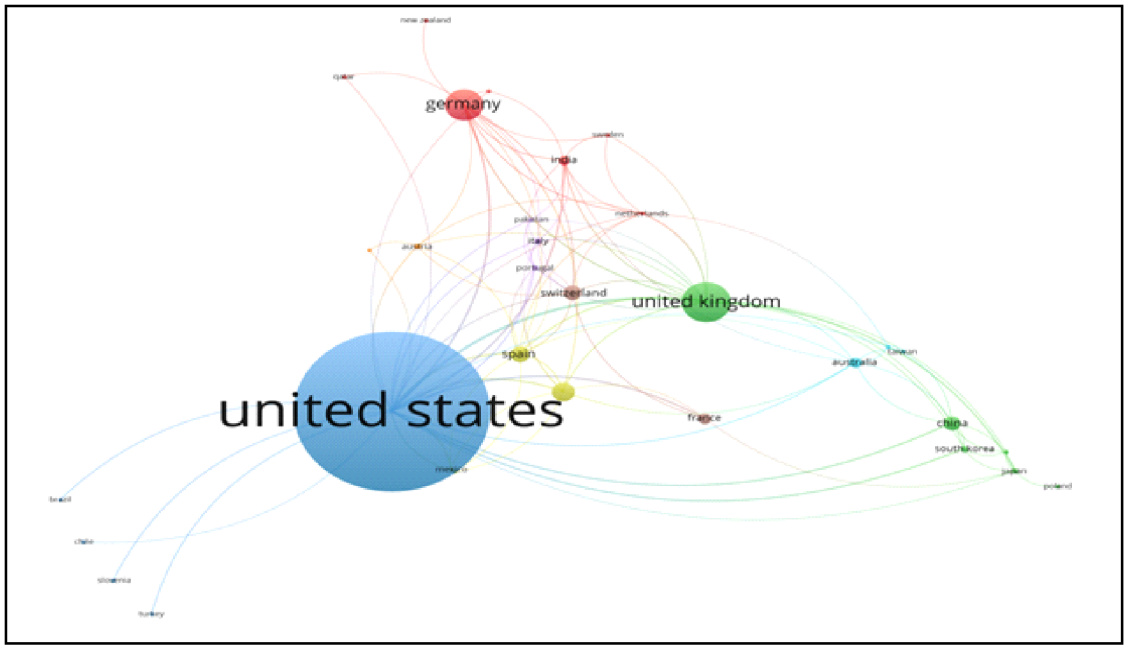

Figure 2: Co-authorship by countries

Figure 2 shows the country co-authorship network map drawn from the sample of 327 papers. A total of 52 countries contributed papers on NLP and Data curation in the year 2003-2022. This means that these countries have published at least one paper on the research topic.Through the map, it is possible to see that the cluster of countries i.e. United States, the United Kingdom, and Germany together account for 74.60% of the publication, followed by Canada (6.11%), Spain (5.19%), Switzerland (4.58%), China (3.97%), France (3.96%), Australia (3.36%) and India (3.05%). Therefore, through the geographical scattering of the publications, it is possible to notice that publications on the research topic are scattered globally.

4.4. Funding/Sponsor-wise publication:

The top 10 funding agencies supporting research in Natural Language Processing for data curation are listed in table 2. The table clearly shows that most of the funding agencies are institutions dedicated to medical science. The leading sponsor for the field articles is the National Institute of Health with 48 articles, followed by the U.S. National Library of Medicine with 36 articles and the National Science Foundation with 25 articles. It is evident from the list that natural language processing models are being implemented in the health science sector and at par supported by the various funding agencies.

Table 2: List of top funding institutions for research on NLP for data curation during 2003-2022

Funding Sponsor	Documents
National Institutes of Health	48
U.S. National Library of Medicine	36
National Science Foundation	25
National Institute of General Medical Sciences	22
U.S. Department of Health and Human Services	17
National Cancer Institute	15
National Human Genome Research	15
Biotechnology and Biological Sciences Research Council	13
National Centre for Research Resources	8
National Centre for Advancing Translational Sciences	7

4.5. Publication by author

This analysis addresses the authors leading in publication and production in the field of the research area. Table 3 identifies the most prolific author in the field. A total of 157 authors were identified who contributed to the publication of natural language processing and data curation. This data shows that at least 2 papers have been published by each author. S. Ananiadou from the University of Manchester, National Centre of Text Mining has 10 publications (3.05%), followed by K. Vijay-Shanker from the University of Delaware, Department of Computer and Information Science with 8 publications (2.44%). Although the research seems quite prevalent in the field of medical science, it is clear that Author contributions are quite fragmented. The list below shows the authors from the field of computer sciences and text mining, Bioinformatics and computational biology lead the list.

Table 3: shows the list of the top ten leading authors in research output on NLP for data curation during 2003-2022.

Rank	Author	Articles	%publication
1	Ananiadou, S.	10	3.05%
2	Vijay-Shanker, K.	8	2.44%
3	Hirschman, L	7	2.14%
4	Valencia, A.	7	2.14%
5	Wu, C.H.	7	2.14%
6	Krallinger, M.	6	1.83%
7	Lu, Z.	6	1.83%
8	Altman, R.B.	5	1.52%
9	Coulet, A.	5	1.52%
10	Bourginoe, P.	4	1.22%

This network visualization shown in the figure represents the bibliographic coupling of authors. Bibliographic coupling is one good way to analyze how authors use and strengthen links among existing literature (Biscaro & Giupponi, 2014).

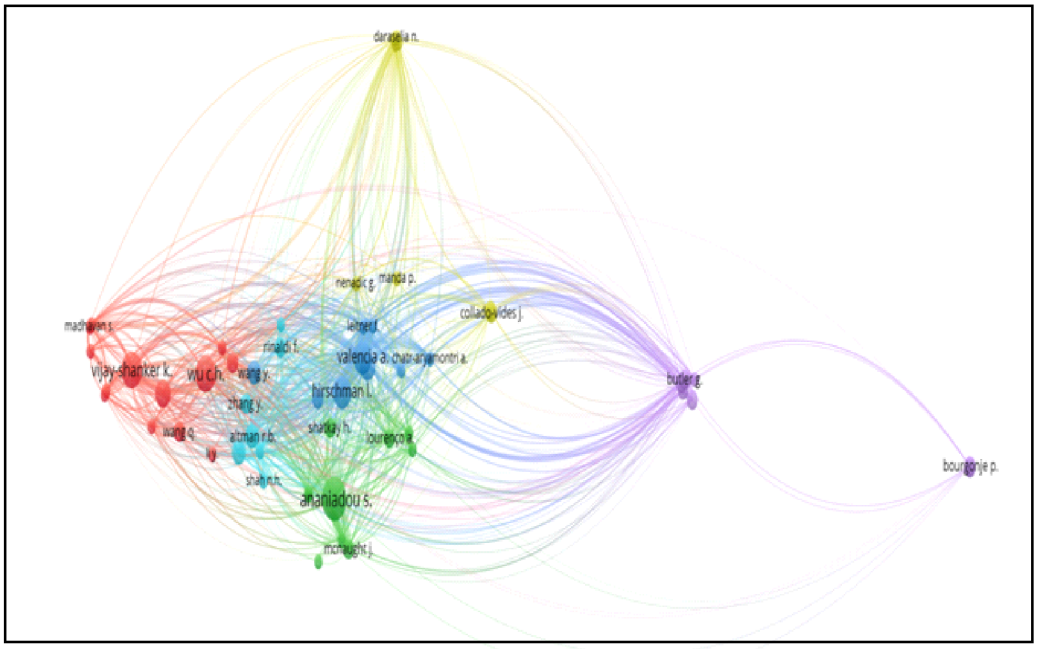

Figure 3: Bibliographic coupling of authors

Figure 3 clearly depicts five major clusters in which publications and authors are co-related through multiple citations. The line between these clusters shows the concomitant citations between authors. The bibliographic coupling can be applied to the author, documents, organization, countries, and source. This is a very effective method to develop deep insights into scientific activityand help solve conjectures that have not yet been explored.

4.6. Author citation network

Figure 4 shows the author citation network map to understand and review the work of the contributing authors towards Natural language processing and data curation.

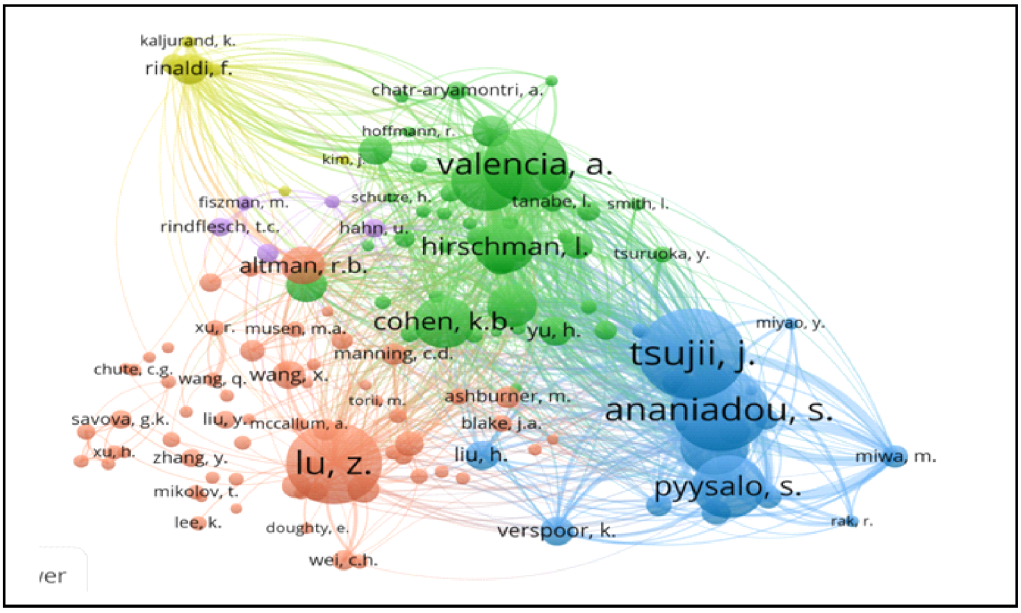

Figure 4: Network visualization of author-citation analysis.

4.7. Co-occurrence of keywords

The analysis of keywords reflects that from the 327 publications, 2866 keywords were identified. Out of these publications, 800 words appeared 2 times which is equivalent to 27.91% occurrence. More than 205 articles meet the threshold when analyzing publications with prevalence order more than 5 times. It is evident from figure 5 that natural language processing is the most prominent keyword used in the articles, i.e. 230 articles, which also summarizes the main subject of this article, followed by information processing with 110 occurrences. The main clusters of the keywords found are natural language processing, information processing, information retrieval, and data mining. The co-occurrence of keywords depicts the effective network of crucial knowledge components that represents the strength of links between keywords that appear in the literature (Mustak et al. 2020).

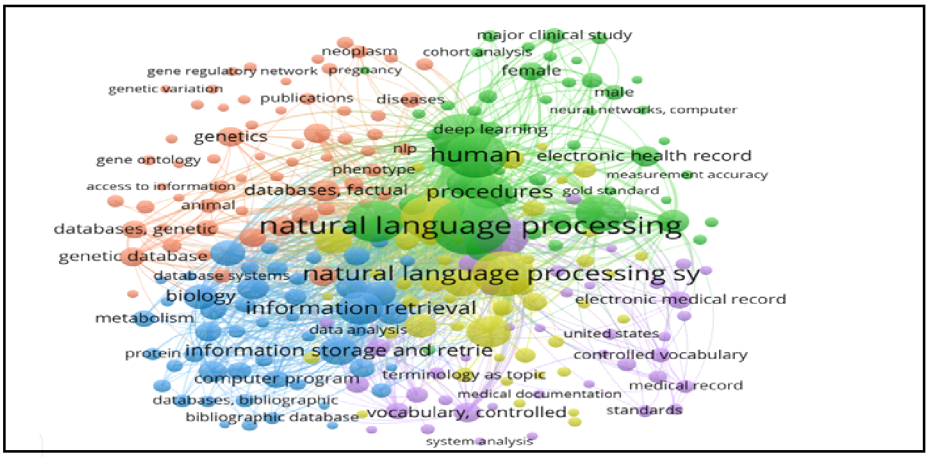

Figure 5: Co-occurrence of keywords

5. Research opportunities on Natural language processing for data curation in libraries

Through review of the literature, it is evident that not much research has been done in the field of libraries concerning the use of Natural language processing for data curation. Data curation is an important process and is one good way for libraries, particularly digital libraries, to organize their data and maintain data efficiently so that they are accessible to the users. The use of NLP in libraries can make this task even more effective, by employing its semantic capabilities which distinguishes it from other information retrieval models. Much research needs to be conducted in this field as their applications in libraries are wide. Here are some points that highlight its applications in libraries for data curation.

5.1. NLP for text summarization

Text summarization has been a topic addressed in research and studies. It is a way of creating summaries of the whole document by taking the key concepts of the document, may it be a single document or multi-documents. It becomes quite a difficult task for the information managers to generate summaries of the data thereby making it difficult for the users to access information efficiently. According to Mishra et al. (2014) studies have shown that natural language processing in combination with machine learning models has been used to generate extractive summaries. These models use algorithms to identify concepts to construct a semantic graph, and then by using complex algorithms they identify main themes and topics to extract salient concepts within. Text summarization is an important feature of data curation and helps in the easy identification of the relevant needs of the users. Therefore, it must be addressed in the scope of libraries.

5.2. Searching and Indexing

NLP uses highly scalable statistical techniques to index documents and its capabilities are beyond imagination for searching voluminous amounts of text efficiently. NLP uses its novel indexing technique to compress

complex datasets into logical syntactic components to capture basic dependency relations like subject-of, object-of, and verb-modification like time, location, etc. (Kao & Poteet, 2007). It uses semantics to extract meaning from the text. Using NLP for data curation in libraries can transform its approaches to deal with specific data. Its higher user friendliness and ease of learning make it highly approachable to non-specialists as it focuses on how to use a particular tool rather than getting deep into the mathematical principles (Nadkarni, 2002).

5.3. Text classification

NLP employs extensive statistical techniques to find word associations among datasets which are summed to require parsing into natural language phrases. Ontologies are an important consideration for most NLP applications. Text classification is an important consideration in the task of data curation. A study by Lytvynenko (2019) revealed that the use of NLP is of great purpose. This technique use text stemming to remove morphological affixes from the text, and also removes copula which is not necessary for further analysis. All these steps help libraries to fulfill the main purpose of data curation and do an accurate analysis of the text data. This method is useful in getting the best clear features for the future classification of the text. It is an important process of extracting relationships between entities. Kumar et al. (2016) introduced the dynamic memory network based on a natural language processing model built on neural network architecture. This model was capable of processing input sequences and questions, forming episodic memories, and presenting relevant answers to the questions. In addition to this, it is capable of doing text classification for sentiment analysis for several types of tasks and datasets.

6. Discussion and Conclusion:

This article contributes to the existing piece of knowledge in the field of data curation in libraries using Natural language processing systems. This study used bibliometric analysis to identify trends, gaps, and research opportunities in the field. In the era of dynamic technological changes and artificial intelligence, it came as a surprise that not much literature is available on the use of NLP for data curation concerning libraries. The results have shown that natural language processing models have been widely used in medical science for managing information from patients and staff. Even the bibliometric study conducted on NLP is in the field of medical science. It is important to explore, understand and discuss the state of the art or present status of NLP for data curation concept. Keeping in view the literature collected through Scopus from the year 2003-2022, the growth has been phenomenal tractioning its growth from 2 in the year 2003 to 36 in 2021. The sources marked this technology quite prevalent in the medical science. The results have shown that there is an increasing trend in collaboration among the authors as well as the countries. Authors from different disciplines have contributed to the publication output. The results have used network visualization maps to better understand the links, citation analysis, and collaboration between authors. The map indicated the cooperative relationship between different countries on NLP and data curation research, revealing the scattering of publications worldwide. This article identifies the scope for

future research work in the field of NLP in libraries as its application in text summarization, indexing, and searching; content chaining, text classification, and text mining are well needed in the process of data curation in libraries. So that it can help in the effective transmission of relevant information to those who need them. Libraries possess skills that are necessary to eradicate the challenges connected to data curation and the user-friendliness of NLP models makes it easier for the non-specialist to interact with the systems by getting deep into the programming process. Librarians are needed to conduct more research in the field for organizing and maintaining data in the libraries and adding value to their services.

References

1. Biscaro, Claudio, and Giupponi, Carlo. (2014). Co-authorship and bibliographic coupling network effects on citations. Plos One, 9(6), 1-12.doi: doi:10.1371/journal.pone.0099502.g002

2. Bishop, Bradley Wade, Grubesic, Tony, and Prasertong, Sonya. (2013). Digital curation and the geo-web: an emerging role for geographical information librarians. Journal of Map & Geography Libraries, 9(3), 296-312. doi:https://doi.org/10.1080/15420353.2013.817367

3. Brackle, Marianne Stowell. (2011). Emerging data curation roles for librarians: a case study of agricultural data.Journal of Agricultural & Food Information, 12(1), 65-74. doi: https://doi.org/10.1080/10496505.2011.539158

4. Chen, Xieling, Xie, Haoran, Wang, Fu Lee, Liu, Ziqing, Xu, Juan, and Hao, Tianyang.(2018). A bibliometric analysis of natural language processing in medical research. BMC Medical Informatics and Decision Making, 18(1), 1-14. doi: https://doi.org/10.1186/s12911-018-0594-x

5. Chen, Xieling, Xie, Haoran, Wang, Fu Lee, Liu, Ziqing, Xu, Juan, and Hao, Tianyong.(2018). A bibliometric analysis of natural language processing in medical research. BMC Medical Informatics and Decision Making, 18(1), 1-14. doi:https://doi.org/10.1186/s12911-018-0594-x

6. Kao, Anne, and Poteet, Steve R. (Ed.). Overview. Natural language processing and text mining, (pp. 1-8). USA: Springer. Springer: Germany, 2007. (1-264) p. Available at http://surl.li/cndrc

7. Kumar, Ankit, Ondruska, Peter, Iyyer, Mohit, Bradbury, James, Gulrajani, Ishaan, Zhong, Victor, Paulus, Romain, and Socher, Richard. Ask me anything: dynamic memory networks for natural language processing. (2016). Proceeding of the 33rd International Conference on Machine Learning. ICML, New York,USA, 20-22 June 2016, Proceedings of Machine Learning Research, 48, 1378-1387. Available at https://arxiv.org/pdf/1506.07285.pdf

8. Kumar, S. (2015). Co-authorship networks: a review of the literature. ASLIB: Journal of Information Management, 67(1), 55 – 73. doi: http://dx.doi.org/10.1108/AJIM-09-2014-0116

9. Lytvynenko, J. (2019). Identify the substantive, attribute and verb collocations in Russian text.Proceedings of the 3rd International Conference Computational Linguistics and Intelligent Systems, Ukraine, 2, 66-68. Retrieved from http://colins.in.ua/

10. Mishra, R., Bian, J., Fiszman, M., Weir, C. R., Jonnalagadda, S., Mustafa, J., and Fiol, G. D. (2014). Text summarization in the biomedical domain: a systematic review of recent research. Journal of Biomedical Informatics, 52, 475-467.doi: https://doi.org/10.1016/j.jbi.2014.06.009

11. Mustak, Mekhail, Salminem, Joni, Ple, Loic, and Wirtz, J. (2021).Artificial intelligence in marketing: bibliometric analysis, top modelling and research agenda. Journal of Business Research, 124, 389-404. doi: http://dx.doi.org/10.1016/j.jbusres.2020.10.044

12. Nandkarni, Prakash M., Ohno-Machado, Lucila, and Chapman, Wendy W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), 544-551. doi: https://doi.org/10.1136/amiajnl-2011-000464

13. Thirumuruganathan, Saravanan, Tang, Nan, Ouzzani, Mourad, and Doan, A. (2020).Data curation with deep learning.Proceedings of 23rd International Conference Database Technology (EDBT), Denmark, 277-286. doi: 10.5441/002/edbt.2020.25

14. Vijaykumar, S., and Sheshadri, Kn. (2019).Applications of artificial intelligence in academic libraries. International Journal of Computer Sciences and Engineering, 7(1), 136-140. doi: http://dx.doi.org/10.26438/ijcse/v7si16.136140

15. Weber, Nicholas, Palmer, Carole, and Chao, Tiffany.(2012). Current trends and future directions in data curation research and education.Journal of Web Librarianship, 6(4), 305-320. doi:https://doi.org/10.1080/19322909.2012.730358

16. Witt, Michael. (2022). Institutional repositories and research data curation in a distributed environment.Library Trend, 57(2), 191-201. doi:https://doi.org/10.1353/lib.0.0029

Keywords: Natural Language Processing; Data Curation; Bibliometrics; Scientific Collaboration

About Authors

Ms. Pooja Rana
Research Scholar
University of Jammu, Jammu, Jammu and Kashmir
Email: pr27.23.0000gmail.com

Dr. Pramod Kumar Singh
Sr. Assistant Professor
University of Jammu, Jammu, Jammu and Kashmir
Email: pksingh22@gmail.com