J Data Sci. Info. Citation Studies., 2023; 2(2):77-89.
https://www.jcitation.org

Research Article

A Bibliometric Study of Natural Language Processing Using Dimensions Database: Development, Research Trend, and Future Research Directions

Nasim Khadivi[1,*], **Sho Sato**[2]

[1]Artificial Intelligence Researcher, Department of Artificial Intelligence, Saman Industrial Group, IRAN.
[2]Associate Professor, Center for License and Qualification, Doshisha University, Kyoto, JAPAN.

ABSTRACT

A division of artificial intelligence known as natural language processing is thought to be a rapidly developing subject. NLP methods enable computers to comprehend human words and translate unstructured data for use by systems. The Dimensions database, which contains papers on natural language processing from 2002 to 2021, is thoroughly examined bibliometrically in this paper. It is the first attempt to perform bibliometric analysis in the NLP area in Dimensions database papers. The information is examined using bibliometric growth indicators, such as growth rate, relative growth rate, doubling time, degree of collaboration, and collaboration index. In addition, the VOSviewer software is used to create bibliometric networks. The primary purposes of this research study include detecting the NLP development pattern, discovering the distribution of co-authorship, influential authors, productive geographical distribution, keyword analysis, highly cited publications, finding recent research hotspots, and determining future research directions based on the challenges that exist in each of the research trends.

Keywords: Natural Language Processing, Bibliometric study, Dimensions database, VOSviewer.

Correspondence:
Ms. Nasim Khadivi
Artificial Intelligence Researcher,
Department of Artificial Intelligence,
Saman Industrial Group, IRAN.
Email: nasimkhadivi.uni@gmail.com
ORCID ID: 0000-0002-0517-3838

Received: 29-01-2023;
Revised: 11-04-2023;
Accepted: 15-06-2023.

INTRODUCTION

Natural language processing is a subfield of artificial intelligence that focuses on effective interaction between humans and computer systems in a natural language. It comprises algorithms and tasks that process and understand unstructured texts. Language processing is regarded as a remarkable part of human intelligence that separates humans from other creatures. Processing natural language is challenging due to the existence of ambiguity in the semantics and structure of human language (Reshamwala, Mishra, & Pawar, 2013). NLP has a long history and is still considered an evolving field of study (Hagiwara, 2021). Several applications of NLP have a notable role in our daily lives, including voice assistants (Hoy, 2018), spam filtering (Garg & Girdhar, 2021), search engines (Cafarella & Etzioni, 2005), and proofreading (L. Li, Song, Zhang, & Zhao, 2022).

A type of study methodology used in information science and libraries is called bibliometrics. It has been described as mathematical and statistical methods that are applied to data for monitoring and analyzing a scientific field's structure, research areas and trends, evaluation of research development, and regional and authorship patterns in publications and citations (Patra, Bhattacharya, & Verma, 2006; Thanuskodi, 2010).

Previous studies have analyzed the applications of NLP in diverse fields. For instance, Wang *et al.* (J. Wang *et al.*, 2020) conducted a bibliometric analysis on NLP data in the medicine field between 1999 and 2018, which is stored in the PubMed database. The authors thoroughly analyzed the collaboration relationships, author orders in publications, publication trends, countries and institutions of authors, and list of diseases that use NLP techniques for study. Chen *et al.* (X. Chen, Ding, *et al.*, 2018) conducted a bibliometric study on NLP in the empowered mobile computing field by using analysis methods such as affinity propagation clustering, latent Dirichlet allocation, social network analysis, geographic visualization, and descriptive statistics. Bibliographic data were retrieved from the Web of Science from 2000 to 2016 to discover geographical distribution, collaboration relationships, publication trends, productive journals, authors and affiliations, and topic distribution.

Martinez *et al.* (Lopez-Martinez & Sierra, 2020) published a bibliometric study on NLP by using Sci2 and the VOSviewer program, in addition to the Web of Science database selected for retrieving data from 2000 to 2019. The authors detected main journals, active counties and institutions, authors, most cited

OPEN ACCESS

EPUBL editing. publishing. technology

DOI: 10.5530/jcitation.2.2.11

Publishing Partner : EManuscript Tech. [www.emanuscript.in]

researchers, and main topics. Based on the PubMed database and by implementing R and Python software, Chen *et al.* (X. Chen, Xie, *et al.*, 2018) have carried out a bibliometric study on NLP medical articles published from 2007 to 2016. In total, 1776 documents in XML format were retrieved from a database. The research concentrates on analyzing publication distribution attributes, scientific collaboration, thematic discovery, and evolution patterns.

The Dimensions database's natural language processing publications were used in this research from 2002 to 2021, a 20-year span. The VOSviewer software analyzes the retrieved data with bibliometric growth indicators, such as growth rate, relative growth rate, doubling time, degree of collaboration, and collaboration index, in addition to creating bibliometric networks. The primary goal of this paper is to achieve the following goals in the field of natural language processing, starting with an examination of research output and a pattern of development that can be used to gauge the field's ups and downs. Then, because of their importance, it is carried out to identify the distribution of co-authorship and collaboration patterns, note the most significant writers, and identify geographic distribution. It's interesting to see how the advantages of study collaboration have increased in significance over time. Collaboration could boost output by allowing for the submission of more papers during a given time period. Although the value of collaborative articles varies depending on the number of authors, they do have the potential to raise publication acceptance rates. It is indicated that the co-authorship could intensify the quality (G. Durden & Gaynor, 1997; G. C. Durden & Perri, 1995). The synergistic benefit of working together stimulates compromise among researchers, and as a result the number of radical ideas reduces (Hudson, 1996). Finally, by analyzing the keywords and highly cited references, the research trends and future research opportunities are discovered.

The remaining study sections are arranged as follows: the data and methodology section develop research questions that are essential to creating a research study and discuss the bibliographic information, Dimensions database, and VOSviewer software. The data analysis and findings include information on the 20-year development pattern of NLP, co-authorship patterns, geographic distribution, references with a high citation count, keyword analysis, research trends, and future research directions. The bibliometric research is concluded in the final section.

METHODOLOGY

Research Questions

The study intends to find out about the development, authorship patterns, productive countries, research hotspots, and research opportunities for the future. To achieve the aims of the study some research questions have been set out.

Q1: How are the studies distributed over the last 20 years?

Answering Q1 would uncover the growth process over a specific timeline. All the ups and downs can be noticed, and important years in development are traced. Since the 2000s and 2010s, several developments have arisen that have led to a new evolution in NLP, such as a large increase in computing power, the accessibility of huge amounts of linguistic data, the advent of several deep learning algorithms, and a greater understanding of the structure of natural language. All of the mentioned factors reveal that the last 20 years have been an important period in NLP and require more investigation.

Q2: What is the most popular authorship pattern?

Answering Q2 would reveal the degree of collaboration in documents. Furthermore, we can find out whether researchers in this field prefer collaborative research work rather than working alone on a project.

Q3: What are the most productive and active countries in the Natural Language Processing field?

Answering Q3 would help researchers identify which countries have paid more attention to this domain and allocate more funds for research. This information would guide them in future collaboration or in seeking a position in a natural language processing-related field.

Q4: What are the research trends in the Natural Language Processing field?

Answering Q4 sheds light on the orientation of research in the field, also enables the prediction of future research directions.

Data Collection

In this study, bibliographic data is collected from the Dimensions database. Dimensions is a new partly free bibliographic database introduced by Digital Science in cooperation with over 100 outstanding research organizations in 2018. The Dimensions provides a large amount of documents—more than 124 million—and allows access to publications, clinical trials, patents, citations, grants, and policy documents. Moreover, it includes datasets from repositories such as PubMed, Figshare, Dryad, Zenodo, Pangaea, the NIH, and many more. It is deliberately launched to assist in contextualizing the research and evaluation environments (Hook, Porter, & Herzog, 2018). Dimensions is regarded as a reasonable alternative to Scopus and Web of Science for citation analyses, almost all Scopus DOIs have existed in Dimensions (Thelwall, 2018). This study analyzed the papers published from 2002 to 2021. By using the keyword ' natural language processing' 10091 records were retrieved from the database. During the study, the excluding task was carried out on some duplicates and a small number of fields that were devoid of authors' names or country names. After excluding these publications, the ultimate dataset

includes 9625 publications, which consist of 8530 articles, 1025 proceedings, and 70 books.

Data Analysis and Visualization

The right tools must be used in order to find appropriate solutions to the research questions that have been put forth. For creating, visualizing, and finding bibliometric maps, a free computer program called VOSviewer was created. By building citation networks between publications or journals, collaboration networks between researchers, and co-occurrence networks between scientific terms, VOSviewer can analyze various bibliometric networks. In addition, it is capable of representing large maps (N. Van Eck & Waltman, 2010). The bibliometric networks in this study are visualized using the VOSviewer program, and the relationships between the items are also found. Additionally, the analysis of bibliographic data is done using Microsoft Excel.

RESULTS AND DISCUSSION

Clarifying Natural Language Processing Development Pattern

Year Wise Distribution and Ratio of Growth

The annual distribution of books on natural language processing from 2002 to 2021 is shown in Table 1. 10091 papers are recovered from Dimensions; some duplicate papers are discovered during preprocessing. The final dataset, consisting of 9625 records, is created after duplicate entries are eliminated. Figure 1 shows the distribution of publications by year and their growth rate.

The graph in Figure 1 shows the annual change in the number of published papers over 20 years. According to the chart, three stages could be identified. During the first stage (2002-2005), the number of published papers increased gradually; it began with 308 papers in 2002 and reached 609 in 2005. In 2004, the growth rate (35.6%) was at its highest level in this stage. The second stage (2006-2017), the fluctuation stage, had great ups and downs, especially in 2017, when the literature number declined dramatically to 194 and the growth rate (-62.04%) was negative and at the lowest level over the past 20 years. In the third stage (2018-2021) the number of papers has significantly risen over the past 4 years; paper numbers reached a peak of 230-821 from 2018-2021. We see the highest growth rate (137.39%) in 2019 over the past 20 years. As stated in the third stage, we expect the increasing trend to continue in the future.

Relative Growth Rate and Doubling Time

Relative growth rate (RGR) and doubling time (DT) are the bibliometric indicators used to calculate the growth of publications. The relative growth rate is an increase in the number of papers or pages per unit of time. It also measures the specific period of the interval (Hunt, 1982). It can be calculated from the following equation:

$$RGR = \frac{W2 - W1}{T2 - T1}$$

Where,

RGR= Growth rate over the specific period of interval

W1= Natural log of the initial number of publications

W2= Natural log of the final number of publications

T1= The unit of the initial time

T2= The unit of the final time

Doubling time is defined as the time required for the number of publications to double its quantity (Mahapatra, 1985). A direct equivalence exists between relative growth rate and doubling time. If the number of articles or pages of subject doubles during the glen period as a result, the difference between the logarithms of numbers at the start and the end of the particular period must

Table 1: Year Wise distribution of literature in Natural Language Processing.

Year	Total Publications	TP(%)	Ratio of Growth
2002	308	3.20%	-
2003	368	3.82%	19.48%
2004	499	5.18%	35.6%
2005	609	6.33%	22.04%
2006	584	6.07%	-4.11%
2007	411	4.27%	-29.62%
2008	574	5.96%	39.66%
2009	374	3.89%	-34.84%
2010	512	5.32%	36.9%
2011	513	5.33%	0.20%
2012	361	3.75%	-29.63%

Year	Total Publications	TP(%)	Ratio of Growth
2013	452	4.70%	25.21%
2014	485	5.04%	7.3%
2015	625	6.49%	28.87%
2016	511	5.31%	-18.24%
2017	194	2.02%	-62.04%
2018	230	2.39%	18.56%
2019	546	5.67%	137.39%
2020	648	6.73%	18.68%
2021	821	8.53%	26.7%
Total	9625	100.00%	

be 2. If natural logarithms are used, the difference is 0.693 (Beaie & Acol, 2009). Doubling time for literature can be measured using the following formula:

$$\text{Doubling Time} = \frac{0.693}{RGR}$$

The relative growth rate and time for doubling the number of published books are shown in Table 2 and Figure 2. RGR and DT trends are indicated by small differences. From 2003 to 2021, RGR decreased from 0.79 to 0.09, as can be seen. The doubling time, on the other hand, has exhibited an increasing tendency; it began at 0.88 in 2003 and increased to 7.77 in 2021. In 2003, the RGR rating peaked at 0.79. However, the lowest number, 0.03, appeared in both 2017 and 2018. The DT values rose to their highest levels in 2017 and 2018 (26.01 and 22.58, respectively). The relationship between the relative growth rate and the doubling time is negative. The DT is at its highest value when the RGR is at its lowest rate because, according to the definition of doubling time, more time is required to raise the number of publications. As a result, it can be said that there was no increasing trend in NLP publications during the study period, and the volume of literature published varied greatly. However, starting in 2018, the relative growth rate began to increase, and it is possible to assume that the increasing trend will persist going forward.

Authorship Pattern Study.

Distribution of Authorship

Table 3 represents the authorship pattern of published papers. It is shown that three authored contributions take the first place and contribute 1769 (18.38%) productivity. The next place is recorded by two authored publications sharing 1652 (17.16%) of the total research contributions. Four-authored papers occupy 1468 (15.25%) publications and take the third place, followed by single-authored contributions with 1108 (11.51%) publications.

The lowest value is recorded by nine authored contributions with 255 (2.65%). According to the data, 40722 authors contributed to the 9625 pieces of literature during the period. The publications of ten authors or more are grouped as a single entity. It could be noted that multi-authored papers outnumbered single-authored papers. It indicates that researchers in the field of natural language processing prefer collaborative research to a solo project. Both mean and median are estimators for the central value, and asymmetrical distributions can cause big differences between average and median. Also, based on the standard deviations of all authors, which are calculated at 2.52, it can be interpreted that the range and spread of the number of authors and their published papers are wide, and the data are not clustered around the average, which means that the number of published papers by each group of authors varied considerably.

Collaboration Patterns

The degree of collaborations, the index of collaborations, and the pattern of writing by year are shown in Table 4. The ratio of collaborative research publications to all papers released in the discipline over a specific time span is used to define the degree of collaboration. It serves as a gauge for the degree of interdisciplinary cooperation (Subramanyam, 1983). The following formula is used to determine the degree of collaboration:

$$DC = \frac{Nm}{Nm+Ns}$$

Where,

DC= is the degree of collaboration in the discipline

Nm = the number of multi authored papers in the discipline published during a year

Ns = the number of single authored papers in the discipline published during a year

The collaboration index is defined as the mean number of authors per paper. Collaboration index is calculated by the following equation (Savanur & Srikanth, 2010).

$$CI = \frac{\sum_{j}^{A}=1 \; j \; fj}{N}$$

Where,

Table 2: **Relative growth rate and doubling time of publications.**

Year	Total Publications	Cumulative Publications	W1	W2	RGR	DT
2002	308	308	0.00	5.73	0.00	0
2003	368	676	5.73	6.52	0.79	0.88
2004	499	1175	6.52	7.07	0.55	1.25
2005	609	1784	7.07	7.49	0.42	1.66
2006	584	2368	7.49	7.77	0.28	2.45
2007	411	2779	7.77	7.93	0.16	4.33
2008	574	3353	7.93	8.12	0.19	3.69
2009	374	3727	8.12	8.22	0.11	6.55
2010	512	4239	8.22	8.35	0.13	5.38
2011	513	4752	8.35	8.47	0.11	6.07
2012	361	5113	8.47	8.54	0.07	9.46
2013	452	5565	8.54	8.62	0.08	8.18
2014	485	6050	8.62	8.71	0.08	8.29
2015	625	6675	8.71	8.81	0.10	7.05
2016	511	7186	8.81	8.88	0.07	9.39
2017	194	7380	8.88	8.91	0.03	26.01
2018	230	7610	8.91	8.94	0.03	22.58
2019	546	8156	8.94	9.01	0.07	10.00
2020	648	8804	9.01	9.08	0.08	9.06
2021	821	9625	9.08	9.17	0.09	7.77

Table 3: **Authorship Pattern of Publications.**

Number of Authors	Total Publications	Total Publications (%)	Total Number of Authors
1	1108	11.51%	1108
2	1652	17.16%	3304
3	1769	18.38%	5307
4	1468	15.25%	5872
5	1100	11.43%	5500
6	761	7.91%	4566
7	542	5.63%	3794
8	362	3.76%	2896
9	255	2.65%	2295
≥10	608	6.32%	6080
Total	9625	100%	40722

Mean number of Authors: 4. 23 Median number of Authors: 3 Mode number of Authors: 3 Standard Deviation number of Authors: 2.52

CI = the mean number of authors per paper

A= the total number of authors per publications

j = the number of authors in publications

fj = the number of j authored publications

N= the total number of published literature in a year

The degree of partnership, as shown in Table 4, varied between 0.71 and 0.93. With 0.96, 2018 and 2020 have the greatest DC. Despite variations in collaboration levels, the field of natural language processing has seen an increase in joint work over the past few years. The collaboration index increased over the course of the research; it peaked in 2021 at 5.37 after starting at 2.77

Table 4: Yearly distribution of authorship pattern, degree of collaboration (DC), and collaboration index (CI).

Year	Number of Authors										TP	DC	CI
	1	2	3	4	5	6	7	8	9	10			
2002	88	80	69	28	17	8	6	4	1	7	308	0.71	2.77
2003	79	98	69	60	23	15	7	5	3	9	368	0.79	3.08
2004	81	116	125	68	54	21	11	7	6	10	499	0.84	3.29
2005	84	142	144	125	52	19	16	7	7	13	609	0.86	3.34
2006	118	116	125	80	59	34	20	15	7	10	584	0.80	3.37
2007	51	75	85	67	64	27	13	8	4	17	411	0.88	3.81
2008	51	110	129	96	70	46	22	15	10	25	574	0.91	3.98
2009	48	78	88	65	37	21	13	9	9	6	374	0.87	3.59
2010	83	89	98	62	59	37	39	20	3	22	512	0.84	3.91
2011	64	90	103	67	58	47	31	21	5	27	513	0.88	4.08
2012	37	74	63	45	48	32	20	14	10	18	361	0.90	4.18
2013	34	75	88	73	60	35	34	15	9	29	452	0.92	4.38
2014	66	61	77	83	48	39	38	21	21	31	485	0.86	4.45
2015	41	102	110	114	76	45	33	30	24	50	625	0.93	4.59
2016	41	93	88	82	60	50	26	18	12	41	511	0.92	4.43
2017	13	24	24	32	24	22	15	12	11	17	194	0.93	5.07
2018	10	28	35	34	27	31	10	14	10	31	230	0.96	5.26
2019	34	44	63	80	74	70	53	32	25	71	546	0.94	5.46
2020	29	76	91	88	80	76	59	40	39	70	648	0.96	5.32
2021	56	81	95	119	110	86	76	55	39	104	821	0.93	5.37
Total	1108	1652	1769	1468	1100	761	542	362	255	608	9625	-	-

in 2002. As was previously stated, CI is the typical number of authors per paper, and based on the data in the Table 4, more collaborations are anticipated in the future.

The Most Influential Authors

According to Table 5, which shows the most influential authors in 20 years, Friedman Carol from Columbia University has achieved the first rank with 2040 citations. The second place belonged to Liu Hongfang from Mayo College of Medicine with 1624 citations, and Hripcsak George placed in the third rank with 1598 citations. It is noticed that the research interests of top authors, apart from natural language processing, also focus on text mining, data mining, biomedical informatics, and artificial intelligence. The higher the citation number of the publications, the more influential the authors are considered.

Geographical Distribution of Collaboration

A co-authorship network is a social network that shows the collaboration of authors, countries, or organizations in research (Melin & Persson, 1996). The co-authorship pattern is an explicit statement of collaboration, and it is used to analyze scientific collaboration patterns. The nodes are the signs of authors, countries, or organizations that are connected by edges. The edge is represented between nodes. If the countries have cooperation in research, in social networks the thickness of the edge can be defined as the strength value of the link; the thicker the link, the larger the number of collaborations(Moosa & Shareefa, 2020; N. J. Van Eck & Waltman, 2013).

Figure 3 has demonstrated the network visualization of the 22 most productive countries in a field of natural language processing; the minimum number of documents per country is set to 50. The size of each node is represented by the number of papers conducted in that country. In addition, the distance between the nodes reveals the relevance of their collaboration. The United States is the most active node in this network; it has conducted the largest collaborative research with the United Kingdom, with a link strength of 124. The second country that has carried out a large amount of cooperation with the United States is China, with a link strength of 115. However, Israel, Brazil, and India are located in a faraway spot from the United States, which shows that their co-authorship and collaboration are less related.

The second-most active nation, the United Kingdom, shows significant collaborative ties with Germany and Japan, with link strengths of 42 and 40, respectively. Greece and Brazil are additionally regarded as the least productive nations, generating 53 and 56 documents, respectively.

Table 5: Ten most influential authors during study period.

Rank	Authors	Affiliation	Research Interest	Citations	Number of Publications
1	Friedman, Carol	Columbia University	Natural language processing- data mining-Pharmacovigilance	2040	87
2	Liu, Hongfang	Mayo College of Medicine	Natural language processing- BioNLP-Clinical informatics	1624	123
3	Hripcsak, George	Columbia University	Biomedical informatics	1598	54
4	Xu, Hua	University of Texas, UTHealth School of Biomedical Informatics	Natural language processing- Text mining	1454	111
5	Chute, Christopher g	Johns Hopkins University	Medical concept representation	1166	61
6	Sohn, Sunghwan	Mayo College of Medicine	Natural language processing-Text analytics-medical informatics	1125	50
7	Lu, Zhiyong	National Center for Biotechnology Information	BioNLP- Biomedical informatics- Artificial intelligence	1099	39
8	Uzuner, Ozlem	George Mason University	Natural language processing- Artificial intelligence- Medical informatics	1080	23
9	Chapman, Wendy	University of Melbourne	Natural language processing- Biomedical informatics- Digital transformation of health	1066	61
10	Denny, Joshua c	National Institutes of Health	Natural language processing- EHR data mining- Biomedical informatics	1018	75

As demonstrated in Figure 3, the nodes are clustered into different groups with distinct colors. Nodes are clustered based on several related co-occurrence terms that are shared by each country. It can be said that the United States, which is a member of the green cluster, is closer to a member of its own cluster in terms of the subject of the research it has carried out. There are 5 clusters in this network, which are listed as follows: green cluster (8 items), red cluster (10 items), dark blue cluster (2 items), yellow cluster (1 item), and purple cluster (1 item).

According to Table 6, which illustrates the top 10 productive countries that dominate the natural language processing field, the United States ranked first and produced 4557 publications during the period of study, sharing 62.18% of citations, which is the maximum value of citations of the total contributions. The second-most productive country is the United Kingdom, with 678 research papers and 10.15% of citations. The third place belonged to China, which produced 579 works of literature and allocated 4.53% of citations. The significant note is that although some countries produce fewer research publications, they receive more citations. As can be seen from the Table 6, Germany got the fourth rank and conducted 358 publications, sharing 6.98% of citations, which is higher than China's citations. Moreover, Canada, which is in seventh place, receives 6.35% more citations than France and Japan.

Analysis of Research Trends and Research Opportunities

Highly Cited References

Table 7 depicts the ten most-cited publications on natural language processing in the last five years. Analyzing the latest frequently cited papers can lead to finding what scopes are attracting more attention and what the main investigation opportunities in the field are. According to the table data, the first place belongs to the paper titled "A guide to deep learning in healthcare," authored by

Table 6: **Top 10 productive countries.**

Rank	Country	Number of Publications	Citations (%)
1	United States	4557	62.18%
2	United Kingdom	678	10.15%
3	China	579	4.53%
4	Germany	358	6.98%
5	Japan	346	2.49%
6	France	304	1.95%
7	Canada	287	6.35%
8	Australia	214	2.06%
9	Spain	209	2.21%
10	South Korea	173	1.11%

Esteva A. *et al.*, published in "Nature Medicine" and getting 826 citations. The main focus of the paper is to investigate how deep learning techniques can be used in natural language processing and computer vision to be effective in the healthcare system. The second-most cited publication is titled 'BIOBERT: a pre-trained biomedical language representation model for biomedical text mining' authored by LEE J. *et al.*, published in 'Bioinformatics Oxford Academic' and has been cited by 711 researchers. BIOBERT is introduced as a bidirectional encoder representation from transformers for biomedical text mining. It can assist biomedical corpora in recognizing complicated biomedical texts.

Figure 4 depicts the paper on natural language processing that is commonly cited. The sources with 60 or more links have been gathered. 41 documents altogether satisfy the requirement. Some papers that are not connected to other nodes are still included. Furthermore, some of the label nodes are unavailable due to duplication. The visualization shows links between several of the widely cited references, indicating some relationship between their research outputs. The citation network is split into 25 clusters, 17 of which contained only one item, and the remaining clusters contained up to five items.

Figure 1: The total publication and growth rate by year.

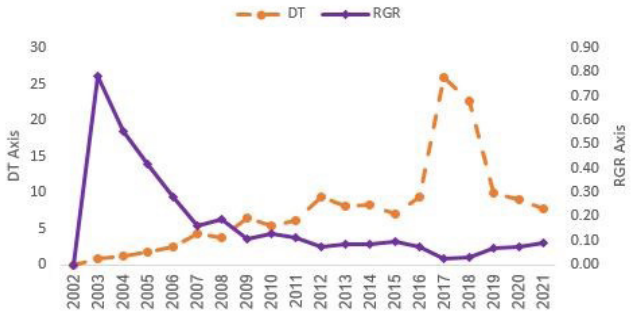

Figure 2: Relative growth rate and doubling time of publications.

Keywords Analysis

Identifying the top keywords can lead to finding hotspots. The top 25 frequently used title keywords, with frequency and relevance, are shown in Table 1. A full-count method is selected for analyzing keywords. Terms that appear at least ten times are taken into account. 370 terms meet the threshold out of 18917. For each of the 370 terms, the relevance score is calculated by VOSviewer, and the most relevant terms are selected. As can be seen from Table 1, the keywords are mainly divided into two groups: 1) Natural language processing methods and tasks 2) Application of natural language processing in the healthcare system.

1)Natural language processing methods and tasks: natural language processing is a subfield of artificial intelligence so the keyword 'artificial intelligence' is included here. 'Entity recognition', ' disambiguation', 'relation extraction', 'similarity', 'semantic', 'normalization', 'text classification', 'machine translation', 'sentiment analysis' are referred to natural language processing methods and tasks keywords which are highly used in publications during the study period.

2) Application of natural language processing in the healthcare system: For instance, natural language processing is applied to 'electronic medical records ' or 'biomedical literature' to discover diseases that might not have been coded previously or uncover computable information. NLP techniques can be used to combat "radiology," "adverse drug events," and even the "COVID-19" pandemic.

Table 7: **Top 10 highly cited refe rences in the last five years.**

Rank	Author(Year)	Journal	Article DOI	Main Idea	Citations
1	Esteva,a, *et al* (2019)	Nature Medicine	10.1038/s41591-018-0316-z	Application of deep learning techniques are presented in natural language processing, computer vision and reinforcement learning for healthcare.	826
2	Lee j, *et al* (2019)	Bioinformatics Oxford Academic	10.1093/bioinformatics/btz682	The BioBERT is introduced which is a domain specific language representation model, and it has facilitated understanding of complex biomedical texts.	711
3	Wang y, *et al* (2017)	Journal of Biomedical Informatics ScienceDirect	10.1016/j.jbi.2017.11.011	The literature review has been conducted on clinical information extraction applications.	281
4	Hoy,m (2018)	Medical Reference Services Quarterly	10.1080/02763869.2018.1404391	The voice assistants and their usage are discussed, also the recent advances in natural language processing which help voice assistants to produce meaningful responses are presented.	259
5	Hashimoto d, *et al* (2018)	Annals of Surgery	10.1097/SLA.0000000000002693	The application of main subfields of AI such as natural language processing in surgical practice are discussed.	257
6	Miller a, *et al* (2017)	American Journal of Medicine ScienceDirect	10.1016/j.amjmed.2017.10.035	The study is focus on the place of the artificial intelligence fields such as natural language processing in medical practice.	245
7	Schmidt a,*et al* (2017)	Fifth International Workshop on NLP for Social Media	10.18653/v1/W17-1101	The study represents a survey on hate speech detection using natural language processing and key techniques which is used to detect utterances.	217
8	Tasnim s, *et al* (2020)	Journal of Preventive Medicine and Public Health	10.3961/jpmph.20.094	The effect of misinformation and fake news on COVID-19 in social media are studied. Also, the importance of the Natural language processing and text mining approaches for eliminating the rumors and fake content are discussed.	198
9	Kreimeyer k, *et al* (2017)	Journal of Biomedical Informatics ScienceDirect	10.1016/j.jbi.2017.07.012	The systematic approach is conducted in order to recognize available clinical NLP systems that produce structured information from unstructured texts.	193
10	Soysal e, *et al* (2017)	Journal of the American Medical Informatics Association (JAMIA)	10.1093/jamia/ocx132	A clinical natural language processing toolkit is developed which helps users to create customized NLP pipelines.	155

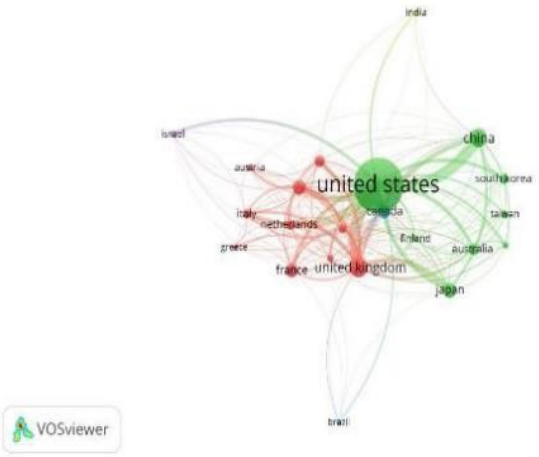

Figure 3: Visualization of country co-authorship network.

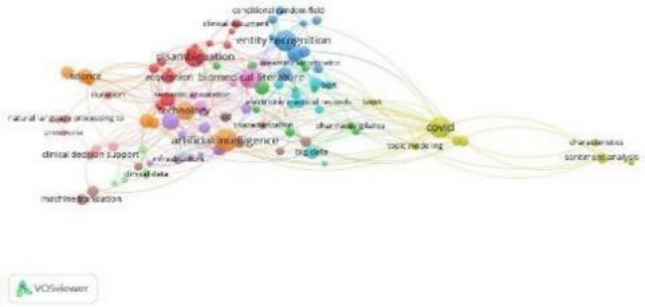

Figure 5: Visualization of co-occurrences of keywords.

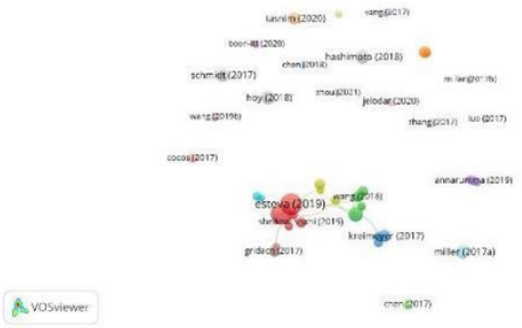

Figure 4: Visualization of publications citation network in the last five years.

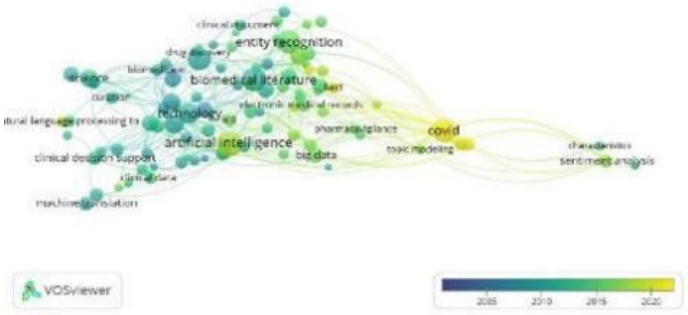

Figure 6: Overlay Visualization of keywords during study period.

In Figure 5, a co-occurrences map, terms were grouped into 11 categories. Disambiguation (for the red cluster), similarity (for the green cluster), entity recognition (for the dark blue cluster), COVID (for the yellow cluster), integration (for the purple cluster), big data (for the light blue cluster), artificial intelligence (for the orange cluster), machine translation (for the brown cluster), biomedical literature (for the pink cluster), clinical decision support (for the light pink cluster), and clinical data are the primary keywords for each cluster (light green cluster). Certain constellation words are related to belonging to particular thematic groups. The placement of a keyword within a cluster provides information about its prevalence and relationships with other words.

Figure 6 illustrates an overlay visualization of keywords over the past 20 years, which can help to notice the most frequently used and appearing keywords. The color of the nodes illustrates the estimated publication year, which begins from dark blue to yellow. Some of the node labels are not available due to overlapping. The keywords that were detected in the last five years' timeline are listed as follows: (1) BERT, (2) convolutional neural network, (3) social medium, (4) tweet, (5) big data, (6) transformer, (7) artificial intelligence, (8) clinical natural language processing, (9) Covid, and (10) sentiment analysis.

Discovering Research Trends and Future Direction

Based on the two previous sections, highly cited references, keyword analysis, research trends, and hot topics can be extracted from the bibliographic data. The last 5 years' data is used for analyzing frequently cited publications; it represents what fields motivate the researcher's curiosity; and it can assist in uncovering the more outstanding and cutting-edge fields in natural language processing. The identified research trends are described as follows: (1) The emergence of deep learning has an important and beneficial impact and enhances the efficiency and accuracy of artificial intelligence subfields, including natural language processing. Deep learning methods have gained more attention, and their performance surpasses that of traditional machine learning methods (Zhang, Schoene, Ji, & Ananiadou, 2022). Machine learning and deep learning techniques help computer systems solve unclear problems by learning from sample data, gaining new insights, making its own decisions, and initiating new intelligence, while improving the responses by using past experiences. Besides, the plentiful availability of data makes machine learning and deep learning popular (Parsons, 2010). Machine learning and deep learning techniques have played a significant role in the success of natural language processing, enhancing its performance and precision in comparison with traditional techniques(Nagarhalli, Vaze, & Rana, 2021). These

dissimilar techniques are used in crucial tasks of natural language processing, including word sense disambiguation, sarcasm detection, syntactic parsing, stemming, part of speech tagging, named entity recognition, and reference resolution(Chiche & Yitagesu, 2022; Jaf & Calder, 2019; P. Kumar, 2020; J. Li, Sun, Han, & Li, 2020; Razali, Halin, Ye, Doraisamy, & Norowi, 2021).

(2) Natural language processing techniques can be used for automatic hate speech detection and identifying fake news. Social media platforms provide online spaces where users can share their opinions, have a conversation, and create content. Consequently, on these online platforms, users can use offensive and hateful speech or disseminate fake news. At this juncture, hate speech is recognized as a major problem by many countries; this is because the obscurity of online social media gives some users the courage to participate in harmful activities (Kumar, Yadav, & Namdeo). Adding barriers to the internet to stop the spread of offensive or fake material. Additionally, online social media platforms encourage the spread of fake news, which can be done for a variety of reasons, including to harm organizations or people, sway people's views, sow discord, or make money (de Oliveira, Pisa, Lopez, de Medeiros, & Mattos, 2021).

(3) The applications of natural language processing, especially ML-based NLP, have gained increasing attention and can be considered a new paradigm in surgical, medical, and mental health practice in recent years. With the evolution of electronic health records in 2015 and the growing amount of information on patients that is kept in the EHR systems, which is the form of free text documents or unstructured data (Evans, 2016), there is a need for the utilization of natural language processing techniques to read the free text data and find hidden information that is useful for clinical decision support, finding the best treatment for patients, diagnosing conditions, and evolving treatment plans. NLP techniques are used to analyze clinical free text to derive meaning and discover concepts in some of the surgical and medical fields in which NLP techniques are implemented, including identification of surgical site infection, prediction of suicide risk, risk prediction of inpatient forensic psychiatric settings, the impact of COVID-19 on mental health, identification of pneumonia, discovering adverse drug events, extracting cancer's phenotypes and identifying cancers, analyzing radiology reports, and many more (Chen *et al*., 2020; Cook *et al*., 2016; Jones *et al*., 2018; Luo & Chong, 2020; Savova *et al*., 2019; Sengupta, Mugde, & Sharma, 2020; Thirukumaran *et al*., 2019; Van Le, Montgomery, Kirkby, & Scanlan, 2018).

(4) Voice assistants are software agents that receive and interpret human language, then respond reasonably based on the received data. These software agents unceasingly prepare for receiving keywords, and the user's voice is recorded and sent to the server, which, after interpretation, sends proper information to the voice assistant to complete the task of the user. Natural language processing is utilized in voice assistants to match the user's voice, which is used as an input to an executable command (Kadali, Prasad, Kudav, & Deshpande, 2020). Voice assistants have been implemented in different fields such as banking and finance, remote healthcare delivery services, education, improving the social skills of patients, and many more (Safi, Al Sadrani, & Mustafa, 2021; Sezgin, Huang, Ramtekkar, & Lin, 2020; Terzopoulos & Satratzemi, 2020).

(5) The presentation of transfer learning and pre-trained language models has reduced the limits of natural language understanding, and they have stood out as one of the most substantial research trends. Nowadays, the BERT (pre-training of deep bidirectional transformers for language understanding) model is one of the most popular language models based on deep learning, which is used in almost every word processing task and provides incomparable precision results (Koroteev, 2021). Several of the language models are listed as follows: BioBERT, GPT, RoBERTa, XLNet, ELECTRA, BART, PEGASUS, and T5 (Kalyan, Rajasekharan, & Sangeetha, 2021).

Every new trend has its challenges, and since natural language processing tasks are fundamental in nature, this fact can be used to carefully suggest the direction of future research. There are a few areas of inquiry that will need further study in the future: (1) It is difficult to mine the medical text data stored in EHR or social media platforms because of imbalanced data, misspellings or ambiguity, high lexical variability, the difficulty of de-identification, the lack of key annotation, and the difficulty of obtaining all the semantic and syntactic attributes from complex sentences due to the lack of efficient methods that can capture all of them (Li *et al*., 2019; Pandey, Pandey, Mishra, & Rhmann, 2021; Sarker, Gonzalez-Hernandez, & Perrone, 2019; X. Wang, Hripcsak, Markatou, & Friedman, 2009).

(2) Although voice assistants are becoming more prevalent, the issue of their security and privacy is still difficult to resolve. Similar to other computing devices, voice assistants are vulnerable to malevolent attacks, such as dolphin attacks, man-in-the-middle attacks, and skill squatting attacks, which rely on flaws in the original design and software of voice assistants. Additionally, sometimes voice assistants (VAs) record users' voices without their consent, and even when they are conscious of it, the VA vendor records a ton of personal information. Therefore, it is difficult to authenticate users without requiring an external device (Bolton, Dargahi, Belguith, Al-Rakhami, & Sodhro, 2021; Kumar *et al*., 2018; Turner, Lovisotto, & Martinovic, 2019).

(3) Pre-training language models are an open research area that guide the acquisition of striking gains on several natural language processing tasks. Some of the language representation models such as BERT (Devlin, Chang, Lee, & Toutanova, 2018) and ELMO have been trained generally for general domain datasets, and they perform weakly on domain-specific text mining tasks (Lee *et al*., 2020), thereby leading to the development of domain-specific

language models like Bio-BERT and ELECTRAMed (Miolo, Mantoan, & Orsenigo, 2021), which are applicable for biomedical text mining, or BERT_SE (Fávero & Casanova, 2021), which is used for noticing software engineering vocabulary, which was initiated a few years ago. Additionally, in recent years, BERT has been used for fake news detection. In the future, an investigation might be continued into evolving the existing language models for domain-specific tasks and enhancing mining performance. (4) Diffusion of harmful content may damage society, specifically people who belong to minority social groups. The process of automatically identifying hateful content has faced several challenges and requires more investigation. These challenges are listed as follows: the usage of sarcasm is common; harmful content might be grammatically accurate and fluent; the evolution of language specifically among young groups takes place rapidly; and discerning emoji-based hate is burdensome.

CONCLUSION

In this research, we presented a meticulous bibliometric analysis of the natural language processing field. A total of 9625 bibliographic records were retrieved from the Dimensions database during the years 2002-2021. Indeed, it is the first attempt on the Dimensions database to conduct a bibliometric analysis on the NLP field of study. According to the findings, 308 papers were reported in 2002 at the start of the study period, and 821 papers were published at the end of it. Significant growth can be seen during the study's final years, and the doubling times for publications over the past four years have decreased, indicating a steady growth of publications in the NLP field. The field of natural language processing (NLP), which is still developing, has the potential to expand in the future. According to analysis, there are several hotspots in the NLP domain, including deep learning-based NLP projects, automatic hate speech detection, automatic fake news detection, ML-based NLP in surgical, medical, and mental health practice, voice assistants, and pre-trained language models. The existing challenges in the hotspots mentioned above open up new prospects for future research projects. Overall, this paper contributed to the following research findings: (1) The degree of collaboration and the collaboration index gradually grew during the last few years, which revealed more collaborations in NLP research. (2) Friedman Carol from Columbia University is the most influential researcher, achieving 2040 citations. (3) In the co-authorship network, the United States identifies as the most productive country, and the United Kingdom and China obtain second and third place, respectively.

ACKNOWLEDGEMENT

This paper has written using bibliographic data from Dimensions.ai database.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Beaie S, Acol P. Population and demographic measures: Concepts and definitions for basic MDG indicators. Kingston Georgetown, Guyana: Bureau of Statistics. 2009.
Bolton T, Dargahi T, Belguith S, Al-Rakhami MS, Sodhro AH. On the security and privacy challenges of virtual assistants. Sensors. 2021;21(7):2312.
Cafarella MJ, Etzioni O. A search engine for natural language applications. Paper presented at the Proceedings of the 14th international conference on World Wide Web. 2005.
Chen L, Gu Y, Ji X, Sun Z, Li H, Gao Y. Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning. Journal of the American Medical Informatics Association, 2020;27(1):56-64.
Chen X, Ding R, Xu K, Wang S, Hao T, Zhou Y. A bibliometric review of natural language processing empowered mobile computing. Wireless Communications and Mobile Computing. 2018.
Chen X, Xie H, Wang FL, Liu Z, Xu J, Hao T. A bibliometric analysis of natural language processing in medical research. BMC medical informatics and decision making. 2018;18(1):1-14.
Chiche A, Yitagesu B. Part of speech tagging: a systematic review of deep learning and machine learning approaches. Journal of Big Data. 2022;9(1):1-25.
Cook BL, Progovac AM, Chen P, Mullin B, Hou S, Baca-Garcia E. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. Computational and mathematical methods in medicine. 2016.
de Oliveira NR, Pisa PS, Lopez MA, de Medeiros DSV, Mattos DM. Identifying fake news on social networks based on natural language processing: trends and challenges. Information. 2021;12(1):38.
Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. arXiv preprint arXiv:1810.04805.
Durden G, Gaynor P. Multi-authored papers in economics. Journal of Economic Perspectives. 1997;11(2):192-3.
Durden GC, Perri TJ. Coauthorship and publication efficiency. Atlantic economic journal. 1995;23(1):69-76.
Evans RS. Electronic health records: then, now, and in the future. Yearbook of medical informatics. 2016;25(S1):S48-61.
Fávero EMDB, Casanova D. BERT_SE: A Pre-trained Language Representation Model for Software Engineering. 2021. arXiv preprint arXiv:2112.00699.
Garg P, Girdhar N. A Systematic Review on Spam Filtering Techniques based on Natural Language Processing Framework. Paper presented at the 2021 11th International Conference on Cloud Computing, Data Science and Engineering (Confluence). 2021.
1Hagiwara M. Real-World Natural Language Processing: Practical Applications with Deep Learning: Simon and Schuster. 2021.
Hook DW, Porter SJ, Herzog C. Dimensions: building context for search and evaluation. Frontiers in Research Metrics and Analytics. 2018;3:23.
Hoy MB. Alexa, Siri, Cortana, and more: an introduction to voice assistants. Medical reference services quarterly. 2018;37(1):81-8.
Hudson J. Trends in multi-authored papers in economics. Journal of Economic Perspectives. 1996;10(3):153-8.
Hunt R. Plant growth analysis. London, Edward Arnold. 1978:67. Studies in.
Hunt R. Plant growth analysis: second derivatives and compounded second derivatives of splined plant growth curves. Annals of Botany. 1982;50(3):317-28.
Jaf S, Calder C. Deep learning for natural language parsing. IEEE Access. 2019;7:131363-73.
Jones B, South BR, Shao Y, Lu C, Leng J, Sauer BC, et al. Development and validation of a natural language processing tool to identify patients treated for pneumonia across VA emergency departments. Applied clinical informatics. 2018;9(1):122-8.
Kadali B, Prasad N, Kudav P, Deshpande M. Home Automation Using Chatbot and Voice Assistant. Paper presented at the ITM Web of Conferences. 2020.
Kalyan KS, Rajasekharan A, Sangeetha S. Ammus: A survey of transformer-based pretrained models in natural language processing. 2021. arXiv preprint arXiv:2108.05542.
Koroteev M. BERT: A review of applications in natural language processing and understanding. 2021. arXiv preprint arXiv:2103.11943.
Kumar C, Yadav RK, Namdeo V. A Review on Hate Speech Recognition on Social Media. 2021.
Kumar D, Paccagnella R, Murley P, Hennenfent E, Mason J, Bates A, Bailey M. Skill squatting attacks on Amazon Alexa. Paper presented at the 27th USENIX security symposium (USENIX Security 18). 2018.
Kumar P. Word sense disambiguation for Punjabi language using deep learning techniques. Neural Computing and Applications. 2020;32(8):2963-73.
Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40.
Li J, Sun A, Han J, Li C. A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering. 2020;34(1):50-70.
Li L, Song P, Zhang D, Zhao D. A Hybrid Model for Chinese Confusable Words Distinguishing in Proofreading. Paper presented at the Workshop on Chinese Lexical Semantics. 2022.

Li Z, Yang Z, Shen C, Xu J, Zhang Y, Xu H. Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text. BMC medical informatics and decision making. 2019;19(1):1-8.
Lopez-Martinez RE, Sierra G. Natural language processing, 2000-2019—a bibliometric study. Journal of Scientometric Research. 2020;9(3):310-8.
Luo JW, Chong JJ. Review of natural language processing in radiology. Neuroimaging Clinics. 2020;30(4):447-58.
Mahapatra M. On the validity of the theory of exponential growth of scientific literature. Paper presented at the Proceedings of the 15th IASLIC conference, Bangalore. 1985.
Melin G, Persson O. Studying research collaboration using co-authorships. Scientometrics. 1996;36(3):363-77.
Miolo G, Mantoan G, Orsenigo C. Electramed: a new pre-trained language representation model for biomedical nlp. 2021. arXiv preprint arXiv:2104.09585.
Moosa V, Shareefa M. Science mapping the most-cited publications on workplace learning. Journal of Workplace Learning. 2020.
Nagarhalli TP, Vaze V, Rana N. Impact of machine learning in natural language processing: A review. Paper presented at the 2021 third international conference on intelligent communication technologies and virtual mobile networks (ICICV). 2021.
Pandey B, Pandey DK, Mishra BP, Rhmann W. A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions. Journal of King Saud University-Computer and Information Sciences. 2021.
Parsons S. Introduction to Machine Learning, Second Editon by Ethem Alpaydin, MIT Press, 584 pp., $55.00. ISBN 978-0-262-01243-0. The Knowledge Engineering Review. 2010;25(3):353.
Patra SK, Bhattacharya P, Verma N. Bibliometric study of literature on bibliometrics. DESIDOC Bulletin of Information Technology. 2006;26(1):27-32.
Razali MS, Halin AA, Ye L, Doraisamy S, Norowi NM. Sarcasm Detection Using Deep Learning with Contextual Features. IEEE Access. 2021;9:68609-18.
Reshamwala A, Mishra D, Pawar P. Review on natural language processing. IRACST Engineering Science and Technology: An International Journal (ESTIJ). 2013;3(1):113-6.
Safi MF, Al Sadrani B, Mustafa A. Virtual voice assistant applications improved expressive verbal abilities and social interactions in children with autism spectrum disorder: a Single-Subject experimental study. International Journal of Developmental Disabilities. 2021:1-13.
Sarker A, Gonzalez-Hernandez G, Perrone J. Towards automating location-specific opioid toxicosurveillance from Twitter via data science methods. Studies in health technology and informatics. 2019;264:333.
Savanur K, Srikanth R. Modified collaborative coefficient: A new measure for quantifying the degree of research collaboration. Scientometrics. 2010;84(2):365-71.
Savova GK, Danciu I, Alamudun F, Miller T, Lin C, Bitterman DS, et al. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer research. 2019;79(21):5463-70.
Sengupta S, Mugde S, Sharma G. An Exploration of Impact of COVID 19 on mental health-Analysis of tweets using Natural Language Processing techniques. 2020. medRxiv.
Sezgin E, Huang Y, Ramtekkar U, Lin S. Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic. NPJ digital medicine. 2020;3(1):1-4.
Subramanyam K. Bibliometric studies of research collaboration: A review. Journal of information Science. 1983;6(1):33-8.
Terzopoulos G, Satratzemi M. Voice assistants and smart speakers in everyday life and in education. Informatics in Education. 2020;19(3):473-90.
Thanuskodi S. Journal of Social Sciences: A bibliometric study. Journal of Social Sciences. 2010;24(2):77-80.
Thelwall M. Dimensions: A competitor to Scopus and the Web of Science? Journal of informetrics. 2018;12(2):430-5.
Thirukumaran CP, Zaman A, Rubery PT, Calabria C, Li Y, Ricciardi BF, et al. Natural language processing for the identification of surgical site infections in orthopaedics. The Journal of Bone and Joint Surgery. American Volume. 2019;101(24):2167.
Turner H, Lovisotto G, Martinovic I. Attacking speaker recognition systems with phoneme morphing. Paper presented at the European Symposium on Research in Computer Security. 2019.
Van EN, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. 2010;84(2):523-38.
Van ENJ, Waltman L. VOSviewer manual. Leiden: Univeristeit Leiden. 2013;1(1):1-53.
Van LD, Montgomery J, Kirkby KC, Scanlan J. Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting. Journal of biomedical informatics. 2018;86:49-58.
Wang J, Deng H, Liu B, Hu A, Liang J, Fan L, et al. Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: bibliometric study on PubMed. Journal of medical internet research. 2020;22(1):e16816.
Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. Journal of the American Medical Informatics Association. 2009;16(3):328-37.
Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. NPJ digital medicine. 2022;5(1):1-13.

Cite this article: Khadivi N, Sato S. A Bibliometric Study of Natural Language Processing using Dimensions Database: Development, Research Trend, and Future Research Directions. Journal of Data Science, Informetrics, and Citation Studies. 2023;2(2):77-89.