COLLNET Journal of Scientometrics and Information Management
ISSN : 0973 – 7766 (Print) 2168 – 930X (Online)
Vol. 17(1) June 2023, pp. 61–79
DOI : 10.47974/CJSIM-2022-0023

Research trends in the field of natural language processing : A scientometric study based on global publications during 2001-2020

B. M. Gupta
S. M. Dhawan
Ghouse Modin N. Mamdapur

The study provides a quantitative and qualitative description of global research in "Natural Language Processing" (NLP) using bibliometric methods. The analysis is based on publications data sourced from Scopus database for the period 2001-2020. The purpose of the study is to understand the status of NLP research at the global, national, institutional, and author level. The study highlights the productivity and performance of NLP research on a series of metrics as well as provides a visual view of collaborative network relationship between authors, research institutions, and leading countries using standard software tools. In addition, the study identified the leading players in NLP research such as key countries, institutions, authors, and areas of research. According to the study, the USA leads in global publications output as well as it leads in terms of relative citation index.

Keywords: *Natural language processing, Global publications, Scientometrics, Bibliometrics.*

B. M. Gupta[@]
Formerly with CSIR-NISTADS
New Delhi 110012
Delhi
India

[@] bmgupta1@gmail.com

S. M. Dhawan[#]
Formerly with CSIR-NPL
New Delhi 110012
Delhi
India

[#] smdhawan@yahoo.com

Ghouse Modin N. Mamdapur[*]
Synthite Industries (P) Ltd
New Product Development & Research
Kolenchery 682311
Kerala
India

*ghouse@synthite.com

[*] (Corresponding Author)

1. Introduction

Computers have been great at working with structured data such as tables and spreadsheets. But we humans cannot interact with computers because humans communicate in words, not in tables. A whole lot of information that is out there in the world cannot be accessed and made intelligible to computers because it is in natural languages and that such a text data is highly unstructured.

The challenge in this context has been how to make computers understand unstructured text and voice data and respond to such a data in much the same way humans do [1]. This is where National Language Processing (NLP) has come to gain importance and play a big role to let computers interact with humans, understand human languages, and get deep insights into business intelligence hidden under unstructured text data [2]. To put it simply, NLP has since evolved as a major technology that enables human-computer interactions in natural languages [3]. The ever growing importance of NLP technology can be judged from some of the features services of Google such as search engine, translate, auto-correct or auto-complete that are now fast becoming popular. Google uses NLP technology to run these services on the internet. Natural Language Processing imbibes important concepts in computer science, machine learning, deep learning, data science and computational linguistics into its algorithms for enabling human-computer interactions and vice versa, in an intelligent and useful way. In this regard, NLP undertakes to perform some basic tasks that are necessary to breakdown human text and voice data in ways that help computer programs to make sense out of it. Some of the most common tasks used for the purpose include semantic analysis, syntactic analysis, morphological analysis, lexical (structure) analysis, etc [4]. Top applications of NLP are machine translation, speech recognition, user interfaces, sentiment analysis, auto- text summarization, auto- text recognition, information extraction, and information search and retrieval etc [5]. Nowadays, businesses and industry intensively use NLP technology to manage their operational tasks online such as hiring and recruitment, advertising, customer service, healthcare and market intelligence. Given the ongoing trends in NLP research, it is believed that future research in natural language processing is going to be trans-formative in nature, driving research from data-driven endeavours to intelligence-driven endeavours in businesses, industry, and academia.

2. Literature Review

Only few bibliometric studies have been published on natural language processing and its applications particularly in medical field. Among such studies, Chen, Ding, Xu, Wang and Hao [6] studied Natural Language Processing (NLP) empowered mobile computing using bibliometric analysis methods. The data for study was retrieved from WoS for the period 2000–2016. The purpose of this study was to explore the status of the research in the field. The main focus was on (i) investigating statistical characteristics of research publications and publication collaborations, (ii) exploring geographical distribution of publications, (iii) visualizing scientific collaboration relationships, and (iv) revealing current hot research topics, themes and research topic changes.

In an another study Chen, Xie, Wang, Liu, Xu and Hao [7] examined NLP-empowered medical research publications (1405) retrieved from PubMed for the period 2007-2016. The analysis focused on three aspects: (i) literature distribution characteristics, (ii) network analysis to reveal scientific collaboration relations, (iii) thematic discovery and evolution using an affinity propagation clustering method.

Wang, Deng, Liu, Hu, Liang, Fan, Zheng, Wang and Lei [8] studied global literature (2336 records) on application of NLP in medical research during the last 20 years (1999-2018), with the main aim to identify publication trends, authors, countries, institutions, collaboration relationships, research hot spots, diseases studied, and research methods. Electronic medical records were the most used research materials, but social media such as Twitter have become important research materials since 2015. Cancer (24.94%, 103/413) was the most common subject area in NLP-assisted medical research on diseases, with breast cancers (23.30%, 24/103) and lung cancers (14.56%, 15/103) accounting for the highest proportions of studies.

Chen, Xie, Cheng, Poon, Leng, and Wang [9] analyzed bibliographical data on natural language processing collected from three databases from 2001 to 2018, using methods such as performance analysis, science mapping and structural topic modeling. Topical trend visualization and test analysis were further employed to quantify the effects of the year of publication on topic proportions. Topical diverse distributions across prolific countries/regions and institutions were also visualized and compared. In, addition, scientific collaborations between countries/regions, institutions, and authors were also explored using social network analysis.

Lopez-Martinez and Sierra [10] analysed 15,639 global bibliographical records on natural language processing collected from different databases with the main aim to identity the main sources where research is published, the most productive and influential countries and research institutions where research is carried out on NLP and its applications, the most important countries and institutions where research is carried out on NLP and its applications, the individual top active researchers, the most cited authors in the field, as well as the main topics and subtopics in the field

The topic of natural language processing of late, has been receiving significant attention by several major countries in terms of R&D planning and investments. Commercial interests are driving NLP applications in businesses, industry and academia. Given such research trends and the ever-growing interest in bibliometric studies as seen from literature review studies, it is considered worthwhile that a bibliometric study be undertaken in NLP research with the objective (i) to examine the past and present status of NLP research, (ii) to identify the leading countries, organizations, authors and journals; (iii) to visually map the collaborative network linkages among top countries, organizations and authors; (iv) to identify key research areas through significant keywords and also by studying their co-occurences using a mapping software; and (v) to study the bibliographic characterstics of high-cited papers.

3. Methodology

The study sourced data on NLP research from the Scopus database (http://www.scopus.com). In its search strategy, the study tagged the keyword “natural language processing” to two fields - “Keyword” and “Title”(Article Title), and confined the initial search to the publication period ‘2001-2020’. The database search returned 2221 pub-

lications. In the subsequent rounds of database search, the global output was refined by the country of publication, one by one, with a view to identify top 15 most productive countries in the subject. The records of search output were downloaded in csv file format, and later analyzed using statistical methods. By using analytical provisions as available in Scopus database, the search output was grouped by broad subject areas, collaborating countries, contributing authors, affiliating research organizations, and source journals, etc. The study used bibliometric techniques and methods to evaluate and assess the performance of the NLP research on select parameters such as most productive countries, organizations, authors and source journals. The VOSviewer and biblioshiny app for bibliometrix were used to evaluate and visualize the interactions among most productive countries/territories, organizations, authors and keywords.

(KEY (natural and language and processing) and TITLE (natural and language and processing)) and PUBYEAR > 2000 and PUBYEAR < 2021

4. Analysis & Results

4.1 *Publication and Citation Analysis*

The global research in the domain of "Natural Language Processing" (NLP) showed an increasing research trend during the 20-year period under study. The annual output in the subject jumped up from 24 in the year 2001 to 456 publications in the year 2020, a 18.64% growth. Its global average productivity was computed as 111.05 publications per year. But in the last five years of this study period, 2016-2020, it soared to 268.4 publications per year, a substantial jump in research productivity. The study also observed that 10-year cumulative NLP research had increased by 315.10%, from 431 in 2001-10 to 1790 publications in 2011-20. Such a big rise in global productivity in the second half of the study period may be attributed either to the emergence of new and additional research groups in NLP research or to a better research performance levels of the already established research groups. The NLP research had accumulated a total of 2221 publications in 2001-2020, during the period of analysis in this study. An analysis of NLP research by publication type reveals that articles and conference papers (51.60% and 40.48% of the total output) accounted for the dominant publication share respectively. The NLP research also accounted for a marginal share in reviews (3.38%), book chapters (1.98%), editorials (0.77%), notes (0.77%), conference reviews (0.36%), letters (0.27%), books (0.23%), retracted (0.09%) and short surveys and errarum (0.05% each).The study found that NLP research received an average of 13.21 citations per paper (CPP) since publication in 2001-2020. But its citation performance computed on a 10-year cumulative window was found to be much higher, a 24.78 CPP citations per paper in 2001-2010. But in 2011-2020 it slipped to 10.42 CPP (Table 1). The study also observed that citation performance of funded publications was 16.44 CPP, marginally above the global average. The study further observed that the productivity of funded research publications (526 out 2221 in this study) had also seen an increasing research trend. When evaluated on the basis of 10-year cumulative research output, it was seen that the productivity of funded publications increased from 61 in 2001-2010 to 465 in 2011-20. Another interesting obser-

Table 1

Global Publications Output and Citations on NLP during 2001-2020

Publication Period	TP	TC	CPP	FP	Publication Period	TP	TC	CPP	FP
2001	24	569	23.71	2	2013	93	1216	13.08	18
2002	21	252	12.00	2	2014	95	1564	16.46	19
2003	27	1242	46.00	5	2015	118	1580	13.39	20
2004	28	514	18.36	6	2016	151	2222	14.72	32
2005	41	487	11.88	6	2017	168	1801	10.72	52
2006	46	899	19.54	10	2018	204	2100	10.29	57
2007	54	972	18.00	9	2019	363	1189	3.28	120
2008	50	3168	63.36	8	2020	456	652	1.43	125
2009	61	1523	24.97	7	2001-10	431	10682	24.78	61
2010	79	1056	13.37	6	2011-20	1790	18649	10.42	465
2011	62	5276	85.10	7	2001-20	2221	29331	13.21	526
2012	80	1049	13.11	15					

TP=Total Papers; TC=Total Citations; CPP=Citations Per Paper; FP=Funded Papers

vation is that a 67% share of the total funded papers in NLP research had appeared the last 4 years of the study period. This sort of abrupt jump in funded research publications in particular in the last 4 years of the study period reflects a fast growing interest in NLP research and its applications. The leading funding agencies in NLP research were: National Institute of Health, USA (238 papers), U.S.Department of Health & Human Service (226 papers), U.S.National Library of Medicine (99 papers), National Science Foundation, USA (48 papers), European Commission (42 papers), National Center for Advancing Translational Sciences (40 papers), National Natural SCience Foundation of China (40 papers), National Cancer Institute (39 papers), National Institute of General Medical Sciences (29 papers) and National Heart, Lung & Blood Institute (24 papers).

4.2 *Most Productive Countries*

NLP research has been found a global research activity involving the participation of 89 countries. However, the distribution of global research by country of publication in the subject is found to be skewed For instance, during the 20 years of study period, 37 countries contributed barely 1-5 publications each, 17 countries 6-10 publications each, 15 countries 11-20 publications each, whereas in contrast 16 countries contributed 21-100 publications each, 3 countries 101-200 publications papers each and one single country 857 publications. The top 15 countries ranked on research productivity together contributed a high 89.87% share to the global research output. The publications by 15 most productive countries received a very high 97.47% share in global citations. The USA leads in global ranking of most productive countries with a 38.50% share, followed

by 14 others but their global publications share is in single digit. For instance, China, India and U.K. contributed 8.55%, 7.88% and 6.08% share respectively. Germany, Japan, France, Canada, Italy, Spain and Australia contributed from 2.16% to 4.14% share, and the rest contributed 1.44% to 1.98% global publications share during the period under study. In the 10-year period between 2001-10 and 2011-20, India, China, Russia Federation, Netherlands and France showed an increasing trend in their contribution to NLP research by 0.10% to 8.62% Whereas the other 10 countries most productive counties under discussion showed a declining trend in NLP research down by 0.02% to 43.76% during the same period. The quality of NLP research by country of publication was evaluated on the relative citation index metric. The global average of relative citation index of NLP research is always taken as 1 (one). Only one country, the USA, (out of the top 15 countries) registered its relative citation index (1.87) above the global average. This trend reflects that most productive countries have been lacking in the quality of research. The top 15 most countries had collaborated in NLP research at international level. They contributed 17.39% to 69.30% share of their respective national out as a share of international collaborative publications , with an average of 25.05% of national output (Table 2).

Table 2

NLP Research: Top 15 Most Productive Countries 2001-2020

S.No	Name of the Country	Number of Papers			Share of Papers			TCP	CPP	ICP	%ICP	RCI	TCL
		2001-10	2011-20	2001-20	2001-10	2011-20	2001-20	2001-20					
1	USA	166	689	855	38.52	38.49	38.5	20556	24	141	16.5	1.9	184
2	China	27	163	190	6.26	9.11	8.55	1815	9.55	44	23.2	0.7	53
3	India	4	171	175	0.93	9.55	7.88	382	2.18	17	9.71	0.2	22
4	U.K.	37	98	135	8.58	5.47	6.08	1087	8.05	68	50.4	0.6	87
5	Germany	24	68	92	5.57	3.8	4.14	780	8.48	40	43.5	0.7	62
6	Japan	28	49	77	6.5	2.74	3.47	930	12.1	15	19.5	0.9	18
7	France	14	60	74	3.25	3.35	3.33	567	7.66	27	36.5	0.6	45
8	Canada	14	55	69	3.25	3.07	3.11	460	6.67	28	40.6	0.5	37
9	Italy	15	50	65	3.48	2.79	2.93	523	8.05	25	38.5	0.6	33
10	Spain	16	48	64	3.71	2.68	2.88	425	6.64	16	25	0.5	23
11	Australia	11	37	48	2.55	2.07	2.16	415	8.65	21	43.8	0.7	29
12	Russia Fed.	0	44	44	0	2.46	1.98	43	0.98	7	15.9	0.1	9
13	Netherlands	5	35	40	1.16	1.96	1.8	434	10.9	27	67.5	0.8	35
14	South Korea	8	28	36	1.86	1.56	1.62	95	2.64	8	22.2	0.2	8

Contd...

15	Brazil	9	23	32	2.09	1.28	1.44	78	2.44	6	18.8	0.2	11
Total of 15 countries		378	1618	1996	87.7	90.39	89.87	28590	14.3	490	25.1	1.1	656
World total		431	1790	2221				29331	13.2				
TCP=Total Citations; CPP=Citations Per Paper; ICP=International Collaborative Publications; RCI=Relative Citation Index; TCL=Total Collaborative Linkages													

4.2.1 Collaborative Linkages among Top 15 Countries

The top 15 most productive countries collaborated with each other in NLP research but the scale of their collaboration measured in terms of collaborative linkages varied from 3 to 137. The top three countries with the most collaborative linkages count are: the USA (137 linkages), U.K. (57 linkages) and Germany (37 linkages). On the other hand South Korea (3 linkages), Russian Federation (4 linkages) and Brazil (10 linkages) accounted for the the lowest collaborative linkages count. The collaborative linkages, between any two countries, ranged from 1 to 20. For example, the USA - China registered the most collaborative linkages (20), followed by the USA - the U.K. (19), the USA - Canada (17), USA - Germany (15), USA – France (13), (Table 3). The top 15 countries collaborative network map is shown in Figure 1, where all countries are depicted in four main clusters (represented by different colors). The degree and strength of research collaboration can be gauged through extent of thickness of links and distance between countries in the collaboration map. As expected, major dominance in research collaboration comes from USA, followed by UK, Germany, France, China and others.

4.2.1.

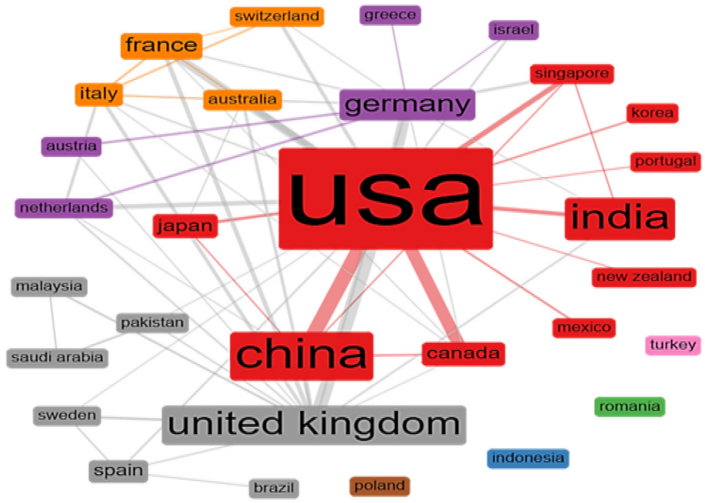

Figure 1

Collaboration networks chart of top 15 countries

Table 3

Collaborative Linkages among Top 15 Countries in NLP Research 2001-2021

S.No	Name of the Country	Country-Wise Distribution of Collaborative linkages	TCL (NOC)
1	USA	2(20), 3(7), 4(19), 5(15), 6(6), 7(13), 8(17), 9(3), 10(3), 11(9), 12(2), 13(9), 14(4), 15(10)	137(14)
2	China	1(20), 4(5), 6(2), 7(1), 8(3), 9(1), 11(1), 13(1)	34(8)
3	India	1(7), 4(3), 5(2), 7(1), 8(1), 12(1), 13(1), 14(1)	17(8)
4	U.K.	1(19), 2(5), 3(3), 5(3), 6(2), 7(6), 8(3), 9(5), 10(2), 11(4), 13(3), 15(2)	57(12)
5	Germany	1(15), 3(2), 4(3), 6(1), 7(3), 8(2), 9(2), 10(2), 11(2), 13(4), 15(1)	37(11)
6	Japan	1(6), 2(2), 4(2), 5(1), 10(1), 11(2), 14(1), 15(1)	16(8)
7	France	1(13), 2(1), 3(1), 4(6), 5(3), 8(3), 9(3), 10(1), 11(2), 13(1), 15(2)	36(11)
8	Canada	1(17), 2(3), 3(1), 4(3), 5(2), 7(3), 9(2), 15(1)	32(8)
9	Italy	1(3), 2(1), 4(5), 5(2), 7(3), 8(2), 11(2), 12(1), 13(4), 15(1)	24(10)
10	Spain	1(3), 4(2), 5(2), 6(1), 7(1), 13(1), 15(2)	12(7)
11	Australia	1(9), 2(1), 4(4), 5(2), 6(2), 7(2), 9(2)	22(7)
12	Russia Fed.	1(2), 3(1), 9(1)	3(4)
13	Netherlands	1(9), 2(1), 3(1), 4(3), 5(4), 7(1), 9(4), 10(1)	24(8)
14	South Korea	3(1), 4(1), 6(1)	3(3)
15	Brazil	4(2), 5(1), 6(1), 7(2), 8(1), 9(1), 10(2)	10(7)

TCL=Total Collaborative Linkages; NOC=Number of Countries

- Under Column 3, data inside brackets refers to linkages count between collaborating counties.
- Data outside the bracket refers to the collaborating country.
- For example, data 2(20) in row 1, column 3 depicts that China collaborated with USA (Country Sr No 1, Column 2) in 20 publications

4.3 *Subject-Wise Distribution*

NLP intersects with eight broad disciplines. Computer Science is the most sought after area of research pursuit in NLP research. Nearly 61% NLP research publications intersected with Computer Science. Besides, Medicine and Engineering representing as broad areas of NLP applications accounted for 27.42% and 22.96% share, are second and third most popular areas, followed by Mathematics & Social Sciences (15.71% and 11.08%), etc. Compared to world average (100) activity index involving all these eight disciplines, Medicine, Engineering, Decision Science and Biochemistry, Genetics & Molecular Biology witnessed an increasing trend in their research activity between 2001-10 to 2011-20. Amongst eight subjects, Mathematics registered the highest citations per paper (20.99 CPP) and Decision Sciences the least (3.38 CPP) (Table 4).

Table 4

NLP Research: Subject-Wise Breakup of Global Publications during 2001-20

S. No	Subject	Publications			Activity Index		TC	CPP	%TP
		2001-10	2011-20	2001-20	2001-10	2011-20	2001-20		
1	Computer Science	292	1073	1365	110.24	97.54	19985	14.64	61.46
2	Medicine	94	515	609	79.54	104.93	9241	15.17	27.42
3	Engineering	95	415	510	95.99	100.97	6066	11.89	22.96
4	Mathematics	96	253	349	141.75	89.95	7324	20.99	15.71
5	Social Sciences	49	197	246	102.64	99.36	2009	8.17	11.08
6	Decision Sciences	17	135	152	57.63	110.20	513	3.38	6.84
7	Arts & Humanities	32	90	122	135.16	91.53	1073	8.80	5.49
8	Biochemistry, Genetics & Molecular Biology	16	84	100	82.45	104.23	1579	15.79	4.50
	Total	431	1790	2221			29331	13.21	

TC=Total Collaborative Linkages; CPP=Citations Per Publication

4.3.1 Significant Keywords

As per bibliometric studies, author keywords either independently or in combination with related keywords invariably represent trending areas of research in a discipline (11). As seen from Scopus database, NLP global publications accumulated a total of 9635 author-keywords. The frequency of their occurence varied from 1 to 1866 times. Top 50 most significant keywords, ranging in frequency of their occurrence from 44 to 1866 times were analyzed.

During 2001-2020, the top trending topics in NLP research were Natural Language Processing (1866 keyword frequency), followed by Natural Language Processing Systems - a sub concept of NLP (1409), Computational Linguistics (443), Semantics (324), Machine Learning (320), Artificial Intelligence (298), Data Mining (249), Learning Algorithms (233), etc. (Table 5). These prominent trends can also be seen from the co-occurrence relationship visual chart of top 50 keywords (Figure 2). The size of each node in the chart is proportional to the total number of publications in which the related keyword had appeared.

Table 5

Top 50 Keywords co-occurrence (Of the Total 9635 Keywords)

S. No.	Keyword	Occur-rences	Total Link Strength	Clus-ter	S. No.	Keyword	Occur-rences	Total Link Strength	Clus-ter
1	Natural Language Processing	1886	6134	1	26	Syntactics	75	308	2
2	Natural Language Processing Systems	1409	4949	2	27	Predictive Value	74	351	1
3	Computational Linguistics	443	1823	2	28	Retrospective Study	74	344	1
4	Semantics	324	1360	2	29	Medical Informatics	73	334	1
5	Machine Learning	320	1429	3	30	Information Processing	72	306	1
6	Artificial Intelligence	298	1416	3	31	Speech Recognition	64	270	2
7	Data Mining	249	1220	3	32	Named Entity Recognition	61	296	2
8	Learning Algorithms	233	1311	3	33	Social Networking (Online)	60	305	3
9	Information Retrieval	228	1134	1	34	Cohort Analysis	58	265	1
10	Natural Languages	184	762	2	35	Search Engines	58	294	2
11	Electronic Health Record	177	794	1	36	Decision Making	58	277	3
12	Linguistics	175	712	2	37	Neural Networks	57	254	2
13	Learning Systems	175	1000	3	38	Social Media	56	230	3
14	Algorithms	167	836	1	39	Software	51	255	1
15	Deep Learning	144	609	2	40	Big Data	51	210	3
16	Text Processing	133	658	2	41	Data Handling	51	261	3
17	Electronic Medical Record	115	514	1	42	Reproducibility	48	263	1
18	Automation	106	527	1	43	Software Engineering	48	198	2

Contd...

19	Sensitivity And Specificity	106	521	1	44	Information Systems	47	199	2
20	Text Mining	104	502	3	45	Websites	47	198	2
21	Sentiment Analysis	96	356	2	46	Modeling Languages	46	237	2
22	Information Storage And Retrieval	93	465	1	47	Diagnosis	44	238	1
23	Ontology	87	401	2	48	Language Processing	44	221	1
24	Information Extraction	85	463	4	49	Radiology	44	225	1
25	Medical Record	75	336	1	50	Support Vector Machines	44	257	3

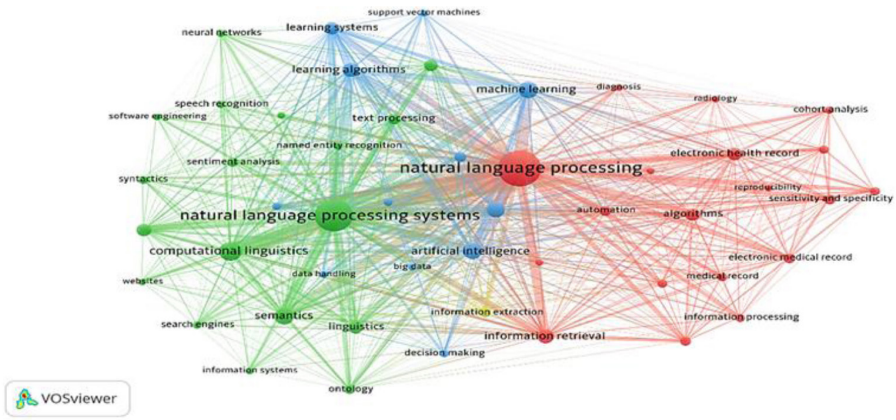

Figure 2

Keywords co-occurrence networks chart of 50 significant keywords

4.3.2 NLP Applications

NLP applications appear in 10 segments. Sentiment analysis tops the list of NLP applications accounting for a 4.41% share (98 publications) of global output, followed by Social Media/Network Monitoring (88 publications and 3.96% share), speech recognition (74 publications and 3.33% share), machine translation (61 publications and 2.75% share), text extraction (53 publications and 2.39% share), character recognition (47 publications and 2.12% share), text classification (43 publications and 1.94% share), etc. In terms of research impact of NLP contributions application-wise, speech recognition registered the highest citation impact per publication (61.42), followed by sentiment analysis (12.84%), text classification (11.47%), etc (Table 6)

Table 6

List of Applications of NLP during 2001-20

S.No	Application Area	TP			% TP	TC	CPP
		2001-10	2011-20	2001-20	2001-20		
1	Machine Translation	15	46	61	2.75	507	8.31
2	Sentiment Analysis	3	95	98	4.41	1258	12.84
3	Text Prediction	0	24	24	1.08	249	10.38
4	Text Classification	1	42	43	1.94	493	11.47
5	Speech Recognition	13	61	74	3.33	4545	61.42
6	Character Recognition	8	39	47	2.12	235	5.00
7	Chatbot/Dialogue System	3	18	21	0.95	145	6.90
8	Text Extraction	3	50	53	2.39	529	9.98
9	Text Summarization	3	11	14	0.63	138	9.86
10	Social Media/Network Monitoring	0	88	88	3.96	1155	13.13
	Global total	431	1790	2221	100.00	29331	13.21

TP=Total Publications; TC=Total Citations; CPP=Citations per Paper

4.4 *Most Productive Research Organizations*

In all, a total of 472 research organizations were found to have participated in NLP research during 2001-2020. The top 35 research organization contributed a 33.63% (747) global publications share and a 57.71% (16928) global citations share during the period. Of the top 35 most productive global organizations, 30 were from USA, 2 from France and U.K. and 1 from China. Their research productivity varied from 12 to 61 publications per organization. On further analysis of top 35 most productive organizations on citation and productivity metrics, it was observed that: (i) 11 organizations contributed above the group average productivity (21.34) of all 35 organizations, and (ii) 13 organizations performed above their group average (22.66 citations per publication and 1.72 relative citation index). It was also seen that the distribution of NLP research by research organization was skewed. For instance, of the total 424 research organizations, 211 barely published 1-5 publications each, 91 organizations 6-10 publications each, 32 organizations 11-20 publications each and 10 organizations 21-32 publications each. A scientimeric profile of top eight research organizations is given in Table 7.

Table 7

Top 8 most productive and top 8 most cited organizations in NLP Research 2001-2020

S. No	Name of the Organization	TP	TC	CPP	HI	ICP	ICP (%)	RCI	TCL
Top 8 Most Productive Organizations									
1	Harvard Medical School, USA	61	1131	18.54	17	11	18.03	1.40	231
2	Columbia University in the City of New York, USA	46	2145	46.63	20	2	4.35	3.53	54
3	Mayo Clinic, USA	41	611	14.90	13	4	9.76	1.13	98
4	Brigham & Women's Hospital, USA	41	861	21.00	14	6	14.63	1.59	165
5	V A Medical Center, USA	34	600	17.65	12	2	5.88	1.34	130
6	University of Utah, USA	34	557	16.38	14	2	5.88	1.24	83
7	Massausetts General Hospital, USA	34	788	23.18	13	3	8.82	1.75	134
8	Massachsetts Institute of Technology, USA	28	538	19.21	15	6	21.43	1.45	112
Top 8 Most Cited Organizations									
1	National Institute of Health, USA	12	846	70.50	10	3	25.00	5.34	28
2	National Library of Medicine, USA	14	834	59.57	10	5	35.71	4.51	34
3	Columbia University in the City of New York, USA	46	2145	46.63	20	2	4.35	3.53	54
4	Stanford University, USA	19	852	44.84	9	7	36.84	3.39	73
5	University of California, San Diago, USA	13	531	40.85	7	2	15.38	3.09	51
6	University of Pittsburg, USA	22	848	38.55	15	2	9.09	2.92	43
7	Yale School of Medicine, USA	12	457	38.08	5	1	8.33	2.88	54
8	Vanderbilt University, USA	14	517	36.93	8	2	14.29	2.80	48

TP=Total Publications; TC=Total Citations; CPP=Citations per Paper; HI= Herch Index; ICP=International Collaborative Publications; RCI=Relative Citation Index; TCL=Total Collaborative Linkages

4.4.1 Institutional Collaboration Linkages among Top 35 Organizations

The top 35 most productive organizations collaborated in NLP research; their collaborative linkages varied from 17 to 231. The top 3 organizations in terms of most collaborative linkages were: Harvard Medical School, USA Brigham & Women's Hospital, USA and Massausetts General Hospital, USA (231, 165 and 134 linkages). The three top organizations in terms of least collaborative linkages were: University of Shieffield, U.K., Vanderbilt University, USA and University of Pennsylavania, Perelman School of Medicine, USA (17, 21 and 22 linkages). The collaborative linkages between one-to-one organizations varied from 1 to 30. Harvard Medical School, USA and Brigham & Women's Hospital, USA both registered the most collaborative linkages (30 each), followed by Harvard Medical School, USA and Massausetts General Hospital, USA (24 linkages), Brigham & Women's Hospital, USA and Massausetts General Hospital, USA (18 linkages), University of Utah, USA and V.A.Medical Center, USA (15 linkages), Harvard Medical School, USA and MIT,

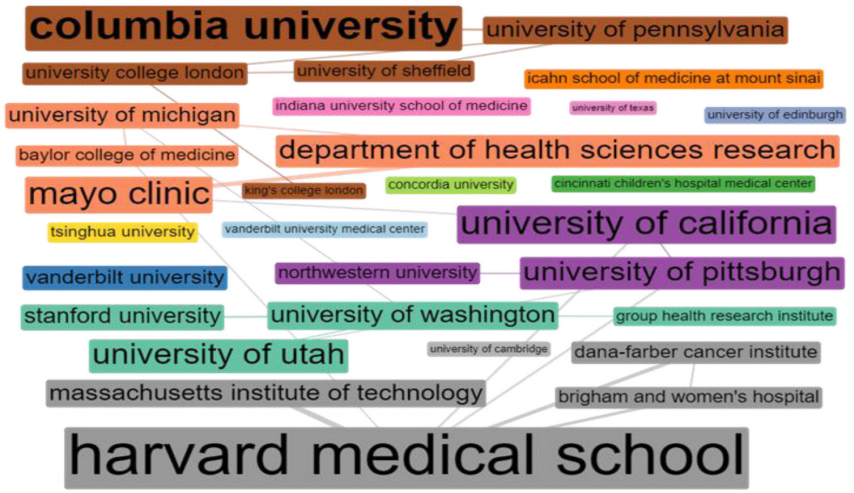

Figure 3

Institutional collaboration networks chart of top 35 research organizations in NLP

USA and Harvard Medical School, USA and Children's Hospital, Boston, USA (12 linkages each), Harvard Medical School, USA and Harvard T.H. Chan School of Public Health, USA and Harvard Medical School, USA and Dana Farber Cancer Institute (10 linkages each), Brigham & Women's Hospital, USA and Harvard T.H. Chan School of Public Health, USA and Massausetts General Hospital, USA and Harvard T.H. Chan School of Public Health, USA (9 linkages each) , CNRS, France and The Computer Science Laboratory for Mechanics and Engineering Sciences (LIMSI), CNRS, Orsay, France (8 linkages)etc. A collaborative networks visual chart of top 35 organizations is presented in Figure 3. The node indicates the number of publications, the more the number of publications the larger the node size. Tsinghua University tops the list of most collaborative publications, followed by University of Cambridge.

4.5 *Top 35 Most Productive Authors*

The 738 authors participitated in 2221 global papers in NLP research during 2001-20: 703 authors contributed 1-5 publications each, 29 authors 6-10 publications each and 6 authors 11-26 publications each. Among 35 top authors, the dominant partcipitation comes from 31 USA authors, followed by 2 from France and 1 each from Canada and U.K. The individual research output of top 35 authors varied from 5 to 26 publications per author and together contributed 318 publications and 8187 citations, which represent 14.32% and 27.91% share of global publications and citations in NLP research. On further analysis of 35 top authors, it was observed that: (I) Only seven authors have contributed publications above their group average productivity (9.08), (ii) Only nine authors have registered citation per publications and relative citation index above their group average citation impact (25.75 and 1.95)

Table 8

Top 8 most productive and top 8 most cited authors in NLP research 2001-2020

S. No	Name of the Author	Affiliation of the Author	TP	TC	CPP	HI	ICP	ICP (%)	RCI	TCL
Top 8 Most Productive Authors										
1	C. Friedman	Columbia University in the city of New York, USA	26	1271	48.88	14	0	0.00	3.70	15
2	H. Liu	Mayo Clinic, USA	26	429	16.50	12	3	11.54	1.25	185
3	W.W. Chapman	University of Utah, USA	18	1042	57.89	11	1	5.56	4.38	101
4	H. Xu	Columbia University in the city of New York, USA	15	329	21.93	9	0	0.00	1.66	101
5	S. Sohn	Mayo Clinic, USA	15	204	13.60	9	2	13.33	1.03	110
6	Y. Yang	Mayo Clinic, USA	11	114	10.36	4	1	9.09	0.78	56
7	A. Neveol	CNRS, France	10	158	15.80	6	5	50.00	1.20	45
8	K.B.Cohen	University of Colarodo, USA	9	123	13.67	4	3	33.33	1.03	39
Top 8 Most Cited Authors										
1	G.Hripcsak	Columbia University in the City of New York	6	617	102.83	6	0	0.00	7.78	22
2	G.K.Savova	Children's Hospital, Boston, USA	7	590	84.29	11	2	28.57	6.38	100
3	W.W. Chapman	University of Utah, USA	18	1042	57.89	11	1	5.56	4.38	101
4	D.S.Carell	Group Healthcare Research Institute, USA	6	307	51.17	8	0	0.00	3.87	95
5	C. Friedman	Columbia University in the city of New York, USA	26	1271	48.88	14	0	0.00	3.70	15
6	P.J.Haug	University of Utah, USA	8	356	44.50	7	0	0.00	3.37	40
7	T.Cai	Brigham & Women's Hospital, USA	8	300	37.50	7	2	25.00	2.84	75
8	S. Meystre	University of Utah, USA	6	215	35.83	4	0	0.00	2.71	17

TP=Total Publications; TC=Total Citations; CPP=Citations per Paper; HI= Herch Index; ICP=International Collaborative Publications; RCI=Relative Citation Index; TCL=Total Collaborative Linkages

4.5.1 Collaborative Linkages among Top 35 authors

Among top 35 authors, except few most authors showed active collaborative links among themselves in NLP research. The combined collaborative author linkages of top 35 authors with others in the group varied from 13 to 185, with (i) highest collaborative linkages depicted by H. Liu (185), S. Sohn (110), W.W.Chapman and H.Xu (100 each) and G.K.Savova (100) and (ii) least collaborative linkages depicted R. Witte (13), C. Friedman, (15), P. Zweigenbaum, S. Meystre and S. Crossley (7 each). Compared to combined collaborative linkages, the author-to author collaborative linkages were comparatively much small and varied from 1 to 13. Here, the largest collaborative linkages (13) were

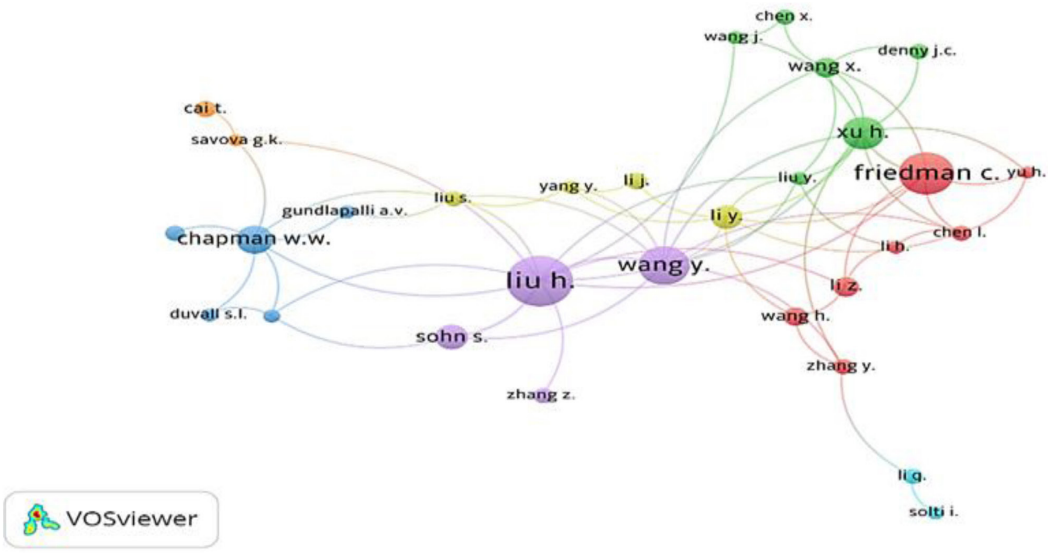

Figure 4

Co-Author Collaboration networks chart of top 35 authors

depicted by H. Liu and S.Sohn authors pairs, followed by H. Liu and Y.Wang and A. Neveol and P. Zweigenbaum (6 linkages each), C. Friedman and H. Xu (5 and J.C.Denny and L.Bastarache (5 linkages each). A top 35 authors collaborative networks map is presented in Figure 4. The thickness of a collaborative link between any two authors is proportional to its number of collaborative publications.

4.6 *Channels of Research Communication*

Of the 2221 global publications in NLP research, 995 (44.8%) appeared in 475 journals, 887 (39.94%) in conference proceedings, 298 (13.42%) in book series, 33(1.49%) as books, and 4 each in trade journal and undefined. Among 475 journals, 452 contributed 1-5 publications each, 9 contributed 6-10 publications each, and 5 contributed 21-52 publications each. The combined publications of top 35 journals accounted for 30.30% share in total journals publications in NLP research. *AMIA Annual Symposium Proceedings* contributed the most publications (52) among top 35 journals, followed by *Journal of the American Medical Informatics Association* (37 papers), *Journal of Biomedical Information* (36 papers), *BMC Medical Informatics & Decision Making* (24 papers), *International Journal of Medical Informatics* (23 papers) and *Journal of Medical Internet Research* (18 papers). Among most impactful journals, measured by citations per publications, the largest contribution comes from *Journal Artificial Intelligence Research* (124.8), *Bioinformatics* (91.5), *Journal of Biomedical Information* (44.97), *Journal of the American Medical Informatics Association* (43.0), *Automation in Construction* (40.25) and *Expert System with Applications* (34.83).

4.7 *Highly-Cited Papers*

Only 35 (1.58%) out of 2221 global publications in NLP research received 100 to 3891 citations per publication. These 35 publications are assumed here as high-cited publications(HCPs) and together received 14507 citations, averaging 414.48 citations per publication. The citations received by 35 HCPs is skewed, with (i) 18 publications in citation range 100-196 and 12 in citations range 224-356 on one hand and (ii) 3 publications in citation range 404-743 and 2 publications in citation range 2792-3891 on the other hand. The 35 HCPs consists of 18 articles, 9 reviews, 7 conference papers and 1 book). Among 35 HCPs, 16 involved one organization (indicating zero collaboration), 19 involve the partcipitation of two or more organizations: 12 national collaborative and 7 international collaborative. USA made the largest contribution (25 publications) in 35 HCPs, followed by 3 publications from Israel, 2 publications each from China, Germany, Japan and Singapore and 1 publication each from Greece, Mexico, Netherland, Sweden, U.A.E. and U.K. 138 authors from 72 organizations partcipitated in 35 HCPs. Columbia University, New York, USA contributed the most highly cited papers (5 papers), followed by Harvard Medical School, USA, Brigham & Women's Hospital, USA, Massachusetts General Hospital, USA, Stanford University, USA, National Library of Medicine, USA and National Institute of Health, USA (2 papers each), etc. Among 35 high-cited papers, C.Friedman (Columbia University in the City of New York) contributed the Most highly cited papers (3), followed by W.W. Chapan (University of Utah, USA)(2 papers), G. Hripcsak (Columbia University in the City of New York)(2 papers), etc. Among 35 high cited papers, 27 papers appeared in 17 journals, of which 4 papers appeared in Journal of Biomedical Informatics, 3 papers in Journal of the American Medical Informatics Association, 2 papers each in Bioinformatics, IEEE Computational Intelligence Magazine, Journal of the Artificial Intelligence and Synthetic Lectureson Human Language Technologies and 1 paper each in 7 other journals.

5. Summary and Conclusion

This study analyzed 2221 global publications in NLP research with the aim to understand research trends in the subject at various geographical, institutional and author levels. The data for the study was sourced from Scopus database covering the period 2001-2020 (2221 publications). The study examined research performance and mapped collaborative network linkages among top countries, organizations, authors and journals. It identified important keywords and studies their co-occurences throuch cluster analysis to identify important themes in NLP research. The NLP research showed an increasing research trend in publication productivity, indicating a 18.64% annual growth, and a 315.31% increase in 10-year cumulative productivity during the period between 2001-10 and 2011-20. The global publication productivity at the level of research organization was 4.7 per organization, and at author level, it was 3 per author. The NLP research revealed that funded research publications are cited more than the non-funded ones. Funded publications received an average of 16.44 citations per paper vis-a-vis the global average of 13.61 CPP for the total research output. The quality of global NLP research, measured in terms of HCPs that actually received alteast 100 citations per paper were

low. It was noted that just a very small, a 1.58% share (35 papers) received 100+ citations per paper. Of the 35 HCPs, the USA contributed the most papers (24), followed by Israel (3 publications papers), China, Germany, Japan and Singapore (2 publications each), Greece, Mexico, Netherlands, Sweden, U.A.E. and U.K. (1 publication each).

Even as NLP research is global, but analysis based on the output of top 15 most productive countries reveals that the USA, China, India and the U.K. lead in terms of research productivity and citation performance. The USA contributed the most publications, a 38.5% share to global output. The USA also collaborated with most other countries, 14 of top 15 and contributed the most international collaborative publications (137), Besides, the USA has been the home country to 30 of top 35 global research organizations, and also home country to 31 of the top 35 global research authors. China, the second most productive country, but it contributed 8.55% share in global output, and collaborated with 8 out of top 15 countries. India, the third most productive country, contributed a 7.88% share in global output, and collaborated with 8 out of top 15 countries. NLP research led to applications in 10 areas of interest -- sentiment analysis, social media/ network monitoring, speech recognition, machine translation, text extraction, character recognition, and text classification.

The study concludes that global NLP research is highly skewed. Of the top 15 most productive countries, only two countries, the USA and China, dominate the global research. The USA and China are also the home countries to centres of excellence in NLP research as well as to authors of excellence. The other productive countries in the top 15 countries in the global NLP research are just distant cousins as their research productivity at the national, institutional, individual author level is comparatively small and insignificant despite their 20-year long participation in the research. The stakeholders from the developing countries need to formulate appropriate policies and guidelines if they wish to promote and encourage NLP research in their academic and research institutions and centres.

References

[1] Banerjee, Dibyendu. (2020, April). Natural language processing (NLP) simplified : A step-by-step guide. A DSF Whitepaper. Retrieved August 3, 2021, from https://datascience.foundation/sciencewhitepaper/natural-language-processing-nlp-simplified-a-step-by-step-guide

[2] IBM Cloud Education. (2020, July) What is natural language processing? Retrieved August 3, 2021, from https://www.ibm.com/cloud/learn/natural-language-processing

[3] Raghavan, V., Gudivada, V., Govindaraju, V. & Rao, C.R. (2016, September 1). Cognitive computing: Theory and applications. (1st Edition). North-Holland.

[4] Rybkin, Yuriy. (2020, April). NLP applications in business. Retrieved August 3, 2021, from https://codeit.us/blog/nlp-applications-in-business.

[5] Reshamwala, A., Mishra, D. S. & Pawar, P. (2013, February). Review of natural language processing. Retrieved August 3, 2021, from https://www.researchgate.net/publication/235788362_REVIEW_ON_NATURAL_ LANGUAGE_PROCESSING.

[6] Xieling, C., Ding, R., Xu, K., Wang, S. & Hao, T. A bibliometric review of natural language processing empowered mobile computing. Wireless Communications and Mobile Computing, 14(1), 1-21 (2018, June). doi: 10.1155/2018/1827074.

[7] Chen, X., Xie, H., Wang, F.L., Liu, Z., Xu, J. & Hao T. (2018). A bibliometric analysis of natural language processing in medical research. BMC Med Inform Decis Mak., 18 (Suppl 1), 14. doi: 10.1186/s12911-018-0594-x.

[8] Wang, J., Deng, H., Liu, B., Hu, A., Liang, J., Fan, L., Zheng, X., Wang, T. & Lei, J. Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: Bibliometric study on PubMed. *J Med Internet Res.*, 22(1) (2020, Jan). e16816. doi: 10.2196/16816.

[9] Chen, X., Xie, H., Cheng, G., Poon, L. K. M., Leng, M., & Fu, L. Wang. Trends and features of the applications of natural language processing techniques for clinical trials text snalysis. Applied Sciences, 10(6), 2157 (2020). https://doi.org/10.3390/app1006215.

[10] Lopez-Martinez, R.E. & Sierra, G., Research trends in the international literature on natural language processing, 2000-2019: A bibliometric Study. *Journal of Scientometric Research,* 9(3), 310-318 (2020). doi:10.5530/jscires.9.3.38.BibTex.

[11] Dhawan, S.M., Gupta, B.M., & Mamdapur, G.M., Quantum machine learning: A scientometric assessment of global publications during 1999-2020. *International Journal of Knowledge Content Development & Technology*, 11(3), 29-44 (2021).