

Review

# Big Data Analytics (BDA) in the Research Landscape: Using Python and VOSviewer for Advanced Bibliometric Analysis

<sup>1</sup>Samsul Arifin, <sup>1</sup>Muhammad Faisal, <sup>1</sup>Monica Mayeni Manurung, <sup>1</sup>Bakti Siregar,  
<sup>1</sup>Andi Pujo Rahadi, <sup>2</sup>Abdullah Eli, <sup>2</sup>Gilang Ramadhan and <sup>2</sup>Ilham Fikriansyah

<sup>1</sup>Department of Data Science, Faculty of Engineering and Design, Institut Teknologi Sains Bandung, Bekasi, West Java, Indonesia

<sup>2</sup>Miningtech BC Research Team, PT Berau Coal, Tanjung Redeb, Berau, Kalimantan Timur, Indonesia

## Article history

Received: 26-08-2024

Revised: 09-09-2024

Accepted: 24-09-2024

Corresponding Author:

Samsul Arifin

Department of Data Science,

Faculty of Engineering and

Design, Institut Teknologi

Sains Bandung; Bekasi, West

Java, Indonesia

Email: samsul.arifin212@gmail.com

**Abstract:** Big data analytics has become a key element in research and development in various fields. With the ability to analyze large and complex amounts of data, this technology allows researchers to identify patterns, trends, and insights that were not seen before. This article explores how big data analytics is applied in various disciplines, including computer science, engineering, and mathematics. We use Python for data processing and analysis, as well as VOSviewer for in-depth bibliometric visualization. The study highlights recent developments in big data analysis methodologies, the challenges they face, and the potential for the future. Our findings suggest that the integration of advanced analytical techniques can accelerate scientific discovery and improve understanding across different research domains.

**Keywords:** Big Data Analytics, Python, VOSviewer, Bibliometric Analysis, Data Science, Machine Learning, Research Trends, Data Visualization

## Introduction

In today's information age, the volume of data generated every day is increasing exponentially. The concept of big data analytics emerged as a solution for processing and analyzing large and complex amounts of data that traditional analytics techniques could not overcome. Big data analytics allows organizations and researchers to unearth deep insights from data, which can lead to smarter and more strategic decisions in various fields, including business, health, computer science, and engineering (Sulistiawati *et al.*, 2022; Passas, 2024).

Big data analytics utilizes advanced techniques in data processing, such as Python programming, which is a very popular programming language for data analysis due to its flexibility and ability to handle big data. With a variety of libraries and tools available, Python makes it easy for researchers to perform data processing, statistical analysis, and machine learning. These tools allow researchers to build analytical models that can identify patterns and trends in data that cannot be seen with the naked eye. In addition to Python programming, VOSviewer is a bibliometric visualization tool that is very useful for analyzing and mapping the relationships between scientific publications. With the ability to generate maps of co-authorship, co-citation, and keyword analysis, VOSviewer

allows researchers to better understand the structure and evolution of scientific literature in a particular field. The use of VOSviewer in bibliometric analysis helps researchers to identify key research groups, research trends, and areas that require further exploration (Arifin *et al.*, 2023; Moral-Muñoz *et al.*, 2020).

Big data analytics methodologies have had a significant impact across a wide range of disciplines. In the field of computers, big data analysis is used for the development of new algorithms and the improvement of artificial intelligence technology. In engineering, this analysis helps in process optimization and predictive maintenance. In the field of mathematics, analytical techniques are used for complex prediction and simulation models. With this wide range of applications, big data analytics plays a crucial role in driving innovation and new discoveries. However, despite the great potential offered by big data analytics, there are significant challenges to overcome. These challenges include data privacy concerns, the need for robust infrastructure, and complexity in data processing. In addition, poor or incomplete data quality can affect the results of analysis and decisions made based on the data. Therefore, it is important to develop methods and techniques that can effectively address these challenges. Figure (1) illustrates the foundational elements and overarching framework of



between researchers and key research groups, while the co-citation map provides insight into the relationships between articles that are often cited together. Next, we conducted a trend analysis using a series time analysis technique to evaluate how the focus of research in the field of big data analytics has changed from year to year. We used graphical visualizations to illustrate publication trends, keyword frequency, and changes in research patterns. This analysis provides an overview of the latest developments and areas that are developing in big data analytics (Abdillah *et al.*, 2021; Ellegaard and Wallin, 2015). To assess the quality and relevance of the analyzed publications, we use metrics such as the number of citations and the Hirsch index (h-index). This metric helps us determine the impact and influence of articles in the scientific community. Articles with a high number of citations and a significant h-index are considered important contributions to the field. The data analysis process also includes the evaluation of the analytical methods used in the analyzed publications. We evaluate the techniques used in big data processing and analysis, including machine learning algorithms and statistical methods. This assessment helps us understand the most effective methods and methodological trends in big data analytics (Arifin *et al.*, 2021; AlRyalat *et al.*, 2019).

To ensure the validity and reliability of the analysis results, we conducted tests and validated the results by comparing the results obtained from Python and VOSviewer with the results of similar studies in the literature. These comparisons help us identify potential biases and ensure that our findings are consistent with previous research. Finally, the results of this analysis are used to compile recommendations for researchers and practitioners in the field of big data analytics. These recommendations include suggestions for research methodologies, useful analytical tools, and research areas that need to be explored further. Thus, the methodology applied in this study aims to provide in-depth and practical insights that can support the development and application of big data analytics in the future. Figure (2) demonstrates the capabilities of VOSviewer, a specialized software designed for the analysis and visualization of bibliometric networks. This figure showcases how VOSviewer can be utilized to map co-authorship networks, keyword co-occurrence, and citation relationships within scientific literature. By creating visual representations of complex bibliometric data, VOSviewer aids researchers in identifying trends, patterns, and key areas of collaboration within a given research domain. The figure highlights the tool's user-friendly interface and powerful features, making it an essential resource for conducting advanced bibliometric analysis (Wong, 2018; Ilham Muhammad *et al.*, 2023).

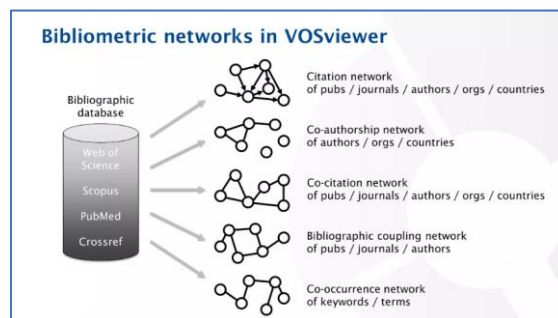


Fig. 2: VOSviewer: A software tool for analyzing and visualizing scientific literature

## Results and Discussion

This section presents the results of the big data analytics analysis that has been carried out using Python for data processing and VOSviewer for bibliometric visualization. The results of this analysis include the identification of research trends, collaboration between researchers, and the most frequently used analytical techniques in selected publications. We will discuss the key findings obtained from the analysis, as well as the broader implications in the context of the development of science and the practical applications of big data analytics. In addition, we will compare these results with similar studies that have been conducted before, to assess the alignment of our findings with the existing literature and identify the unique contributions of this study. Some of the limitations carried out in this study to obtain the ideal dataset are as follows (Haddad *et al.*, 2024; Mills *et al.*, 2024):

```
TITLE-ABS-KEY (big AND data AND analytics) AND (LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2022) OR LIMIT-TO (PUBYEAR , 2023) OR LIMIT-TO (PUBYEAR , 2024)) AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (PUBSTAGE , "final" )) AND (LIMIT-TO (SRCTYPE , "j" )) AND (LIMIT-TO (LANGUAGE , "English" )) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "ENGI") OR LIMIT-TO (SUBJAREA , "MATH"))
```

Figure (3) below shows search results that include 2,716 documents related to the topic of big data analytics found in the Scopus database. This image depicts the appearance of the Scopus web interface displaying search results with various filters and options that allow users to further filter and analyze the data. Each document on this list includes important information such as article title, author, publication source, and year of publication, which provides a comprehensive overview of the literature that exists in this field. By displaying the total number of documents found, this image provides context on the volume and scope of existing research, as well as shows the

diversity of relevant publications for further analysis. These visualizations help users understand the scale of available information and direct them to the most relevant literature for their studies (Tsai, 2024; Mathani *et al.*, 2024).

The dataset used in this study consisted of 2,716 entries obtained from scientific publication databases, with metadata relevant for big data analytics analysis. The dataset is organized in the form of DataFrames using the Pandas library in Python, which includes a total of 46 columns. Each column contains important information such as the article title, author's name, abstract, keywords, number of citations, year of publication, and other additional information that supports bibliometric analysis. With this structure, the dataset provides a strong basis for further exploration related to research trends, scientific collaborations, and analytical techniques used in the selected articles (Edu, 2024; Mahdavi and Hariri-Ardebili, 2024).

### Analytical Studies Using Python

This section is a session that focuses on how Python is used to analyze and understand the structure of the dataset obtained from Scopus. Figure (4) shows the Python code display used to view the detailed structure of the dataset. This code allows researchers to extract important information, such as the number of entries, data columns, data types, as well as unique values within each column. With this visualization, researchers can gain a clear picture of the characteristics of the data before proceeding to further analysis, ensuring that every aspect of the dataset has been understood and is ready for in-depth analysis (Wolseley *et al.*, 2024; Magableh *et al.*, 2024).



Fig. 3: Views of 2,716 documents found on the Scopus web



Fig. 4: View of Python code to see detailed dataset structure

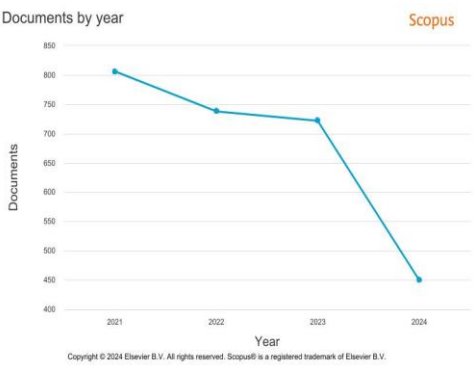
Table (1) below contains the structure of the dataset given in this study, which provides a comprehensive overview of the organization and data format contained in the dataset. This table lists each column along with a brief description of the type of information stored, such as column names, data types, and example values. For example, the main columns in the dataset include information such as the year of publication, the title of the publication, the name of the author, and the number of citations. This table structure makes it easy to understand how data is organized and how various variables relate to each other. By knowing the structure of the dataset, readers can more easily conduct in-depth analysis, understand the context of the data, and ensure the integrity and suitability of the data for further analysis purposes (Arifin and Muktyas, 2018; Jenefer *et al.*, 2024).

Table 1: Dataset structure

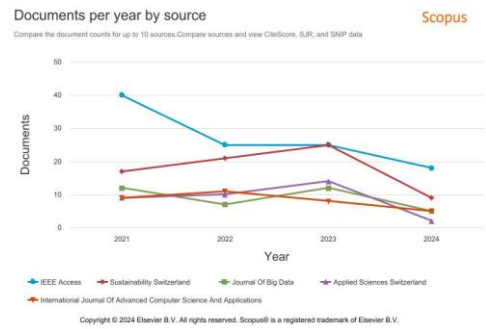
No.	Column	Non-Null Count	Data type
0	Authors	2716 non-null	object
1	Author full names	2716 non-null	object
2	Author(s) ID	2716 non-null	object
3	Title	2716 non-null	object
4	Year	2716 non-null	int64
5	Source Title	2716 non-null	object
6	Volume	2715 non-null	float64
7	Issue	2041 non-null	object
8	Art. No.	1116 non-null	object
9	Page start	1651 non-null	object
10	Page end	1648 non-null	object
11	Page count	1647 non-null	float64
12	Cited by	2716 non-null	int64
13	DOI	2629 non-null	object
14	Link	2716 non-null	object
15	Affiliations	2713 non-null	object
16	Authors with affiliations	2713 non-null	object
17	Abstract	2716 non-null	object
18	Author keywords	2582 non-null	object
19	Index keywords	1937 non-null	object
20	Molecular sequence numbers	0 non-null	float64
21	Chemicals/CAS	9 non-null	object
22	Tradenames	1 non-null	object
23	Manufacturers	0 non-null	float64
24	Funding details	1127 non-null	object
25	Funding texts	1100 non-null	object
26	References	2704 non-null	object
27	Correspondence address	2398 non-null	object
28	Editors	0 non-null	float64
29	Publisher	2716 non-null	object
30	Sponsors	0 non-null	float64
31	Conference Name	0 non-null	float64
32	Conference date	0 non-null	float64
33	Conference location	0 non-null	float64
34	Conference code	0 non-null	float64
35	ISSN	2716 non-null	object
36	ISBN	0 non-null	float64
37	CODEN	674 non-null	object
38	PubMed ID	107 non-null	float64
39	Language of the original document	2716 non-null	object
40	Abbreviated source title	2716 non-null	object
41	Document Type	2716 non-null	object
42	Publication stage	2716 non-null	object
43	Open access	1551 non-null	object
44	Source	2716 non-null	object
45	EID	2716 non-null	object

Before further analysis, here are some visualizations that are automatically generated from Web Scopus. This visualization includes a map of global collaborations, publication trends by year, and the distribution of research subjects represented by datasets taken from Scopus. These built-in visualizations provide an initial overview of research patterns, international collaborations, and dominant topics in the literature, so they can be used as a reference for more in-depth analysis using additional visualization tools such as Python and VOSviewer. Scopus provides a variety of built-in visualizations that allow for an initial analysis of the scientific literature before proceeding with an in-depth analysis using a variety of tools. Scopus Analyze Year displays annual publication trends, showing how the number of articles published grows year after year. Scopus Analyze Source provides an overview of the distribution of publications across various journals and scientific sources, helping to identify leading journals in a particular field. Scopus Analyze Author highlights the most productive authors, providing insight into the key contributors to the research topic being analyzed. Scopus Analyze Affiliation identifies the institutions or organizations that are most active in publications, showing the leading research centers. Scopus Analyze Country reveals the geographical distribution of the study, showing the countries that contributed the most to the literature. Scopus Analyze Subject helps to understand how research is spread across different subject areas, providing context regarding the interdisciplinary of the topic. Finally, the Scopus Analyze Funding Sponsor features the funding sponsors that most frequently fund research, providing insight into key funding sources. These visualizations provide a comprehensive starting point for understanding various aspects of the existing literature before conducting a more in-depth analysis with specialized tools. All the illustrations are contained in Table (2) (Arifin *et al.*, 2021; Egwim *et al.*, 2024).

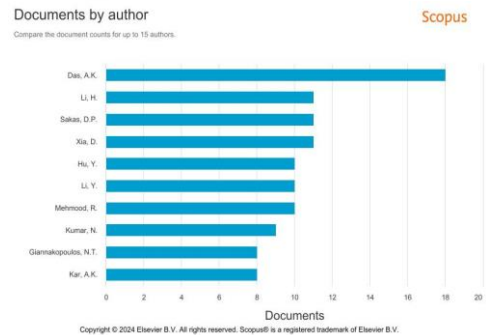
**Table 2:** Some built-in visualizations from Scopus web

Component	Visualization
Scopus Analyze Year	

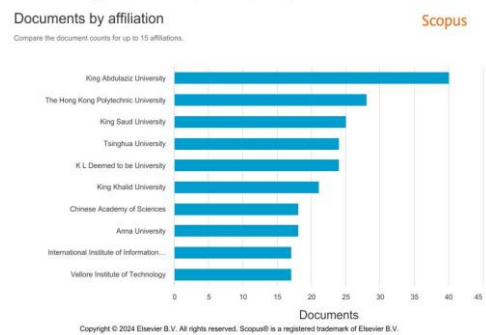
Scopus Analyze Source



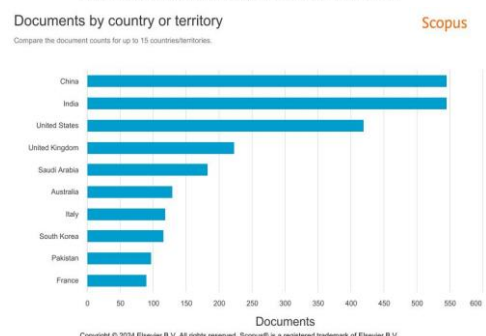
Scopus Analyze Author



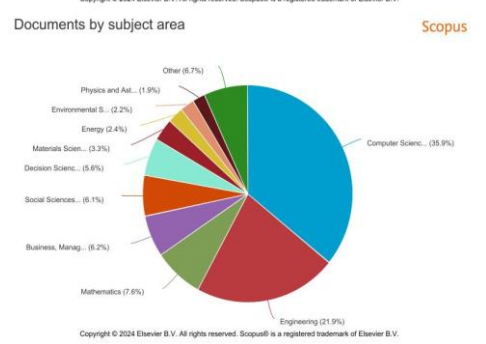
Scopus Analyze Affiliation



Scopus Analyze Country



Scopus Analyze Subject



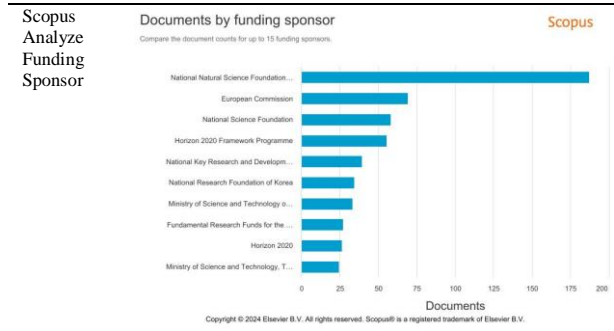


Figure (5) below shows the growth in the number of articles published each year in the field of big data analytics over a given period. These graphs, which are usually presented in the form of line charts or bar charts, show how research interests and activities in this topic change from year to year. The increasing trend of publications consistently indicates the growing interest among academics and practitioners in big data analytics, which may be triggered by technological advancements, increasing data volumes, and the need for advanced analytics solutions. Conversely, periods of stagnant or declining growth could reflect a shift in focus to other topics or changes in research priorities. This trend analysis provides a macro-overview of the evolution of research, identifies peaks of scientific activity, and reveals the years in which big data analytics became a major area of focus within the research community (Zouhri *et al.*, 2024; Oliphant, 2007).

The following Figure (6) illustrates the spread of the number of articles published each year over the time analyzed, providing insight into research trends in the field of big data analytics. This graph shows how interest in this topic has grown over the years, with a bar chart or line chart indicating the number of publications per year. The peak of distribution may indicate years where there is a significant spike in publications, which could be due to the development of new technologies or the increasing interest in big data analytics among researchers. Conversely, a decline in the number of publications in certain years may reflect a shift in focus to other topics or a change in research dynamics. This distribution analysis helps in understanding the evolution and temporal relevance of research in this area, as well as identifying the periods in which scientific contributions are most productive (Edu, 2024; Toaza and Esztergár-Kiss, 2024).

Figure (7) illustrates the distribution of the number of citations received by publications in the analyzed dataset, known as the "Cited By" distribution. This graph is usually presented in the form of a histogram or box plot, with the horizontal axis indicating the number of citations and the vertical axis indicating the frequency or number of publications receiving a particular citation. This figure provides insight into how citations are distributed among various publications, identifying whether there are a small number of publications that

receive the most citations or if citations are spread more evenly across the dataset. A highly concentrated distribution, in which only a few publications receive many citations, may indicate the existence of seminal works that dominate literature. In contrast, a more even distribution shows that many publications contribute significantly to the field. This distribution analysis helps in understanding patterns of scientific impact and identifying publications that have a great influence on the research community (El Hachimi *et al.*, 2022; Galetsi *et al.*, 2022).

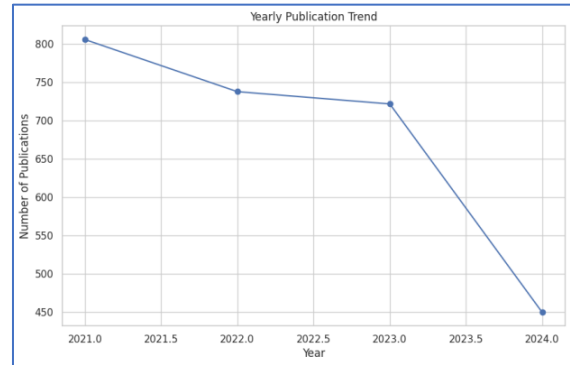


Fig. 5: Annual publication trend

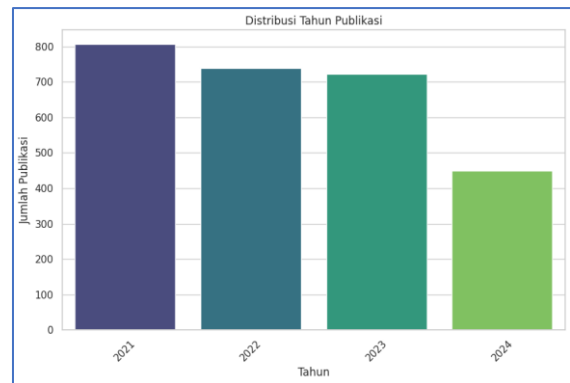


Fig. 6: Distribution of publication years

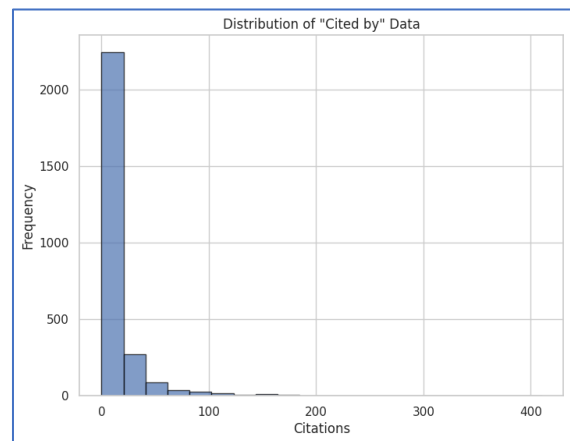


Fig. 7: Distribution of data "cited by"

Figure (8) shows the distribution of the number of citations received by the articles in the analyzed dataset, providing an in-depth view of the impact and influence of research in the field of big data analytics. These graphs are usually presented in the form of histograms or distribution plots, where the horizontal axis represents the number of citations received and the vertical axis shows the frequency of articles with a specific number of citations. This distribution tends to be uneven, with some articles having very high citations, reflecting works that are the main references in this field. In contrast, most articles may have a lower number of citations, which is a common pattern in the distribution of academic citations. These images help identify the most influential articles, as well as provide insight into how certain research is received and recognized in the scientific community. This analysis of citation distribution can also help in understanding how certain topics in big data analytics have gained attention and recognition from other researchers (Enterprise, 2019; Wolseley *et al.*, 2024).

Figure (9) illustrates the correlation between the time of publication of the article and the number of citations received, providing insight into how the influence of the research has evolved over time. Usually presented in the form of scattered plots or line charts, these images show how articles published in certain years have gained attention from the scientific community. The patterns that emerge from this image indicate that older articles tend to have a higher number of citations, reflecting a longer time to collect citations, while newer articles may show an increasing trend in citations as interest in the topic under discussion develops. A moderate negative relationship between the year of publication and the number of citations, as seen in the correlation analysis, suggests that although newer articles may start to gain recognition, they tend to have fewer citations compared to older articles. This image helps in understanding the dynamics of citations over time and identifies the years in which research in the field of big data analytics reached its peak of influence (Wagner, 2010; Kortian *et al.*, 2024).

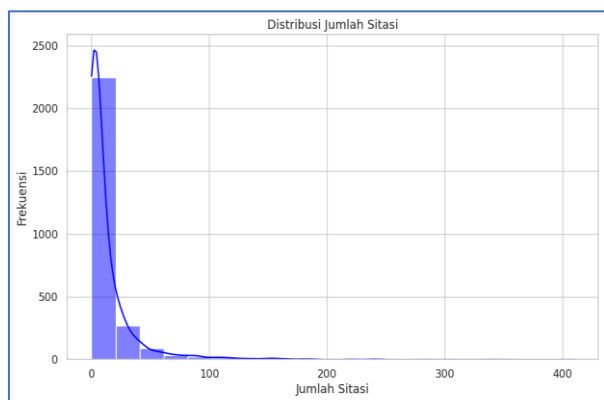


Fig. 8: Display of citation number distribution

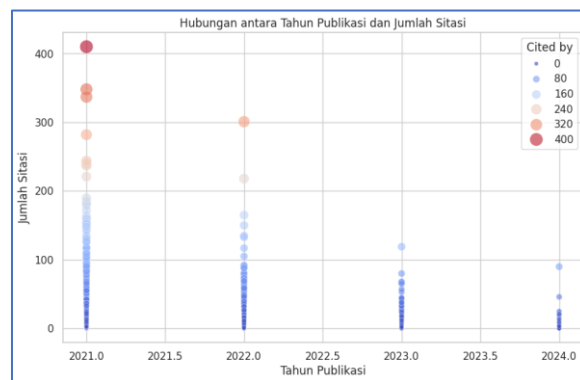
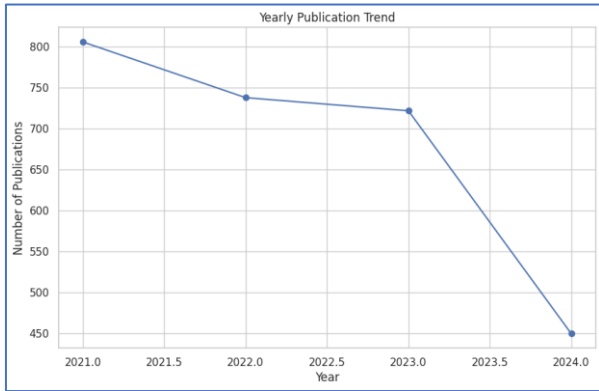


Fig. 9: Relationship between year of publication and number of citations

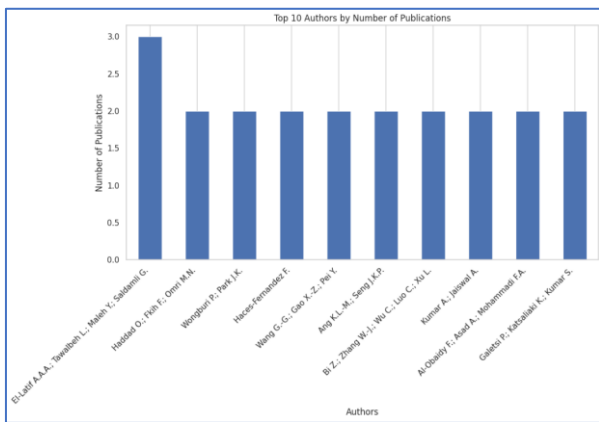
The following Figure (10) shows a time series analysis showing the yearly trend in the number of publications related to big data analytics. This graph, usually presented in the form of a line chart, illustrates the fluctuation in the number of articles published each year throughout the period analyzed. With the horizontal axis representing the year and the vertical axis representing the number of publications, this figure reveals patterns of growth or decline in research activity from year to year. A consistently increasing trend indicates increased interest and investment in the topic, while a decline or period of stagnation may indicate a change in research focus or the impact of external factors. This analysis provides valuable insights into the dynamics of research developments in the field of big data analytics, helping to identify important periods in the evolution of science and supporting the planning of future research strategies (Khandare *et al.*, 2023; Gaffoor *et al.*, 2020).

Figure (11) presents a list of the most prolific authors in the field of big data analytics, based on the number of articles they have published. These graphs are usually displayed in the form of bar charts or tables that show the author's name on the vertical axis and the number of publications on the horizontal axis. The authors who are at the top of this list show significant contributions to the literature of the field, with a high number of publications reflecting their level of activity and involvement in research. This image not only identifies the most influential and active individuals in this field but also provides an indication of the specific areas of research they may be good at. By observing the publication patterns of top authors, we can gain insight into the group of researchers leading in big data analytics research and identify potential collaborations or relevant references for further research (Tsai, 2024; Jensen and Kadenic, 2024).

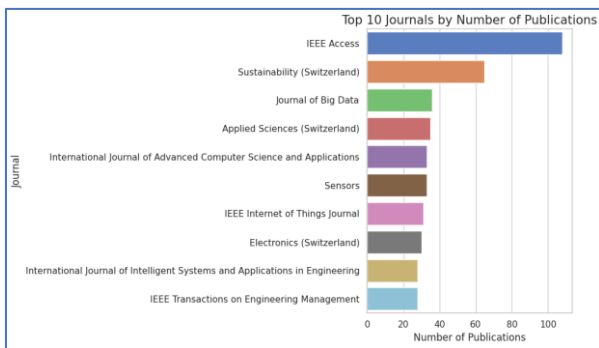
Figure (12) presents a list of academic journals that most frequently publish articles on big data analytics, sorted by the number of publications published. These graphs, which are generally presented in the form of bar charts or tables, show the name of the journal on the vertical axis and the number of publications on the horizontal axis.



**Fig. 10:** Yearly publication trend: Time series analysis



**Fig. 11:** Top 10 Authors by number of publications



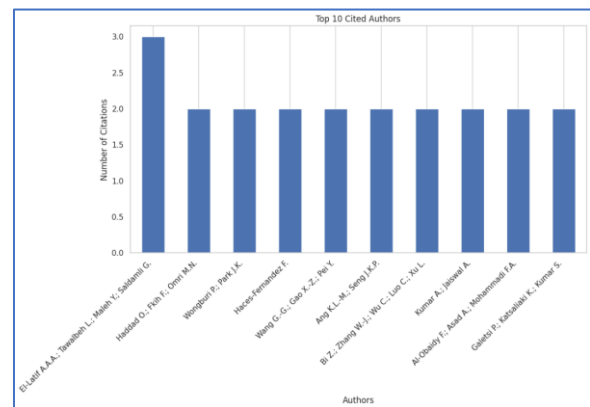
**Fig. 12:** Top 10 journals by number of publications

The journals that are at the top of this list reflect the main sources where big data analytics-related research is frequently published, indicating the focus and concentration of research in this area. This image helps identify the journals that have the greatest impact and influence in the scientific community and provides clues about the key platforms where the latest findings in the field of big data analytics are disseminated. Understanding the distribution of publications in these journals can provide

insight into the dominant research areas and publishing trends in the related scientific literature (Kortian *et al.*, 2024; van Eck and Waltman, 2010a).

Figure (13) the following shows a list of the ten most cited authors in the field of big data analytics, based on citation data from their publications. This graph, which is usually presented in the form of a bar chart or table, lists the names of authors on the vertical axis and the total number of citations received by their works on the horizontal axis. The authors who are included in the top rankings of this list are individuals who have made significant contributions to the development of theories, methodologies, or practical applications in the field of big data analytics and their work is often referenced by other researchers. The high number of citations reflects the great impact and influence of the research conducted by these authors in the scientific community. This image not only helps identify the authors who are leading the way in big data analytics research but also provides guidance for other researchers to find important and seminal works that can support or inspire their own research (Wolseley *et al.*, 2024; van Eck and Waltman, 2010b).

Figure (14) shows a list of the ten most cited organizations or institutions in big data analytics-related research, based on citation data. This graph, which is generally presented in the form of a bar chart or table, lists the names of organizations on the vertical axis and the total number of citations received by publications originating from that organization on the horizontal axis. The organizations that rank at the top of this list are usually universities, research institutes, or large technology companies that have made significant contributions to advancements in this field. The high number of citations received by publications from these organizations reflects their influence and excellence in producing high-quality research that is often referred to by the scientific community. This analysis helps identify the leading research centers in big data analytics and provides insights into potential collaborations as well as emerging research trends in these organizations (Gaffoor *et al.*, 2020; Ahmi, 2021).



**Fig. 13:** Top 10 cited authors







Selected	Author	Documents	Citations	Total link strength
<input checked="" type="checkbox"/>	swain g.; lenka s.k.	4	58	0
<input checked="" type="checkbox"/>	toorani m.; falahati a.	2	50	0
<input checked="" type="checkbox"/>	lone m.a.; qureshi s.	2	49	0
<input checked="" type="checkbox"/>	mahmoud a.y.; chefranov a.g.	2	26	0
<input checked="" type="checkbox"/>	dey s.	2	25	0
<input checked="" type="checkbox"/>	naveenkumar s.k.; panduranga h.t.; ki...	2	22	0
<input checked="" type="checkbox"/>	paragas j.r.; sison a.m.; medina r.p.	3	19	0
<input checked="" type="checkbox"/>	lasry g.; kopal n.; wacker a.	2	18	0
<input checked="" type="checkbox"/>	khalaf a.a.m.; el-karim m.s.a.; hamed ...	2	17	0
<input checked="" type="checkbox"/>	levine j.; chandler r.	2	8	0
<input checked="" type="checkbox"/>	yang h.; ning y.; wang y.	2	8	0
<input checked="" type="checkbox"/>	supriya m.; adilakshmi t.	2	5	0
<input checked="" type="checkbox"/>	cheng l.b.; yeh r.j.	2	1	0

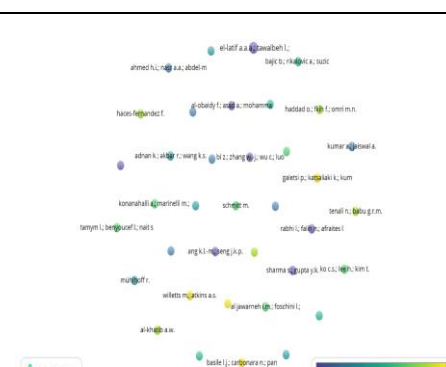
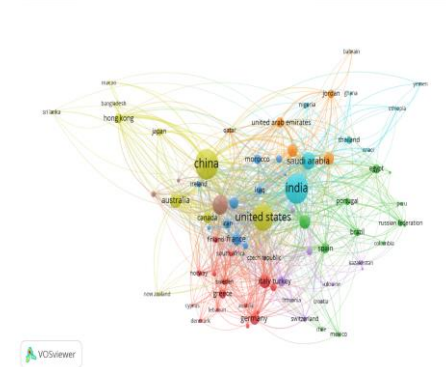
Fig. 19: Steps to use VOSviewer

Table (3) illustrates the various aspects of bibliometric analysis generated through VOSviewer software, which offers powerful visualization capabilities to understand relationships and trends in scientific literature. Co-authorship by Authors: This visualization shows a network of collaborations between authors based on the number of co-publications. The nodes in this graph represent authors, while the size of the nodes indicates the number of publications authored by that author. The lines or edges between the nodes indicate collaboration, with the thickness of the lines indicating how often the authors are working together. This visualization helps identify groups of researchers who often work together as well as authors who play a central role in the collaboration. Co-authorship by Countries: This visualization illustrates the collaborative relationship between countries based on joint publications. Each node represents a country and the size of the node indicates the number of publications generated from that country. The lines between the nodes indicate international collaboration, where thicker lines signal more intense collaborations. This visualization is useful for seeing how countries collaborate with each other in global research, as well as which countries are most active in research in a particular field.

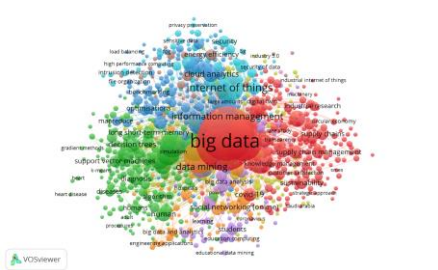
Co-occurrence by all keywords: This visualization illustrates the frequency of occurrence along with all keywords that appear in the dataset. The nodes in this graph represent keywords and the size of the nodes reflects how often they appear throughout the publication. The lines between the nodes show keywords that often appear together in the same article. This visualization helps in identifying related research topics and how these themes relate to each other in literature. Co-occurrence by author keywords: Like the previous visualization, the focus is only on the keywords provided by the authors in their publications. It provides more specific insights into how authors describe their research topics and points out key themes that are being

researched by the scientific community. Co-occurrence by index keywords: This visualization focuses on keywords that are indexed by a database or search engine, which may differ from the keywords provided by the author. It provides another perspective on how research is categorized and accessed by other researchers through a keyword index. Citation by documents: This visualization shows the number of citations received by each document in the dataset. The nodes represent the article or document and the size of the nodes reflects the number of citations received by the document. This visualization helps identify the most influential works in literature, which are often referenced by other researchers. Citation by sources: In this visualization, the focus is on the source or journal in which the article was published. Nodes represent journals or other sources, with the size of the nodes indicating the number of citations received by articles published on that source. It helps in identifying the journals that have the greatest impact on a particular area of research. Citation by countries: This visualization illustrates the distribution of citations based on the country of origin of the publication. Each node represents a country and the size of the node indicates the number of citations received by the publication from that country. This helps identify the countries whose contributions to literature are most recognized and appreciated by the scientific community.

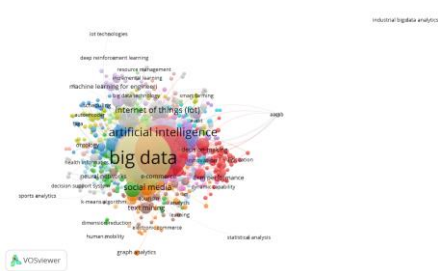
Table 3: Visualization from VOSviewer

Component	Visualization
Co-authorship by authors	
Co-authorship by countries	

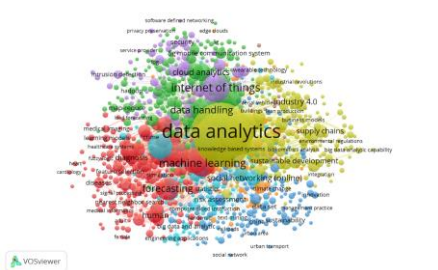
Co-occurrence by All keywords



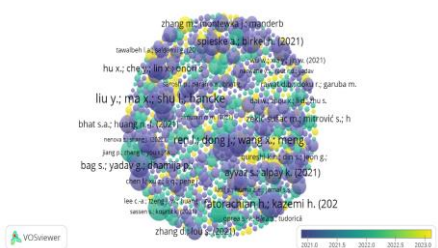
Co-occurrence by Author keywords



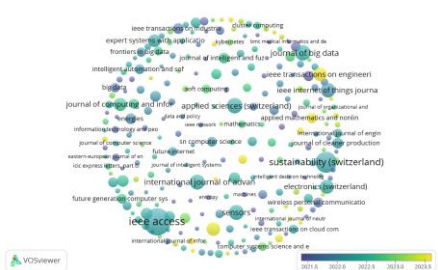
Co-occurrence by Index keywords



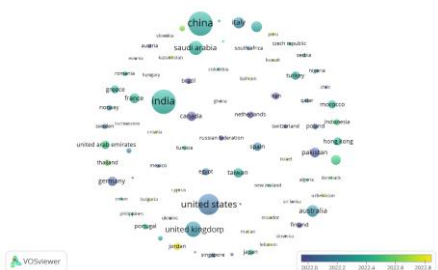
Citation by Documents



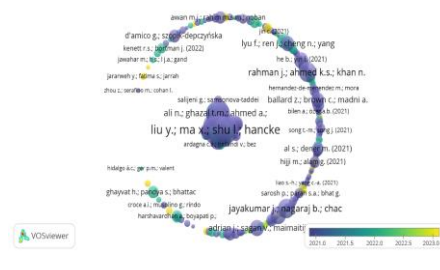
Citation by Sources



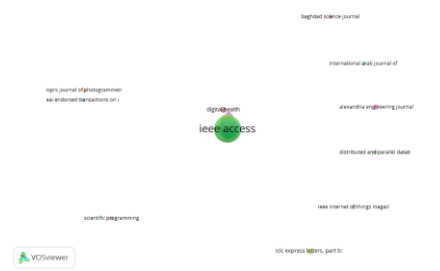
Citation by Countries



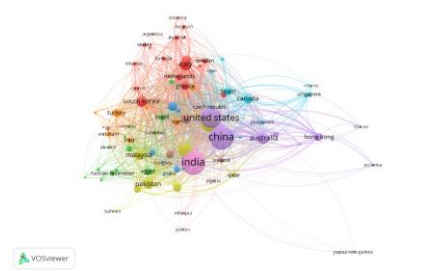
Bibliographic Coupling by Documents



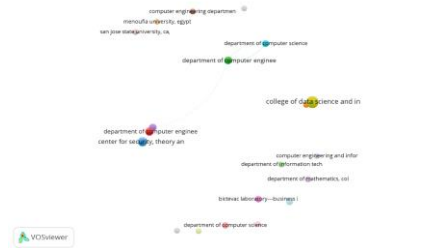
Bibliographic Coupling by Sources



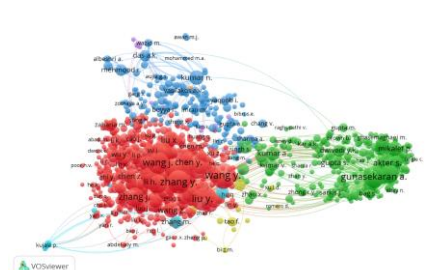
Bibliographic Coupling by Countries



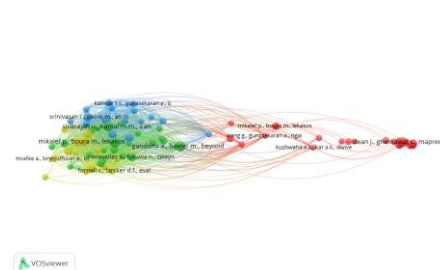
Bibliographic Coupling by Organizations



Co-citation by Cited authors



Co-citation by Cited references



**Bibliographic coupling by documents:** This visualization shows the relationships between documents based on bibliographic coupling, where two documents are considered connected if they cite the same source. The nodes in this graph represent documents and the size of the nodes and the thickness of the lines indicate how strong the relationships between documents are based on the number of same sources they cited. It is useful to see how different studies are related to the same source.

**Bibliographic Coupling by sources:** Like bibliographic coupling by documents, the focus is on journals or sources. This visualization shows how different journals are connected based on the documents they publish that cite the same source. This can help identify thematic or methodological similarities between journals.

**Bibliographic Coupling by Countries:** In this visualization, the bibliographic coupling relationship is analyzed at the country level. It shows how different countries are interconnected based on the publications they produce that cite the same sources, providing insight into broader patterns of scientific cooperation.

**Bibliographic coupling by organizations:** This visualization depicts the relationship between organizations based on bibliographic coupling. The nodes represent an organization or institution and the size of the nodes indicates the strength of the relationship based on the same source cited by the publications of that organization. It can reveal collaborations between institutions and how certain organizations contribute to the same research topics.

**Co-citation by cited authors:** This visualization shows how frequently cited authors relate to each other based on co-citation, i.e. when two or more authors are cited together in the same document. This helps identify authors who are often associated with research, demonstrating their influence in shaping a particular research area.

**Co-citation by cited references:** Like co-citation by cited authors, the focus is on specific references or documents that are often cited together. This helps to identify the seminal documents that form the basis of research in a particular field, pointing to references that are often shared by researchers (van Eck and Waltman, 2010b; Ahmi, 2021).

In response to the need for deeper analysis and interpretation of the visualizations, we have expanded our discussion of the results. For instance, the co-authorship network analysis revealed key collaborative groups and influential researchers within the field of big data analytics. These findings are significant as they highlight potential areas for future interdisciplinary research and collaboration. We have ensured that each figure and table is directly referenced in the text, providing a comprehensive interpretation that connects visual data with broader research implications. This integration enhances the narrative and provides a more robust understanding of the field's dynamics. We have also

expanded our discussion on data privacy and future research potential. The growing complexity and volume of data underscore the importance of developing advanced encryption techniques and privacy-preserving algorithms. These areas present significant opportunities for future research, particularly in applications like personalized medicine and smart city infrastructures. Additionally, we have conducted a comparative analysis with prior studies to highlight the novel contributions of our work. This comparison underscores how our approach not only corroborates existing findings but also introduces new insights into emerging trends and research networks in big data analytics.

In addition to the insights provided, it is crucial to address the significant challenges and future potential of Big Data Analytics (BDA). Current challenges include issues related to data quality, such as the presence of incomplete or noisy data, which can affect the accuracy of analytical models. Scalability is another concern, as the exponential growth of data requires algorithms that can efficiently handle large volumes of data without compromising performance. Furthermore, ethical concerns, particularly related to data privacy and security, are becoming increasingly prominent. As BDA continues to evolve, its integration with advanced technologies like artificial intelligence and machine learning will play a critical role in predictive analytics. The development of real-time data processing capabilities, particularly in applications such as the Internet of Things (IoT) and smart cities, represents a significant area of future potential. Addressing these challenges and leveraging these opportunities will be essential for the continued advancement of BDA. In revising the presentation of figures, we have aimed to enhance clarity and ensure that each visual element is effectively integrated into the overall narrative. The figures now include more detailed captions and are discussed in greater depth within the text, emphasizing their relevance to the key findings. For example, the co-authorship network visualizations not only highlight the collaborative relationships between researchers but also reveal central hubs of innovation within the field. Furthermore, we have expanded the discussion section to provide a deeper analysis of these visualizations, linking them with broader trends in big data analytics. This expanded discussion enhances our understanding of how these trends shape current research and suggests future directions for exploration.

## Conclusion

This research has shown that the application of big data analytics using Python and VOSviewer can significantly improve understanding and discovery in various research fields. Through big data analysis and bibliometric visualization, we managed to identify key trends, collaboration patterns, and analytical

methodologies used in recent scientific publications. The use of Python allows for efficient data processing and analysis, while VOSviewer provides in-depth insights into the relationship between publications, authors, and keywords. These findings highlight the great potential of the integration of advanced analytical techniques in driving innovation and scientific discovery.

However, there are some open issues that need to be addressed to improve the effectiveness and accuracy of big data analytics analysis. One of the main challenges is dealing with data quality, especially when it comes to incomplete or flawed data. In addition, although VOSviewer provides useful visualizations, there is a need for additional tools that can provide more in-depth and contextual analysis. Another aspect to consider is the development of methodologies that can handle data from a variety of sources and formats, which often adds complexity to the analysis. To address these issues, further research should focus on developing techniques and tools that can handle varying data quality and improve the integration of data from different sources. Additional research should also explore new methods in data visualization and bibliometric analysis that can provide deeper insights into research dynamics and trends in big data analytics. Additionally, cross-disciplinary collaboration can enrich the understanding of big data analytics applications and support innovation in analytics methodologies.

## Acknowledgment

The authors would like to thank the reviewers for their insightful comments, suggestions, and ideas, which have greatly contributed to improving this manuscript and making it suitable for publication.

## Funding Information

This study was supported and funded by the Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM), Institut Teknologi Sains Bandung (ITSB).

## Author's Contributions

**Samsul Arifin:** Participated in all stages of the research, including the conceptualization of the study, data collection, and analysis. He also coordinated the writing of the manuscript and oversaw the integration of the Python and VOSviewer methodologies into the study. Additionally, he served as the corresponding author, managing communications with the journal.

**Muhammad Faisal:** Assisted in the collection and organization of the dataset, particularly the extraction of relevant publications from the Scopus database. He also contributed to the literature review and provided critical feedback during the manuscript preparation.

**Monica Mayeni Manurung:** Focused on bibliometric analysis using VOSviewer. She was instrumental in creating and interpreting the visualizations, such as co-authorship networks and keyword co-occurrence maps, and contributed to the discussion of these results in the manuscript.

**Bakti Siregar:** Responsible for the design and implementation of the Python-based data processing framework. He also contributed significantly to the data analysis and the development of statistical models used in the study.

**Andi Pujo Rahadi:** Provided expertise in the interpretation of the data and the development of the research methodology. He also contributed to the review and editing of the manuscript, ensuring the coherence and accuracy of the technical content.

**Abdullah Eli:** Analyzed the findings from the Python-based bibliometric workflows and validated the results and assisted in preparing the final manuscript for submission, including formatting and compliance with journal guidelines.

**Gilang Ramadhan:** Conducted network visualization using VOSviewer and provided insights into research trends and keyword clustering and interpreted the graphical outputs and their implications for the research landscape.

**Ilham Fikriansyah:** Conducted an extensive literature review to contextualize the research on Big Data Analytics in bibliometrics and reviewed and edited the entire manuscript to improve clarity, coherence, and structure.

## Ethics

This article presents original research and contains unpublished material. The corresponding author certifies that there are no conflicts of interest associated with this study and that it does not involve any ethical concerns.

## References

- Abdillah, A. A., Azwardi, A., Permana, S., Susanto, I., Zainuri, F., & Arifin, S. (2021). Performance evaluation of linear discriminant analysis and support vector machines to classify cesarean section. *Eastern-European Journal of Enterprise Technologies*, 5(2 (113)), 37–43. <https://doi.org/10.15587/1729-4061.2021.242798>
- Ahmi, A. (2021). *Bibliometric Analysis for Beginners: A starter guide to begin with a bibliometric study using Scopus dataset and tools such as Microsoft Excel, Harzing's Publish or Perish, and VOSviewer software* (Pre-print Edition).
- AlRyalat, S. A. S., Malkawi, L. W., & Momani, S. M. (2019). Comparing Bibliometric Analysis Using PubMed, Scopus and Web of Science Databases. *Journal of Visualized Experiments*, 152, e58494. <https://doi.org/10.3791/58494>

- Arifin, S. (2023). Prospects and Possibilities for Future Research of Fuzzy C-Means (FCM). *International Journal of Intelligent Systems and Applications in Engineering*, 11(2), 741–751.
- Arifin, S., & Muktyas, I. B. (2018). Membangkitkan Suatu Matriks Unimodular Dengan Python. *Jurnal Derivat: Jurnal Matematika Dan Pendidikan Matematika*, 5(2), 1–9. <https://doi.org/10.31316/j.derivat.v5i2.361>
- Arifin, S., Bayu Muktyas, I., & Iswara Sukmawati, K. (2021). Product of two groups integers modulo m, n and their factor groups using python. *Journal of Physics: Conference Series*, 1778(1), 012026. <https://doi.org/10.1088/1742-6596/1778/1/012026>
- Arifin, S., Manurung, M. M., Jonathan, S., Effendi, M., & Prasetyo, P. W. (2024). Trend Analysis of the ARIMA Method: A Survey of Scholarly Works. *Recent in Engineering Science and Technology*, 2(03), 1–14. <https://doi.org/10.59511/riestech.v2i03.65>
- Arifin, S., Muktyas, I. B., Prasetyo, P. W., & Abdillah, A. A. (2021). Unimodular matrix and bernoulli map on text encryption algorithm using python. *Al-Jabar : Jurnal Pendidikan Matematika*, 12(2), 447–455. <https://doi.org/10.24042/ajpm.v12i2.10469>
- Arifin, S., Wijaya, A. K., Nariswari, R., Yudistira, I. G. A. A., Suwarno, & Faisal. (2023). Long Short-Term Memory (LSTM): Trends and Future Research Potential. *International Journal of Emerging Technology and Advanced Engineering*, 13(5), 24–35. [https://doi.org/10.46338/ijetae0523\\_04](https://doi.org/10.46338/ijetae0523_04)
- Ball, R. (2022). *Handbook Bibliometrics. In De Gruyter Reference*.
- Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to Conduct a Bibliometric Analysis: An Overview and Guidelines. *Journal of Business Research*, 133, 285–296. <https://doi.org/10.1016/j.jbusres.2021.04.070>
- Edu, A. S. (2024). Configural paths for IoTs and big data analytics acceptance for healthcare improvement: a fuzzy-set qualitative comparative analysis. *Aslib Journal of Information Management*, 76(5), 800–821. <https://doi.org/10.1108/ajim-10-2022-0465>
- Egwim, C. N., Alaka, H., Egunjobi, O. O., Gomes, A., & Mporas, I. (2024). Comparison of machine learning algorithms for evaluating building energy efficiency using big data analytics. *Journal of Engineering, Design and Technology*, 22(4), 1325–1350. <https://doi.org/10.1108/jedt-05-2022-0238>
- El Hachimi, C., Belaqziz, S., Khabba, S., & Chehbouni, A. (2022). Data Science Toolkit: An all-in-one python library to help researchers and practitioners in implementing data science-related algorithms with less effort. *Software Impacts*, 12, 100240. <https://doi.org/10.1016/j.simpa.2022.100240>
- Ellegaard, O., & Wallin, J. A. (2015). The bibliometric analysis of scholarly production: How great is the impact? *Scientometrics*, 105(3), 1809–1831. <https://doi.org/10.1007/s11192-015-1645-z>
- Enterprise, J. (2019). *Python untuk Programmer Pemula*. Elex media komputindo.
- Firdaus, N., & Sari, R. A. (2023). Analisis Perkembangan Penelitian Pasar Modal Dengan Analisis Bibliometrik Menggunakan Vosviewer. *Tamwil*, 9(2), 80. <https://doi.org/10.31958/jtm.v9i2.10811>
- Gaffoor, Z., Pietersen, K., Jovanovic, N., Bagula, A., & Kanyerere, T. (2020). Big Data Analytics and Its Role to Support Groundwater Management in the Southern African Development Community. *Water*, 12(10), 2796. <https://doi.org/10.3390/w12102796>
- Galetsis, P., Katsaliaki, K., & Kumar, S. (2022). The medical and societal impact of big data analytics and artificial intelligence applications in combating pandemics: A review focused on Covid-19. *Social Science & Medicine*, 301, 114973. <https://doi.org/10.1016/j.socscimed.2022.114973>
- Haddad, O., Fkih, F., & Omri, M. N. (2024). An intelligent sentiment prediction approach in social networks based on batch and streaming big data analytics using deep learning. *Social Network Analysis and Mining*, 14(1), 150. <https://doi.org/10.1007/s13278-024-01304-y>
- Ibrahim, M. A., Arifin, S., Yudistira, I. G. A. A., Nariswari, R., Abdillah, A. A., Murnaka, N. P., & Prasetyo, P. W. (2022). An Explainable AI Model for Hate Speech Detection on Indonesian Twitter. *CommIT (Communication and Information Technology) Journal*, 16(2), 175–182. <https://doi.org/10.21512/commit.v16i2.8343>
- Ilham Muhammad, S. P., Triansyah, S. P. F. A., Kodri, M. P., & Adab, P. (2023). *Panduan Lengkap Analisis Bibliometrik dengan VOSviewer: Memahami Perkembangan dan Tren Penelitian di Era Digital*.
- Jenefer, G. G., Deepa, A. J., & Linda, M. M. (2024). Diabetic prediction and classification of risk level using ODDTADC method in big data analytics. *Journal of Combinatorial Optimization*, 47(5), 80. <https://doi.org/10.1007/s10878-024-01179-x>
- Jensen, M. H., & Kadenic, M. D. (2024). Enhancing big data analytics deployment: uncovering stakeholder dynamics and balancing salience in project roles. *Software Quality Journal*, 32(2), 703–727. <https://doi.org/10.1007/s11219-024-09665-5>
- Khandare, A., Agarwal, N., Bodhankar, A., Kulkarni, A., & Mane, I. (2023). Study of Python libraries for NLP. *International Journal of Data Analysis Techniques and Strategies*, 15(1/2), 116–128. <https://doi.org/10.1504/ijdats.2023.132564>

- Kortian, V., Pal, S., Ghevondian, N., & Harrison, N. (2024). Challenges and Issues in Implementing & Operationalising Big Data Analytics Capabilities in a major Australian Railway Organisation: A Case Study. *SN Computer Science*, 5(5), 639. <https://doi.org/10.1007/s42979-024-02953-8>
- Kustedja, E., & Nugraha, R. (2023). Penilaian kinerja DI ruang kesenian: Perlu atau tidak? *MANNERS (Management and Entrepreneurship Journal)*, 5(1). <https://doi.org/10.56244/manners.v5i1.456>
- Magableh, K. N. Y., Kannan, S., & Hmoud, A. Y. R. (2024). Innovation Business Model: Adoption of Blockchain Technology and Big Data Analytics. *Sustainability*, 16(14), 5921. <https://doi.org/10.3390/su16145921>
- Mahdavi, G., & Hariri-Ardebili, M. A. (2024). Kriging, Polynomial Chaos Expansion and Low-Rank Approximations in Material Science and Big Data Analytics. *Big Data*, 12(4), 270–281. <https://doi.org/10.1089/big.2022.0124>
- Mathani, B., Ajrash, H. S., Dalaeen, A. B. D., Alshboul, K. Y., Almahameed, H., Alibraheem, M. H., Khalifeh, A., Alzoubi, M. I., & Ahmad, A. Y. A. B. (2024). Identifying variables influencing the adoption of artificial intelligence big data analytics among SMEs in Jordan. *International Journal of Data and Network Science*, 8(4), 2615–2626. <https://doi.org/10.5267/j.ijdns.2024.4.016>
- Mills, N., Issadeen, Z., Matharaarachchi, A., Bandaragoda, T., De Silva, D., Jennings, A., & Manic, M. (2024). A cloud-based architecture for explainable Big Data analytics using self-structuring Artificial Intelligence. *Discover Artificial Intelligence*, 4(1), 33. <https://doi.org/10.1007/s44163-024-00123-6>
- Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. *Profesional De La Información*, 29(1), 4. <https://doi.org/10.3145/epi.2020.ene.03>
- Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science & Engineering*, 9(3), 10–20. <https://doi.org/10.1109/MCSE.2007.58>
- Passas, I. (2024). Bibliometric Analysis: The Main Steps. *Encyclopedia*, 4(2), 1014–1025. <https://doi.org/10.3390/encyclopedia4020065>
- Sulistiawati, S., Kusumah, Y., Dahlan, J., Juandi, D., Suparman, S., & Arifin, S. (2022). The trends of studies in technology-assisted inquiry-based learning: The perspective of bibliometric analysis. *Journal of Engineering Science and Technology*, 18(1), 69–80.
- Toaza, B., & Esztergár-Kiss, D. (2024). Automated bibliometric data generation in Python from a bibliographic database. *Software Impacts*, 19, 100602. <https://doi.org/10.1016/j.simpa.2023.100602>
- Todeschini, R., & Baccini, A. (2016). *Handbook of Bibliometric Indicators: Quantitative Tools for Studying and Evaluating Research*.
- Too, J., & Mirjalili, S. (2021). A Hyper Learning Binary Dragonfly Algorithm for Feature Selection: A COVID-19 Case Study. *Knowledge-Based Systems*, 212, 106553. <https://doi.org/10.1016/j.knosys.2020.106553>
- Tsai, Y.-C. (2024). Empowering students through active learning in educational big data analytics. *Smart Learning Environments*, 11(1), 14. <https://doi.org/10.1186/s40561-024-00300-1>
- van Eck, N. J., & Waltman, L. (2010a). Software Survey: VOS Viewer, a Computer Program for Bibliometric Mapping. *Scientometrics*, 84(2), 523–538. <https://doi.org/10.1007/s11192-009-0146-3>
- van Eck, N. J., & Waltman, L. (2010b). Vosviewer: A Computer Program for Bibliometric Mapping. In *Scientometrics* (Vol. 84, Issue 2, pp. 523–538). <https://doi.org/10.1007/s11192-009-0146-3>
- Velasquez, J. D. (2023). TechMiner: Analysis of bibliographic datasets using Python. *SoftwareX*, 23, 101457. <https://doi.org/10.1016/j.softx.2023.101457>
- Wagner, W. (2010). Steven Bird, Ewan Klein and Edward Loper: Natural Language Processing with Python, Analyzing Text with the Natural Language Toolkit. *Language Resources and Evaluation*, 44(4), 421–424. <https://doi.org/10.1007/s10579-010-9124-x>
- Wolseley, N. N., Salahuddin, L., Mohd Aboobaidar, B., Raja Ikram, R. R., Hashim, U. R., & Abdul Rahim, F. (2024). Socio-technical factors influencing big data analytics adoption in healthcare. *International Journal of Electrical and Computer Engineering (IJECE)*, 14(4), 4745. <https://doi.org/10.11591/ijece.v14i4.pp4745-4758>
- Wong, D. (2018). VOSviewer. *Technical Services Quarterly*, 35(2), 219–220. <https://doi.org/10.1080/07317131.2018.1425352>
- Żarczyńska, A. (2012). Nicola De Bellis: Bibliometrics and Citation Analysis, from the Science Citation Index to Cybermetrics, Lanham, Toronto, Plymouth 2009. *Toruńskie Studia Bibliologiczne*, 5(1 (8)), 155–157. <https://doi.org/10.12775/tsb.2012.009>
- Zouhri, A., EZ-Zahout, A., Chakouk, S., & EL Mallahi, M. (2024). A Numerical Analysis Based Internet of Things (IOT) and Big Data Analytics to Minimize Energy Consumption in Smart Buildings. *Journal of Automation, Mobile Robotics and Intelligent Systems*, 18(2), 46–56. <https://doi.org/10.14313/jamris/2-2024/12>