The International Journal of Engineering and Information Technology

journal homepage:www.ijeit.misuratau.edu.ly

Improving Web Search Results Using KWSSWS System

Fakroun, E .
ebtfakroon@hotmail.com
Misurata University

Sullabi, M .
m.sullabi@it.misuratau.edu.ly
Misurata University

Abstract— **In today's digital age, efficient web searching and PDF retrieval have become crucial for various applications and industries. Development Operations (DevOps) plays a vital role in enhancing web search capabilities and streamlining PDF retrieval processes. This research paper explores the implementation of the Keyword Search System on Web Searching (KWSSWS) on Google PDFs retrieval to improve both DevOps and web searching for PDF documents. The study investigates the effectiveness of the (KWSSWS) system in enhancing the search experience and the performance of DevOps processes. The research methodology involves an experimental design and data analysis to evaluate the system's performance using real-world data. The results demonstrate the potential of the (KWSSWS) system to enhance web searching and PDF retrieval, thus providing valuable insights for further advancements in this domain. The advantage of the proposed (KWSSWS) system is that the collected files are very accurate and match the entered keywords hundred percent. This system is very useful primarily for all researchers around the world in various disciplines, where consider time and quality as the most important factors.**

Index Terms— **Web searching, Python, K-means algorithm, My-SQL database, Clustering, Weka visualizations.**

I. INTRODUCTION

In the digital era, the ability to efficiently search for information on the web and retrieve PDF documents has become increasingly important[1],[2],[3]. Development Operations (DevOps) practices aim to streamline software development and deployment processes, including web searching and PDF retrieval[4]. This research paper investigates the implementation of the Keyword Search on Web searching (KWSWS) system on Google PDFs retrieval to enhance both DevOps and web searching for PDF documents[5]. The data blast on the net has put levels of popularity on web search tools. However, individuals are a long way from being happy with the presentation of the current web search tools, which frequently return a huge number of reports in light of a user query[6]. A large number of the returned records are unessential to the client's needs. It is well known that search engine quality in its entirety cannot be measured with the use of traditional retrieval measures[7], [8]. But the development of new, search engine-specific measures, is not sufficient, either. Search engine quality must be defined more extensively and integrate factors beyond retrieval performance such as index quality and the quality of the search features. Clustering is one of the most relevant features of networks representing real systems[9], [10]. There exist various networks, including (WWW), the Internet, social networks, economic network, power network, traffic networks, neural networks, and so on. It is vital, as more and more researchers have shown that those appeared different networks have striking similarities with each other[11], [12]. The web can be seen as the largest intent to store all human knowledge, either explicitly or implicitly. The Internet actually is a stupendous graph, in which web pages are nodes and hyperlinks are edges[13], [14]. This abundance of related information with the dynamic and heterogeneous nature of the web makes information retrieval a difficult process for the average user[7],[15]. A technique is required that can help the users to organize, summarize and browse the available information from the web with the goal of satisfying their information needs effectively. The clustering process organizes the collection of objects into related groups. Web page clustering is the key concept for getting desired information quickly from the massive storage of web pages on (WWW)[16].

A. Original K-means Algorithm

From a practical point of view, clustering analysis is one of the main tasks of data mining. It is now used in many areas like knowledge discovery, pattern recognition, and so on. Many clustering analysis algorithms are available of which the most well-known is the K-means algorithm which is based on division. Clustering can enable users to find relevant documents more easily. This research study aimed to investigate the research articles and documents that are at the top in one cluster and other sites in the second cluster for top ranking, this research need URL, back-links, in-links, length of title are required in Weka data sate. “Clustering based on k-means” that it is closely related to a number of other clustering and location problems which include the Euclidean k-medians which minimize the sum of

Received 29 Apr , 2024; revised 10 May, 2024; accepted 15 Mar 2024. AVAILABLE ONLINE 08 AUG, 2024.

distances to the nearest center, and the geometric k-center problem, which aimed to minimize the maximum distance from every point to its closest center[17], [18].

B. Clustering and K-Means Algorithm

Calculation may be expressed concisely and precisely inside a defined formal language to determine a function, making it a feasible approach. The rules outline a computational process that starts from an initial state and may lack introduction information. This process progresses through a finite number of well-defined stages, ultimately producing output at the final state of completion. The transition from one state to another is non-deterministic; certain algorithms, referred to as randomized algorithms, use random data.[19]–[21].

C. k-means Algorithms clustering with WEKA

The k-means clustering with WEKA sample data set used for this example is based on the "bank data" which remains available in comma-separated format (bank-data.csv). Furthermore, this document assumes that appropriate INFO preprocessing has been performed. Moreover, a version of the initial data set has been created in which the (ID) field has been removed as well as the "children" attribute has been converted towards categorical. In the same way, the resulting INFO file is "bank. arff" as well as contains six hundred instances. In addition, as performing clustering in WEKA, this research has utilized its implementation of the K-means algorithm towards clustering the customers in this bank data set, to characterize the resulting **customer segments** [22]. Kapil and Chawla, 2016 have reported that clustering is one of the methods that has been proposed to be utilized in the area of data mining behind clustering is to assign objects to clusters; other than those belonging to other clusters by K-means clustering dataset[23].

D. WEKA and a couple of popular text file formats

There are some sorts of file formats that can be accepted in the Weka environment such as the ARFF file format, also, WEKA supports a couple of popular text file formats such as CSV, JSON, and MATLAB ASCII files to import data along with their own file format ARFF. They also have support to import data from databases through JDBC. Besides importing data, they have a wide collection of supervised as well as unsupervised filters to apply to the data to facilitate further analysis[24]–[26]. According to the above, this research study has used CSV format because it is the most suitable format for this research for the reason that it is easy to use and fast data transformation besides it is supported by Weka[27].

Hamoud and Atwell, 2016 have recommended that created for data mining with Waikato Environment for Knowledge Analysis (WEKA). Then the data was cleaned to improve data quality to the level required by the WEKA tool, and then converted to a comma-separated value (CSV) format to provide a suitable corpus dataset that can be loaded into WEKA. Then, the String To Word Vector filter was used to process each string into a bag or vector of word frequencies for further analysis with different data mining techniques. After that, we applied a clustering algorithm to the processed attributes and WEKA cluster visualizer[28].

Attwal and Dhiman,2020 WEKA is a Java-based software suite that implements a large number of machine learning algorithms. It can be used for performing different Data Mining tasks such as Data Preprocessing, Classification, Clustering, Association Rule mining and Visualization[29]. Vijayakamal and Narendhar, 2012 has declared that a machine learning data mining tool used for different analysis[30].

II. THE RESEARCH METHODOLOGY

The research theoretical model process flow diagram which explains and summarizes the research methodology, shows the proposed solution steps, As shown in the research process flow diagram Figure.1. as follows:

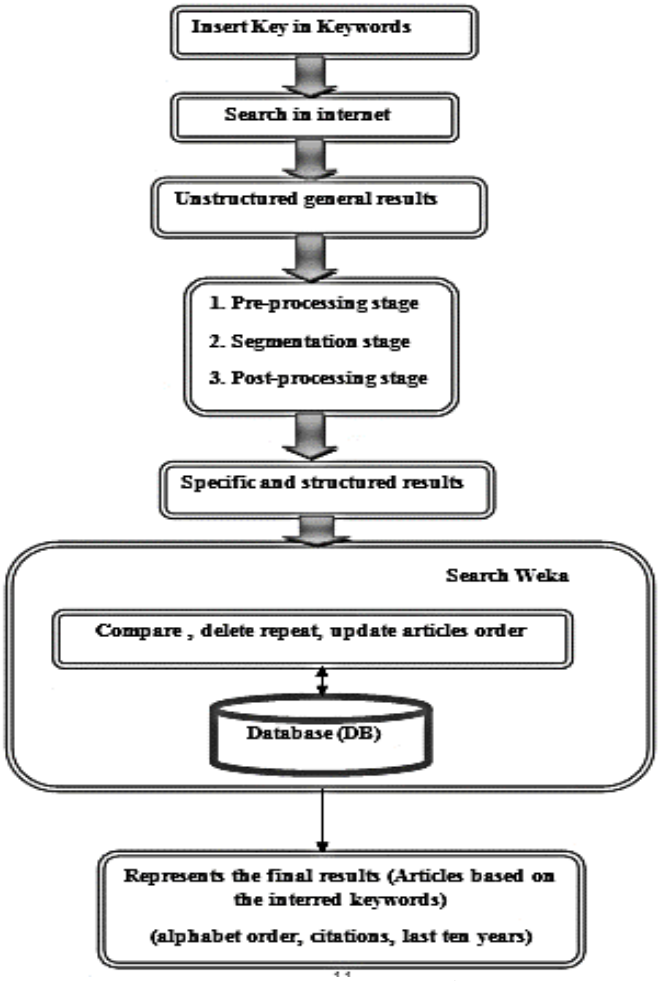

Figure 1: The research theoretical model process flow diagram.

This research method has to interact with the users to emulate the process of web searching from Google and Google Scholar. Furthermore, this research approach has explained the process of solving the accurate problems that can appear when researchers have access to the web and electronic databases to retrieve the PDFs (books, journal papers, conference papers, etc.) according to the interned Keywords. Moreover, this research methodology has used a simple and accurate systematic process based on searching and solving the research problem statement by applying some processes as an optimal solution. Also, this research method has focused on some main processes such as inserting keywords,

Searching on the internet, and unstructured general results this stage is called the pre-processing stage, segmentation stage, Post-processing stage, Specific and structured results, and after collecting PDFs and PDFs information such as PDFs topic, Title, Author name, publication year, and paper URL and store in the database according to the last ten years the final PDFs information has been ordered according to the last ten years of publication. Furthermore, the Weka dataset process (compared, eliminating duplication and visualization). In addition, the final results that the research has gotten from the search on the internet are compared with the Weka database, and the purification process takes place, in Weka to eliminate duplication, and then, visualize the results. So, the results of this method get at the first time while searching on the internet are not displayed it is sent to the Weka dataset after comparison; also will be filtered by using the Weka database lastly, the visualization stage of the Weka results is in the last stage. the proposed solution steps, as follows:

A. Insert keywords

In this step, the entered keywords must be matching the research domain, which is entered to obtain the required PDF files, in the same way, this step is divided into two main stages as follows:

- The first stage is to ask the user to enter keywords according to the research topic.
- The second stage is required to ask the user to addition enter main keywords, so, the user should enter keywords; this stage has been made to make the user pay attention to enter the main keywords according to the research topic, which returns a request, enter the important keywords for the research title to collect files for. This step is too tactful to warn the user to re-enter the keywords according to the research topic. In order to increase the accuracy of the search process.

B. Search on the internet

This step goes on searching on Google to retrieve the related PDFs which are according to the interred keywords. Search on **Google** and **electronic databases** such as **Google Scholar,** and Collect unstructured PDFs (**pre-processing stage**).

C. *Unstructured general results*: This step has been made to collect unstructured as well as no ordered PDFs from the electronic database, also, this stage is called the pre-processing stage.

D. *Segmentation stage***:** In this section, the proposed system collects PDFs and PDF information from an electronic database and the PDF information such as PDFs topic, Title, Author name, publication year, and paper URL and stored in the database;\

E. *Post-processing stage:* In this section, the proposed system collects the PDFs and at the same time collects PDF information such as PDFs topic, Title, Author name, publication year, and paper URL as well as stores in the file in the database according to the last ten years;

F. *Specific and structured results*: After collecting PDFs and PDFs information such as PDFs topic, the PDF title, the author name, publication year, and the paper URL and store in the database according to the last ten years the final PDFs information that has been ordered according to the last ten years of publication.

G. The Weka dataset process (compared, eliminating duplication and visualization):The proposed system is present the final results that the research has been gotten from the search on the internet which are compared with the Weka database, and the purification process takes place, in the Weka application the clustering by using KMeans algorithm will eliminate duplication in the PDFs titles, then, visualize the results. In the same way, the results of this system will get at first time while searching on the internet which are not displayed it is sending it to Weka dataset after comparison; also will be filtered by using Weka database. Starting K-Means Algorithm by selecting the number of cluster center and set initial cluster center randomly. Put object to closest cluster center; Recalculate the new cluster center and create cluster based on smallest distance; object move to cluster and show the results. In addition, the visualization stage of the Weka results is in the last stage. Furthermore, the research design will focus on the utilization of Weka Explorer to determine the common classification problems and towards identify the dataset visualization for presenting the research efficient as well as effective outcomes.

III. RESULTS AND ANALYSIS

A. A Comparison Between This Ystem And Traditional Web Searching

TABLE 1: A COMPARISON BETWEEN THIS SYSTEM AND TRADITIONAL WEB SEARCHING

A comparative factor	**The Search using the proposed system**	**The traditional Google Scholar search**
The main first interred keywords	1- Deep learning 2- Clustering 3- Algorithm	1- Deep learning 2- Clustering 3- Algorithm
The second interred keywords	1- Deep learning 2- Clustering	1- Deep learning 2- Clustering
Execution time	0:00:02.524801 part of the seconds	0.13 seconds
Number of the retrieved PDFs files	The number of files retrieved from the Internet search process ranges from 10 to 20 files	The number of files retrieved from the Internet search process ranges 0.41400 .
The mostly retrieved PDFs files	10 targeted PDF files	The researcher will move between pages to search manually for the target PDFs
The web search accuracy	The total number of PDFs estimated by Google Scholar related to the entered keywords: 900,000 PDFs. The number of PDFs retrieved from the Internet search	The number of PDFs estimated by Google Scholar related to the entered keywords: 900,000 PDFs. The number of PDFs retrieved from the Internet search process: 41400 PDFs. To calculate the

	process: 20 PDFs. Percentage of Retrieved PDFs = Number of Retrieved PDFs / Total Estimated PDFs) * 100% Given that 20 retrieved PDFs and a total estimated 900,000 PDFs, the calculation becomes: Percentage of Retrieved PDFs = (20 / 900,000) * 100% ≈ 0.0022% The result here is extremely low (0.0022%), and it seems like there might be an issue with your calculations or the data provided. Please ensure that the values you have provided are accurate and that the calculations are done correctly. If you intended for the calculation to result in 100%, I recommend revisiting the numbers and the formula to ensure their accuracy.	accuracy of the retrieved PDFs, The number of PDFs estimated by Google Scholar related to the entered keywords: 900,000 PDFs. The number of PDFs retrieved from the Internet search process: 41400 PDFs. To calculate the accuracy of the retrieved PDFs, which you're stating should be 80%. Accuracy = (Number of Relevant Retrieved PDFs / Total Number of Retrieved PDFs) * 100% Given that 41400 PDFs were retrieved and you're assuming the accuracy to be 80%, you can rearrange the formula to solve for the number of relevant PDFs: Number of Relevant Retrieved PDFs = (Accuracy / 100%) * Total Number of Retrieved PDFs Plugging in the values: Number of Relevant Retrieved PDFs = (80 / 100) * 41400 = 33120 PDFs 33120 PDFs out of 41400 PDFs are relevant. The percentage of relevant retrieved PDFs can be calculated as: (33120 / 41400) * 100% = 80% Rounded to two decimal places, this is approximately 80%. So, the required accuracy of the retrieved PDFs is indeed close to 80%, given the assumptions you've provided. However, please double-check the accuracy of your data and calculations to ensure their correctness.
The retrieved PDFs by the system according to the publication year	The system already has justified that the retrieved PDFs must be in the last ten years	Has been done manually by the researcher
The search country	Libya so the speed of search low	Libya so the speed of search low
Very important notes	1.Execution time depends on the speed of the computer processor, 2. The speed of the Internet connection to make the request from the server 3. The speed of the Internet connection to respond from the server Response	1.Execution time depends on the speed of the computer processor, 2. The speed of the Internet connection to make the request from the server 3. The speed of the Internet connection to respond from the server Response

B. Advantages of KWSSWS system

- It collects scientific papers and articles from the electronic library and puts them in a database to be used while when using traditional web searching the presented PDFs are presented just one time and then will be removed from the Google list to show another results or another web search [31].
- Collects a number of scientific papers in a fast time.
- Collects scientific papers in light of the entered keywords, which may help researchers for different disciplines to collect their scientific research papers according to their needs in the fastest time[32],[33].
- Store the collected files in the database to be used many times with the same Keyword condition.
- Alphabetical order and data display and presentation of any scientific papers in the Weka application.
- This system is scalable consequently, it works on most electronic databases such as the scholar database as recorded in[32].

IV. DISCUSSION

The study investigates the application of the KWSSWS system in obtaining Google PDFs to boost the efficiency of online searches in the context of Development Operations (DevOps). This section presents the main discoveries and consequences of the study, emphasizing the importance of the (KWSSWS) system in enhancing online searches by incorporating DevOps concepts. The incorporation of the (KWSSWS) system with DevOps concepts provides several advantages in the realm of online searching. Firstly, the use of Keyword search enables customers to precisely define their search queries and get more precise and pertinent outcomes from Google PDFs. By employing a focused strategy, the accuracy and effectiveness of online searches are improved, resulting in time and energy savings for consumers. The (KWSSWS) system utilizes sophisticated algorithms and approaches to extract keywords and obtain PDFs that are in line with users' search queries. Furthermore, incorporating DevOps ideas into the (KWSSWS) system improves its efficiency and overall performance. DevOps fosters cooperation, communication, and automation between development and operations teams, resulting in expedited deployment, ongoing integration, and smooth upgrades of the (KWSSWS) system. Organisations may provide a

seamless and effective web browsing experience for users by implementing DevOps methods[34]. This approach reduces downtime and maximizes system performance. The research findings suggest that the (KWSSWS) system has the capacity to greatly enhance online searches by utilizing Google PDFs. It allows users to get vital information from PDF documents that may not be readily available through conventional web search engines. The (KWSSWS) system enhances online searches by obtaining PDFs and extracting pertinent material, so offering consumers a complete and varied information retrieval experience. Nevertheless, there are certain obstacles and factors to take into account while implementing the (KWSSWS) system. First and foremost, the functionality of the system is dependent on the presence and ease of access to Google PDFs, which might potentially be constrained or subject to limitations in certain situations[35]. The researchers must guarantee that the system is intended to effectively manage diverse circumstances and seamlessly adjust to fluctuations in the accessibility and organization of Google PDFs. Moreover, the incorporation of DevOps principles necessitates meticulous strategizing, synchronization, and allocation of infrastructure and resources[36]. Organizations must build unambiguous communication channels between development and operations teams, implement automated deployment and testing procedures, and consistently monitor and optimize the system's performance. The research highlights the significance of integrating DevOps methods at the initial stages of developing the (KWSSWS) system to guarantee its scalability, stability, and efficiency.

V. CONCLUSION

This research paper explores the implementation of the (KWSSWS) system on Google PDF retrieval to enhance web searching and Development Operations (DevOps) processes. The results demonstrate the effectiveness of the (KWSSWS) system in improving search capabilities and streamlining PDF retrieval. The study provides valuable insights for further advancements in web searching and PDF retrieval techniques. In conclusion, the research on the (KWSSWS) system practices on Google PDFs retrieval to improve web searching Development Operations (DevOps)" highlights the significance of integrating the (KWSSWS) system with DevOps principles. The utilization of the (KWSSWS) system enhances web searching efficiency by leveraging keyword search and retrieving relevant information from Google PDFs. The adoption of DevOps practices ensures the seamless integration and continuous improvement of the KWSWS system. Future research can focus on further optimizing the (KWSSWS) system, exploring additional data sources, and evaluating its performance in real-world web searching scenarios.

REFERENCES

[1]W. J. Ripple *et al.*, "World scientists' warning of a climate emergency 2022." Oxford University Press, 2022.

[2]L. A. McGuinness and J. P. T. Higgins, "Risk- of- bias VISualization (robvis): an R package and Shiny web app for visualizing risk- of- bias assessments," *Res. Synth. Methods*, vol. 12, no. 1, pp. 55–61, 2021.

[3]R. Kinney *et al.*, "The semantic scholar open data platform," *arXiv Prepr. arXiv2301.10140*, 2023.

[4]A. Kumar, M. Nadeem, and M. Shameem, "Multicriteria decision- making–based framework for implementing DevOps practices: A fuzzy best–worst approach," *J. Softw. Evol. Process*, p. e2631.

[5]Q. Ai *et al.*, "Information retrieval meets large language models: a strategic report from chinese ir community," *AI Open*, vol. 4, pp. 80–90, 2023.

[6]C. Shah and E. M. Bender, "Envisioning information access systems: What makes for good tools and a healthy Web?," *ACM Trans. Web*, vol. 18, no. 3, pp. 1–24, 2024.

[7]N. F. Liu, T. Zhang, and P. Liang, "Evaluating verifiability in generative search engines," *arXiv Prepr. arXiv2304.09848*, 2023.

[8]M. Schaefer and G. Sapi, "Complementarities in learning from data: Insights from general search," *Inf. Econ. Policy*, vol. 65, p. 101063, 2023.

[9]H. H. Shang *et al.*, "Histopathology Slide Indexing and Search—Are We There Yet?," *NEJM AI*, vol. 1, no. 5, p. AIcs2300019, 2024.

[10] A. Henzinger, E. Dauterman, H. Corrigan-Gibbs, and N. Zeldovich, "Private web search with Tiptoe," in *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023, pp. 396–416.

[11] C. A. Yeung, I. Liccardi, K. Lu, O. Seneviratne, and T. Berners-Lee, "Decentralization: The future of online social networking," in *Linking the World's Information: Essays on Tim Berners-Lee's Invention of the World Wide Web*, 2023, pp. 187–199.

[12] S. P. Borgatti, F. Agneessens, J. C. Johnson, and M. G. Everett, "Analyzing social networks," 2024.

[13] R. Schober, *Spider Web, Labyrinth, Tightrope Walk: Networks in US American Literature and Culture*, vol. 82. Walter de Gruyter GmbH & Co KG, 2023.

[14] C. J. Lim and L. Angers, *Dreams+ Disillusions*. Taylor & Francis, 2024.

[15] Q. Qiu *et al.*, "Integrating NLP and Ontology Matching into a Unified System for Automated Information Extraction from Geological Hazard Reports," *J. Earth Sci.*, vol. 34, no. 5, pp. 1433–1446, 2023.

[16] V. K. Veparala and V. Kalpana, "Big Data and Different Subspace Clustering Approaches: From social media promotion to genome mapping," *Salud, Cienc. y Tecnol.*, vol. 3, p. 413, 2023.

[17] V. Shenmaier, "Linear-size universal discretization of geometric center-based problems in fixed dimensions," *J. Comb. Optim.*, vol. 43, no. 3, pp. 528–542, 2022.

[18] H. Esfandiari, A. Karbasi, V. Mirrokni, G. Velegkas, and F. Zhou, "Replicable clustering," *Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.

[19] T. M. Ghazal, "Performances of k-means clustering algorithm with different distance metrics," *Intell. Autom. Soft Comput.*, vol. 30, no. 2, pp. 735–742, 2021.

[20] K. P. Sinaga and M.-S. Yang, "Unsupervised K-means clustering algorithm," *IEEE access*, vol. 8, pp. 80716–80727, 2020.

[21] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, p. 1295, 2020.

[22] "No Title." http://facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/k-means.html).

[23] S. Kapil, M. Chawla, and M. D. Ansari, "On K-means data clustering algorithm with genetic algorithm," in *2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, IEEE, 2016, pp. 202–206.

[24] Ü. Demirbaga, G. S. Aujla, A. Jindal, and O. Kalyon, "Big Data Analytics Platforms," in *Big Data Analytics*, Springer, 2024, pp. 79–126.

[25] M. S. Husain, M. Z. Khan, and T. Siddiqui, *Big Data*

Concepts, Technologies, and Applications. CRC Press, 2023.

[26] A. K. Feroz, G. F. Khan, and M. Sponder, *Digital Analytics for Marketing*. Taylor & Francis, 2024.

[27] O. Llaha and A. Aliu, "Application of Data Visualization and Machine Learning Algorithms for Better Decision Making.," in *RTA-CSIT*, 2023, pp. 97–102.

[28] B. Hamoud and E. Atwell, "Quran question and answer corpus for data mining with WEKA," in *2016 Conference of Basic Sciences and Engineering Studies (SGCAC)*, IEEE, 2016, pp. 211–216.

[29] K. P. S. Attwal and A. S. Dhiman, "Exploring data mining tool-Weka and using Weka to build and evaluate predictive models," *Adv. Appl. Math. Sci.*, vol. 19, no. 6, pp. 451–469, 2020.

[30] M. Vijayakamal and M. Narendhar, "A Novel Approach for WEKA & Study On Data Mining Tools," *Int. J. Eng. Innov. Technol. Vol.*, vol. 2, 2012.

[31] K. Aslett, Z. Sanderson, W. Godel, N. Persily, J. Nagler, and J. A. Tucker, "Online searches to evaluate misinformation can increase its perceived veracity," *Nature*, vol. 625, no. 7995, pp. 548–556, 2024.

[32] R. A. Saputra, "Economic Improvement, Environmental Sustainability, and Community Empowerment in Indonesia: Bibliometric Analysis (Smart City and Smart Tourism) Year 2013-2022," in *E3S Web of Conferences*, EDP Sciences, 2023, p. 1006.

[33] K. St. Amant and W. Giordano, "Expanding Communication Expectations: Examining Audience Understanding of Scripts Through Fold and Swap Strategies," *J. Tech. Writ. Commun.*, p. 00472816231216911, 2023.

[34] G. Kim, J. Humble, P. Debois, J. Willis, and N. Forsgren, *The DevOps handbook: How to create world-class agility, reliability, & security in technology organizations*. IT Revolution, 2021.

[35] A. F. Schiopu, R. I. Hornoiu, A. M. Padurean, and A.-M. Nica, "Constrained and virtually traveling? Exploring the effect of travel constraints on intention to use virtual reality in tourism," *Technol. Soc.*, vol. 71, p. 102091, 2022.

[36] N. Azad and S. Hyrynsalmi, "DevOps critical success factors—A systematic literature review," *Inf. Softw. Technol.*, vol. 157, p. 107150, 2023.