Bibliometric Insights into Web Scraping and Advanced AI-Based Models for Valuable Business Data

Barba Giuliana[a], Lazoi Mariangela[b] and Lezzi Marianna[c]
Department of Engineering for Innovation, University of Salento, Campus Ecotekne Via Monteroni, Lecce 73100, Italy

Keywords: Web Scraping, Artificial Intelligence, Natural Language Processing, Business Data Analysis, Sentiment Analysis.

Abstract: The integration of advanced Artificial Intelligence (AI) based models with web scraping technique opens new opportunities for businesses, streamlining the extraction of valuable insights from the huge amounts of online data. This integration is strategic in overcoming the challenges of extracting dirty data and retrieving missing information, which could otherwise compromise the reliability of business decisions. Despite the growing importance of integrating AI-based models and web scraping techniques in the business context, there exists a significant gap in understanding the specific implications. To address this gap, our study uses a systematic literature review (SLR) and bibliometric analysis to examine the implications of the combined use of advanced AI-based models and web scraping in business contexts. The study highlights four distinct clusters that suggest potential research areas in the areas of "Machine Learning (ML) for sentiment analysis", "Artificial Intelligence and Natural Language Processing (NLP) integration", "Data intelligence and optimization", "NLP and Deep Learning (DL) integration". The paper offers both theoretical and practical contributions, providing a clear overview of emerging research directions in the field of AI-based models and web scraping integration and guiding managers in adopting advanced AI-based models to enhance the value of web data obtained through scraping.

1 INTRODUCTION

Online data are a crucial tool for knowledge generation and knowledge-based decision support (Rejeb et al., 2020). In particular, online data enable companies to consolidate information by transforming it into useful insights for marketing and service decisions, such as optimising pricing decisions based on consumer behaviour (Jorge et al., 2020) or segmenting customers based on their perceptions (Rejeb et al., 2020). Moreover, gathering online data not only facilitates a comprehensive analysis of web users' perceptions of corporate products and services (Bisconti et al., 2019), but also contributes positively to the enrichment of the company's information assets (M. A. Khder, 2021).

Companies that extract insights from online data prove to be more agile and more adaptive to market dynamics (Rejeb et al., 2020), highlighting the crucial importance of the strategic use of web data to drive business decisions and maintain a distinctive competitive position in the global landscape.

Web scraping is a valuable technique for automatic data collection from the internet (Tanasescu et al., 2022) about customer feedback (Bisconti et al., 2019) and sentiment about particular products (Jorge et al., 2020). However, web data from various online sources such as websites, blogs and social media is often unstructured and consists of a wide range of heterogeneous information such as text, images and video (Eberendu, 2016). To enhance the value of unstructured data, advanced Artificial Intelligence (AI) based models can extract hidden insights from news and opinions on the web providing useful in understanding customer attitudes (Chan et al., 2022). This information can help decision-making strategies (Tanasescu et al., 2022) and refining customer targeting (Rejeb et al., 2020).

[a] https://orcid.org/0009-0004-9138-2063
[b] https://orcid.org/0000-0003-4280-1597
[c] https://orcid.org/0000-0003-3526-8421

Giuliana, B., Mariangela, L. and Marianna, L.
Bibliometric Insights into Web Scraping and Advanced AI-Based Models for Valuable Business Data.
DOI: 10.5220/0012686900003690

In *Proceedings of the 26th International Conference on Enterprise Information Systems (ICEIS 2024) - Volume 1*, pages 321-328
ISBN: 978-989-758-692-7; ISSN: 2184-4992

The implementation of advanced AI-based models, together with the web scraping, presents novel opportunities for companies as it facilitates the management of vast amounts of data from the internet with the objective of extracting valuable insights (Arjunan, 2022). This integration is particularly important for overcoming the challenges of mining dirty data and retrieving missing information that could otherwise affect the reliability of business decisions (Tanasescu et al., 2022). Several studies reflect the growing interest in integrating advanced models based on artificial intelligence and web scraping. For instance, (Kumar et al., 2021) focus on estimating the relevance of online documents; while, (Sahu et al., 2022) address the evaluation of customer reviews in the context of e-commerce. Moreover, (Arjunan, 2022) study highlights the potential of integrating advanced artificial intelligence-based models to enrich web data. However, the lack of analysis on implications of integrating AI-based models and web scraping in business contexts reveals a gap that needs to be filled.

Through a Systematic Literature Review (SLR) and bibliometric analysis, the objective of this study is to investigate the current key research directions in the combined use of advanced AI-based models and web scraping applied in business contexts. This promises to highlight the main implications in this emerging field and to help explore a new frontier in the literature, which still seems to be developing. In particular, the paper aims to provide an answer to the following research question: "What are the implications of applying web scraping integrated with advanced AI-based models in the business context?". To answer this research question, a network examination and visual analysis of textual bibliographic data acquired from Scopus database were conducted, with the aim of creating a map of interconnections using VOSviewer (a software tool for bibliometric analysis).

2 BACKGROUND

2.1 Web Scraping and Advanced AI-Based Models for Business

Web scraping - also known as web crawling (M. A. Khder, 2021), web harvesting (Zhao, 2017), web data extraction (Zhao, 2017), web data scraping (Barbera et al., 2023), or screen scraping (Arjunan, 2022) - is an advanced technique for systematically extracting unstructured data from websites and subsequently transforming them into structured data to be stored in a file or database (M. A. Khder, 2021s).

The significant advantage of web scraping lies in the automation of the process of searching and extracting data, eliminating the need to manually copy information from a website to a file (M. A. Khder, 2021; Tanasescu et al., 2022). Moreover, in the context of data science, web scraping emerges as a tool of considerable scientific interest, recognized as an efficient method for collecting big data (Barbera et al., 2023), thanks to low execution and maintenance costs (M. A. Khder, 2021).

The applications of web scraping span various sectors, significantly contributing to the business and marketing world. In particular, web scraping is widely employed in the analysis of online user feedback, such as for the support of event management (Bisconti et al., 2019), the prediction of consumers ‘choices (Corallo et al., 2020), the monitoring changes in product prices (Jorge et al., 2020) or to investigate employee opinions (Tanasescu et al., 2022). Furthermore, in the job search engine sector, there are several applications of web scraping to define job profiles (De Mauro et al., 2018) and conduct labour market surveys (Vankevich & Kalinouskaya, 2021). On the other hand, AI has revolutionised the business landscape, by enabling the use of models capable of handling an enormous stream of heterogeneous data identifying hidden patterns or trends in large datasets (Afandizadeh et al., 2023). The application of AI-based models is increasingly significant in analysing online data (such as customer reviews or sales information) to detect behavioural patterns and identify preferences (Chan et al., 2022).

The application of these models allows companies to delineate detailed customer segmentation strategies, opening the door to personalised marketing campaigns, individualised product suggestions and customer experiences tailored to specific needs (Afandizadeh et al., 2023). Therefore, companies can gain extensive and contextual knowledge from online data, enabling them to make informed decisions and develop precise business strategies (Chan et al., 2022).

AI-based models are pivotal in audio, video, image and text processing, significantly contributing to the optimisation of diverse business operations and demonstrating how innovative technologies have been transformed into practical commercial solutions.

For instance, (Arslan & Cruz, 2022) use an AI-based model to exploit large corpora of textual documents while (Sarica et al., 2020) propose the TechNet model, demonstrating how AI can be used to extract relational knowledge from complex patent documents or also to find optimal candidates facilitating recruitment process (Sridevi & Suganthi, 2022).

AI-based models are also used to analyse images for hot rolling mill process defect detection (Latham & Giannetti, 2023) or to improve the performance of harvesting robots (Tang et al., 2020).

2.2 Integration of Web Scraping and Advanced AI-Based Models in the Business Domain

The combination of advanced AI-based models and web scraping presents a modern strategy for obtaining high quality data from the web by optimizing the data collection process, eliminating redundant information and ensuring the delivery of more relevant and clean data (Arjunan, 2022).

The application of this integration plays a key role in providing crucial information for understanding consumer sentiment, thus helping to drive informed decisions in the business environment (Sahu et al., 2022).

The synergies of web scraping and AI-based models integration are evident in various contexts, as shown by (Hao et al., 2023), which extract vulnerable online contracts using web scraping, and trained an AI-based model to enhance the security of smart contracts. (Thuan et al., 2022) use this integration to simplify the assignment of professional roles based on employee competencies. Furthermore, (Kinne & Resch, 2018) propose an AI-based model that assesses the companies' degree of digital innovation by analysing web scraped data from their websites.

The joint use of AI-based models for sentiment and opinion recognition, together with web scraping, is another demonstration of how AI can enhance the value of data collected from online posts and reviews (M. A. Khder, 2021). A tangible example is provided by (Kafeza et al., 2023), which exploit these data from social media to model customers' actions in order to discover their behavioural patterns. Additionally, (Tanasescu et al., 2022) collect data via web scraping on employee feedback and integrate it into an AI-based model to facilitate decision-making processes based on the opinions of company employees.

Moreover, other studies show that AI-based models improve the accuracy of web scraping by providing more relevant information and cleaner data. In particular, (Reddy et al., 2021) present an AI-based model that supports the scraping process by summarizing meaningful insights of online documents. (Arjunan, 2022) show that AI-based model can be specialized in recognising the relevance of web and in eliminating superfluous information. On the other hand, (M. Lee & Na, 2023) use this integration to select relevant companies information by retrieving customised market data from the web; whereas (Sahu et al., 2022) propose an AI-based model to evaluate the honesty of online scrapped reviews.

The combination of AI-based models and web scraping holds promise, but poses significant challenges. First, further exploration in the field of AI-based models is necessary to improve the performance and accuracy of results as the current implementation is in its early stages of development (Sahu et al., 2022; Tanasescu et al., 2022). Moreover, the malleability of website design and layout presents a significant obstacle, as scripts can quickly become obsolete, and adapting them to hundreds of sites is nearly impossible (Patnaik & Babu, 2021).

Furthermore, the analysis of sentiments poses even greater challenges, as it involves identifying complex phenomena like negation or irony in the analysed texts (Tanasescu et al., 2022).

3 METHODOLOGY

The research strategy was designed to gain the current key research directions in the combined use of advanced AI-based models and web scraping for business applications. In particular, the research aims to address the key question: "What are the implications of applying web scraping integrated with advanced AI-based models in the business context?".

To address this question, a SLR was conducted as the main methodology for data collection. This process, known for its transparency, scientific rigor, and replicability, allows for the identification, highlighting, and evaluation of various sources of information, facilitating the cataloguing and structured comparison of results (Del Vecchio et al., 2023; Tranfield et al., 2003).

Additionally, a bibliometric analysis was conducted to examine the collected sample and gain an exploratory understanding of the main themes of interest. Bibliometric analysis, introduced by (Pritchard, 1969), is recognized as a crucial methodology for exploring research in various disciplines, highlighting its nature, and providing particularly valuable insights in fields still in development (Donthu, Kumar, Mukherjee, et al., 2021). As suggested by (Del Vecchio et al., 2023), the research process is structured into three main phases: definition of the search framework and data sample, preliminary analysis of the sample, and data analysis.

3.1 Definition of Search Schema and Data Sample

In line with the objective of this study and following (Corallo et al., 2021) SRL procedure, the keywords were identified for the investigation of the fields of interest: web scraping, artificial intelligence, text mining and business.

The choice of these keywords was based on the consideration of similar concepts, synonyms, and acronyms identified in the theoretical context. The combination of these keywords was implemented using accurate mathematical logical connectors (Boolean and Proximity operators) as it is provided in Table 1.

Table 1: Query structure used for Scopus.

SCOPUS
TITLE-ABS-KEY
("web") W/2 ("scrap*" OR "craw*" OR "harvest*" OR "data extract*" OR "data harvest*" OR "data scrap*" OR "automat* scrap*" OR "information extract*")) OR (“screen scrap*" OR "automat* web scrap*" OR "spider*"))
AND
("Machine learning" OR "ML" OR "artificial intelligence" OR "AI" OR "natural language process*" OR "NLP" OR "natural language generat*" OR "NLG" OR "artificial neural network" OR "ANN" OR "neural network" OR "deep neural network" OR "DNN" OR "recurrent neural network" OR "RNN" OR "unsupervised learn*" OR "convolutional neural network" OR "CNN" OR "supervised learn*" OR "learning algorithm" OR "convolutional neural network" OR "CNN" OR "SVM" OR "support vector machine")
AND
("text data" OR "Text mining" OR "data mining" OR "web mining" OR "web data mining" OR "knowledge discovery from text" OR "KDT" OR "information extraction" OR "IE" OR "information retrieval" OR "IR" OR "data enrich*" OR "data augment*" OR "data extract*" OR "data collect*" OR "semantic analysis" OR "sentiment analysis" OR "sentiment recognition" OR "text analysis" OR "data analysis" OR "information retrieval" OR "IR" OR "big data" OR "Text Summarization" OR "Text Generation" OR "Market Intelligence" OR "Predictive Analytics" OR "Reputation Monitoring")
AND
("Business" OR "firm" OR "enterprise" OR "company" OR "organization" OR "manufactur*" OR "industry" OR "market*" OR "account*" OR "sales" OR "decision making" OR "knowledge manag*" OR "strategic planning" OR "financial" OR "HRM" OR "human resource manag*" OR "SCM" OR "supply chain manag*" OR "risk management" OR "retail" OR "CRM" OR "customer relationship manag*" OR "insurance")

The query used for Scopus (Table 1) was modified in the field tag and proximity operator to adapt for Web of Science (WOS).

Papers containing the keywords in the title, abstract, and keywords were searched in the Scopus and Web of Science electronic scientific database in December 2023, chosen for their robustness as bibliographic data sources (Pranckutė, 2021). Subsequently the initial sample of 455 and 129 articles, respectively from Scopus and WOS, was refined to 459 unique by setting English language limitation as the only filter.

All results were exported in RIS format, retaining all necessary information for conducting the subsequent analysis, including title, year, abstract, and keywords.

3.2 Preliminary Analysis of the Sample

The second step of the adopted methodology focuses on the statistical analysis through MS Excel of metadata associated with the selected papers. Initially, we explored the quantity of publications over the years (Figure 1). The graph reveals that from 1998 to 2008, the production of studies in this field is negligible. However, from 2009 to 2017, a gradual increase occurs, indicating a growing interest in data management and its connection to the web within the scientific community. It's interesting to note that from 2017 to 2023, there is a rapid growth in the number of studies related to this domain, with a remarkable growth rate of 100% over six years.

Additionally, the distribution of papers based on their respective disciplinary areas was examined. "Computer Science" emerges as the predominant area, representing 36.5% of the total articles published on this subject. Following in order are the disciplinary areas of "Engineering" (15.4%), and "Decision Science" (10.3%), confirming the multidisciplinary nature and broad scope of web scraping and AI-based models in the business context.

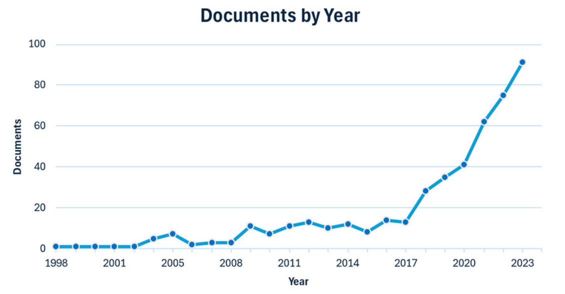

Figure 1: Distribution of papers by year.

3.3 Data Analysis

In the third phase of this research methodology, an in-depth bibliometric analysis has been chosen as a suitable method for macro-level evaluations of the 459 sample of selected papers. Additionally, since it is not influenced by researcher subjectivity, it can help to decrease reviewer bias during research (van Oorschot et al., 2018).

In the context of this study, bibliometric analysis played a crucial role in delineating implications and future research directions related to combined use of AI-based models and Web Scraping within business applications.

Specifically, through this analysis, the objectives were to: i) comprehend the most recurring topics in the fields of analysis; ii) identify relationships among them; iii) highlight emerging themes and trends over time. Network analysis and graphical exploration of textual data were conducted using VOSviewer software, known for its ability to create easily interpretable co-occurrence maps on a large scale (van Oorschot et al., 2018). Following the guidelines of (Donthu, Kumar, Mukherjee, et al., 2021), the analysis focused on counting recurring terms in title, keywords and abstracts, with a minimum frequency of 10. By carrying out a preliminary merge of the extracted keywords and creating a specific thesaurus, out of the 3658 terms examined, 67 meet the threshold, contributing to a clear definition of conceptual relationships within papers keywords.

4 EXPLORING FUTURE RESEARCH DIRECTIONS IN THE COMBINED USE OF WEB SCRAPING AND ADVANCED AI-BASED MODELS

For the data analysis phase, VOSviewer offers three visualizations, namely the network, overlay, and density visualizations which are commented in next sections.

4.1 Network Visualization

Bibliometric analysis enables a network analysis that relies on term co-occurrence. The co-occurrence map, presented in Figure 2, clearly displays four main clusters generated by binding a minimum of one item per cluster.

Cluster 1 (red), comprising 22 terms, focuses on the implementation of “Machine Learning (ML) for sentiment analysis”, with an emphasis on big data, learning algorithms, web scraping and online social networks. Additionally, the topics cover aspects of financial markets, risk assessment, e-commerce, investment, data analysis, and human impacts. This implies an interest in comprehending human and market dynamics using sophisticated data analysis techniques.

Cluster 2 (green), consisting of 19 terms, addresses the “AI and NLP integration” to optimise the management and analysis of online information. Key topics include classifying information, semantics, search engines, extracting web information and managing knowledge, with a specific emphasis on ontologies and semantic analysis.

Cluster 3 (blue), with 16 terms, focuses on “Data intelligence and optimization” to improve the extraction and management of information through data mining and text mining approaches. This indicates an emphasis on improving data collection and manipulation operations, alongside decision-making systems.

Cluster 4 (yellow), consisting of 10 terms, focuses on the “NLP and DL integration”, such as deep neural networks and convolutional neural networks. This cluster highlights an interest in advanced natural language comprehension to facilitate informed decision-making, as well as emphasizing the analysis of opinions through deep learning techniques.

Moreover, the position of the largest nodes in Figure 2 confirms that the extraction and analysis of data, both structured and unstructured, are crucial aspects for research in the considered field. Furthermore, the presence of these central nodes suggests that technologies, such as web scraping, data mining and text mining, are strongly integrated with big data and machine learning, indicating an emphasis on extraction and analysis of web data as an integral part of the decision-making process and knowledge generation.

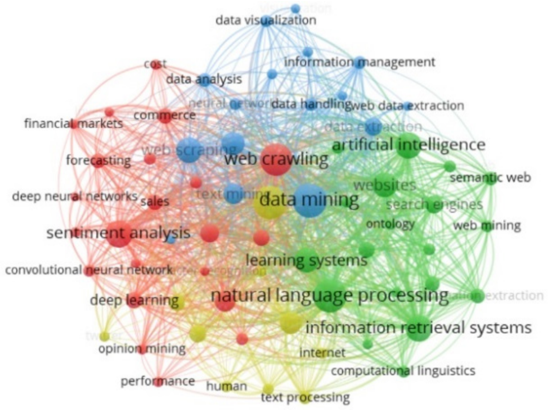

Figure 2: Network visualization.

4.2 Overlay Visualization

The overlay map (Figure 3), a variant of the co-occurrence map, broadens the temporal perspective by introducing a distinctive chronological dimension. The use of distinct colours highlights the older terms in blue, intermediate terms in green, and more recent terms in yellow. This temporal representation spans from 2014 to 2022. The choice of this time interval is crucial as it provides an optimal balance for clearly perceiving variations in terms over the years.

It is interesting to note that the older terms, covering the period from 2014 to 2017, such as information retrieval, information extraction, semantic web, and AI, reflecting the early stages of development when the focus was on information organisation and extraction.

Over the interim period from 2017 to 2019, there has been a move towards more research in NLP, learning systems, web crawling and data mining. This suggests an awareness of the increasing significance of online information and a drive to create more sophisticated systems to handle it.

Finally, the recent terms, from 2019 to the present, reflect the ongoing development of technologies and areas of interest. Terms such as web scraping, deep learning, sentiment analysis, social media, and convolutional neural networks highlight a greater emphasis on advanced data analysis and the application of advanced technologies across various domains (i.e. sales, financial, risk assessment), encompassing those associated with online social dynamics.

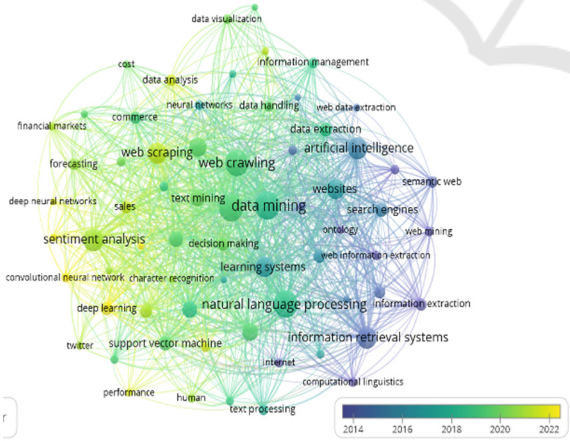

Figure 3: Overlay visualization.

4.3 Density Visualization

Density visualization provides a perspective on the distribution and relevance of terms within the extensive range of analysed papers (Figure 4). Each identified label and the surrounding area, takes on a colour and intensity that reflect the density of papers in that specific position.

The presence of key terms such as web crawling, data mining, AI, NLP systems, ML, sentiment analysis, big and learning systems highlights points of dense concentration where scientific research has significantly focused. In these clusters, the colour tends to shift towards dark blue, indicating a higher density and suggesting that these topics are at the centre of academic debate.

Conversely, terms like deep learning, sales, and knowledge management, while important, exhibit a more widespread distribution characterized by lighter shades of yellow, suggesting a lower concentration in specific research areas compared to key terms. This could indicate that these concepts are addressed in various contexts and sectors without a specific focus in a single research area.

The keyword density highlights central and current topics in scientific analysis, reflecting a continued interest in crucial areas such as NLP, ML and data mining from web sources, with a keen eye on new developments and emerging challenges.

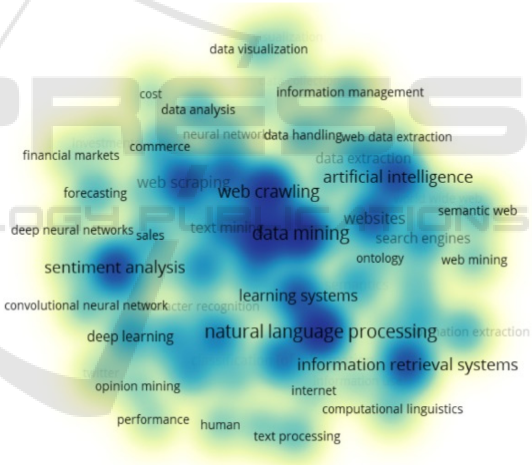

Figure 4: Item density visualization.

5 CONCLUSIONS

The bibliometric analysis conducted offers a detailed perspective on the main areas of research and development in the field of web scraping and AI-based models. The four clusters identified (“ML for sentiment analysis”, “AI and NLP integration, “Data intelligence and optimization”, “NLP and DL integration”) reflect different approaches of analysis and usage of online information. Several topics emerge from these clusters, such as the implementation of ML for sentiment analysis, the integration between

AI and NLP to optimize information management, the use of data mining and text mining techniques to improve data extraction and management, and the application of advanced technologies such as deep neural networks to understand and analyse complex online opinions. Furthermore, the temporal view highlights a progression in research focus over the years, with an increase in attention towards advanced data analysis and the application of these innovative technologies.

From an academic perspective, the identification of emerging research trends is one of the most relevant contributions. By analysing the temporal distribution of keywords, the evolution of topics of interest over time can be identified, providing academics with a clear overview of emerging research directions. Furthermore, the analysis provides insights into the intersections between different disciplines and research fields.

From a managerial perspective, this study provides greater clarity of key trends and themes in the field of web scraping and AI-based models, which can inform strategic decisions regarding investments in technologies to understand the online data and optimise their extraction and analysis.

However, this research has some limitations. The interpretation of the results could be affected by the subjectivity of the observers and their previous knowledge in the field, introducing potential biases into the analysis. Finally, bibliometric analysis may not fully capture the dynamism and complexity of the context such as socio-cultural and political factors that could influence research trends. Future research should integrate theoretical and empirical approaches to obtain a more complete and in-depth understanding of the dynamics and challenges in the field of web scraping combined with AI-based models, with the aim of contributing to the development of innovative and impactful solutions.

REFERENCES

Afandizadeh, S., Sharifi, D., Kalantari, N., & Mirzahossein, H. (2023). Using machine learning methods to predict electric vehicles penetration in the automotive market. Scientific Reports, 13(1), 8345. https://doi.org/10.1038/s41598-023-35366-3

Arjunan, T. (2022). Building Business Intelligence Data Extractor using NLP and Python. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 10(X).

Arslan, M., & Cruz, C. (2022). Semantic taxonomy enrichment to improve business text classification for dynamic environments. 2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), 1–6. https://doi.org/10.1109/INISTA55318.2022.9894173

Babu, S., Pragathi, B. S., Chinthala, U., & Maheshwaram, S. (2020). Subject Tracking with Camera Movement Using Single Board Computer. Proceedings of 2020 IEEE-HYDCON International Conference on Engineering in the 4th Industrial Revolution, HYDCON 2020, 0–5. https://doi.org/10.1109/HYDCON48903.2020.9242811

Barbera, G., Araujo, L., & Fernandes, S. (2023). The Value of Web Data Scraping: An Application to TripAdvisor. Big Data and Cognitive Computing, 7(3), 121. https://doi.org/10.3390/bdcc7030121

Bisconti, C., Corallo, A., Fortunato, L., & Spennato, A. (2019). Influence parameters correlation in a Twitter event network. Entrepreneurship and Small Business.

Chan, L., Hogaboam, L., & Cao, R. (2022). Artificial Intelligence for Business. https://doi.org/10.1007/978-3-031-05740-3_1

Corallo, A., Del Vecchio, V., Lezzi, M., & Morciano, P. (2021). Shop floor digital twin in smart manufacturing: A systematic literature review. Sustainability (Switzerland), 13(23). https://doi.org/10.3390/su132312987

Corallo, A., Fortunato, L., Spennato, A., Errico, F., & Pedone, A. (2020). Predicting the Consumer's Purchase Intention of Food Products. 2020 9th International Conference on Industrial Technology and Management (ICITM), 181–185. https://doi.org/10.1109/ICITM48982.2020.9080404

De Mauro, A., Greco, M., Grimaldi, M., & Ritala, P. (2018). Human resources for Big Data professions: A systematic classification of job roles and required skill sets. Information Processing and Management, 54(5), 807–817. https://doi.org/10.1016/j.ipm.2017.05.004

Del Vecchio, V., Lazoi, M., & Lezzi, M. (2023). Digital Twin and Extended Reality in Industrial Contexts: A Bibliometric Review. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 14218 LNCS. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43401-3_18

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(April), 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Eberendu, A. C. (2016). Unstructured Data: An overview of the data of Big Data. International Journal of Computer Trends and Technology, 38(1), 46–50. https://doi.org/10.14445/22312803/ijctt-v38p109

Hao, Z., Zhang, B., Mao, D., Yen, J., Zhao, Z., Zuo, M., Li, H., & Xu, C.-Z. (2023). A novel method using LSTM-RNN to generate smart contracts code templates for improved usability. Multimedia Tools and Applications, 82(27), 41669–41699. https://doi.org/ 10.1007/s11042-023-14592-x

Jorge, O., Pons, A., Rius, J., Vintró, C., Mateo, J., & Vilaplana, J. (2020). Increasing online shop revenues with web scraping: A case study for the wine sector. British Food Journal, 122(11), 3383–3401. https://doi.org/10.1108/BFJ-07-2019-0522

Kafeza, E., Rompolas, G., Kyriazidis, S., & Makris, C. (2023). Time-Series Clustering for Determining Behavioral-Based Brand Loyalty of Users Across Social Media. IEEE Transactions on Computational Social Systems, 10(4), 1951–1965. https://doi.org/10.1109/TCSS.2022.3219781

Khder, M. A. (2021). Web scraping or web crawling: State of art, techniques, approaches and application. International Journal of Advances in Soft Computing and its Applications, 13(3), 144–168. https://doi.org/10.15849/ijasca.211128.11

Kinne, J., & Resch, B. (2018). Generating Big Spatial Data on Firm Innovation Activity from Text- Mined Firm Websites. GI_Forum, 1, 82–89. https://doi.org/10.1553/giscience2018_01_s82

Kumar, S. A., Nasralla, M. M., García-Magariño, I., & Kumar, H. (2021). A machine-learning scraping tool for data fusion in the analysis of sentiments about pandemics for supporting business decisions with human-centric AI explanations. PeerJ Computer Science, 7, e713. https://doi.org/10.7717/peerj-cs.713

Latham, S., & Giannetti, C. (2023). A Tool to Combine Expert Knowledge and Machine Learning for Defect Detection and Root Cause Analysis in a Hot Strip Mill. SN Computer Science, 4(5), 628. https://doi.org/10.1007/s42979-023-02104-5

Lee, M., & Na, I. (2023). Enhancing Similar Business Group Recommendation through Derivative Criteria and Web Crawling. KSII Transactions on Internet and Information Systems, 17(10). https://doi.org/10.3837/tiis.2023.10.012

Patnaik, S. K., & Babu, C. N. (2021). Information Retrieval from web with Faster R-CNN Deep Learning Networks: A New Perspective. Proceedings - 6th International Conference on Computer Science and Engineering, UBMK 2021, i, 61–66. https://doi.org/10.1109/UBMK52708.2021.9558956

Pranckutė, R. (2021). Web of science (Wos) and scopus: The titans of bibliographic information in today's academic world. Publications, 9(1). https://doi.org/10.3390/publications9010012

Pritchard, A. (1969). Statistical Bibliography or Bibliometrics? Journal of Documentation, 25, 348–349.

Reddy, K. K. C., Anisha, P. R., Nguyen, N. G., & Sreelatha, G. (2021). A Text Mining using Web Scraping for Meaningful Insights. Journal of Physics: Conference Series, 2089(1). https://doi.org/10.1088/1742-6596/2089/1/012048

Rejeb, A., Rejeb, K., & Keogh, J. G. (2020). Potential Of Big Data For Marketing: A Literature Review. Management Research and Practice, 12(3).

Sahu, S., Divya, K., Rastogi, D. N., Yadav, P. K., & Perwej, D. Y. (2022). Sentimental Analysis on Web Scraping Using Machine Learning Method.

Sarica, S., Luo, J., & Wood, K. L. (2020). TechNet: Technology semantic network based on patent data. Expert Systems with Applications, 142, 112995. https://doi.org/10.1016/j.eswa.2019.112995

Sridevi, G. M., & Suganthi, S. K. (2022). AI based suitability measurement and prediction between job description and job seeker profiles. International Journal of Information Management Data Insights, 2(2), 100109. https://doi.org/10.1016/j.jjimei.2022.10 0109

Tanasescu, L. G., Vines, A., Bologa, A. R., & Vaida, C. A. (2022). Big Data ETL Process and Its Impact on Text Mining Analysis for Employees' Reviews. Applied Sciences, 12(15), 7509. https://doi.org/10.3390/app121 57509

Tang, Y., Chen, M., Wang, C., Luo, L., Li, J., Lian, G., & Zou, X. (2020). Recognition and Localization Methods for Vision-Based Fruit Picking Robots: A Review. Frontiers in Plant Science, 11(May), 1–17. https://doi.org/10.3389/fpls.2020.00510

Thuan, N. D., Nhut, N. M., Quan, D. M., & Khanh, L. M. D. (2022). Using Blockchain and Artificial Intelligence to build a Job Recommendation System for Students in Information Technology. 2022 RIVF International Conference on Computing and Communication Technologies (RIVF), 364–369. https://doi.org/10.11 09/RIVF55975.2022.10013916

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. British Journal of Management, 14(3), 207–222. https://doi.org/10.1111/1467-8551.00375

van Oorschot, J. A. W. H., Hofman, E., & Halman, J. I. M. (2018). A bibliometric review of the innovation adoption literature. Technological Forecasting and Social Change, 134(June), 1–21. https://doi.org/10.1016/j.techfore.2018.04.032

Vankevich, A., & Kalinouskaya, I. (2021). Better understanding of the labour market using Big Data. Ekonomia i Prawo, 20(3), 677–692. https://doi.org/10.12775/eip.2021.040

Zhao, B. (2017). Web Scraping. In L. A. Schintler & C. L. McNeely (A c. Di), Encyclopedia of Big Data.