Ciencia & Turismo

ISSN: 3028-8665
Vol. 3 No. 4. pp. 23 – 40, octubre – diciembre 2024
Revista multidisciplinar

www.cienciayturismo.org

Machine Learning applications in bibliometrics. Scopus bibliometric analysis of consumer behavior

Aplicaciones del Machine Learning en bibliometría. Análisis bibliométrico Scopus del comportamiento del consumidor.

[1] Emilio Enrique Guerra Castellón
Investigador independiente
emilito042@gmail.com
https://orcid.org/0009-0005-2436-7186

[2] Edgar Núñez Torres
Investigador independiente
enunez8609@gmail.com
https://orcid.org/0000-0003-3354-8024

[3] Yasser Vázquez Alfonso
Investigador independiente
yalfos1@gmail.com
https://orcid.org/0000-0002-4074-0711

Artículo de Investigación Científica y Tecnológica
Enviado: 11/07/2024
Revisado: 19/08/2024
Aceptado: 10/09/2024
Publicado:05/10/2024
DOI: **https://doi.org/10.33262/ct.v3i4.51**

Cítese: Guerra Castellón, E. E., Núñez Torres, E., & Vázquez Alfonso, Y. (2024). Aplicaciones del Machine Learning en bibliometría. Análisis bibliométrico Scopus del comportamiento del consumidor. Ciencia & Turismo, 3(4), 23-40. https://doi.org/10.33262/ct.v3i4.51

CIENCIA & TURISMO, *es* una revista multidisciplinaria, **t**rimestral, que se publicará en soporte electrónico tiene como misión contribuir a la formación de profesionales competentes con visión humanística y crítica que sean capaces de exponer sus resultados investigativos y científicos en la misma medida que se promueva mediante su intervención cambios positivos en la sociedad.
La revista es editada por la Editorial Ciencia Digital (Editorial de prestigio registrada en la Cámara Ecuatoriana de Libro con No de Afiliación 663) www.celibro.org.ec

Ciencia & Turismo

ISSN: 3028-8665
Vol. 3 No. 4. pp. 23 – 40, octubre – diciembre 2024
Revista multidisciplinar

www.cienciayturismo.org

Palabras claves:
Análisis bibliométrico, machine Learning, comportamiento del consumidor, aplicaciones, bibliometría.

Resumen

Este artículo explora las aplicaciones del aprendizaje automático en bibliometría, específicamente aplicando este conocimiento a un estudio bibliométrico del comportamiento del consumidor. El crecimiento exponencial de las publicaciones científicas ha superado las capacidades de las herramientas bibliométricas tradicionales, creando la necesidad de técnicas más avanzadas para analizar grandes cantidades de datos. El aprendizaje automático, una rama de la inteligencia artificial, permite a las computadoras aprender de los datos y hacer predicciones sin instrucciones explícitas. El estudio utiliza una metodología bibliométrica descriptiva, utilizando 2563 artículos de Scopus y algoritmos de aprendizaje automático como Latent Dirichlet Allocation (LDA) para la clasificación de artículos. Se identificaron cinco temas principales: investigación de mercado, influencia en línea en el consumo de alimentos y la percepción de la marca, implicaciones sociales del consumo, implicaciones para la salud y la relación entre el comportamiento de compra y el marketing. Los resultados revelan una gran red de colaboración internacional con Estados Unidos y el Reino Unido a la cabeza y autores clave como Rifkin, Almotairi, Gruber, Kunz, Bonnevie, Buil, Hieke, Lal, Wang y Carlson. El aprendizaje automático se perfila como una herramienta clave para la bibliometría, permitiendo automatizar procesos, identificar patrones complejos y mejorar la eficiencia del análisis de grandes bases de datos. Este enfoque innovador facilita la comprensión de las tendencias y el desarrollo de conocimientos sobre el comportamiento del consumidor.

Keywords:
Bibliometric analysis, machine Learning, consumer behaviour, applications, bibliometrics.

Abstract

This article explores the applications of machine learning in bibliometrics, specifically by applying this knowledge to a bibliometric study of consumer behaviour. The exponential growth of scientific publications has outstripped the capabilities of traditional bibliometric tools, creating a need for more advanced techniques to analyse large amounts of data. Machine learning, a branch of artificial intelligence, allows computers to learn from data and make predictions without explicit instruction. The study uses a descriptive bibliometric methodology, using 2563 articles from Scopus and machine learning algorithms such

Ciencia & Turismo
ISSN: 3028-8665
Vol. 3 No. 4. pp. 23 – 40, octubre – diciembre 2024
Revista multidisciplinar
www.cienciayturismo.org

as Latent Dirichlet Allocation (LDA) for article classification. Five main topics were identified: market research, online influence on food consumption and brand perception, social implications of consumption, health implications, and the relationship between shopping behaviour and marketing. The results reveal a large international collaborative network with the US and UK leading the way and key authors including Rifkin, Almotairi, Gruber, Kunz, Bonnevie, Buil, Hieke, Lal, Wang and Carlson. Machine learning is emerging as a key tool for bibliometrics, allowing processes to be automated, complex patterns to be identified and the efficiency of analysis of large databases to be improved. This innovative approach facilitates the understanding of trends and the development of knowledge about consumer behaviour.

1. Introduction

Bibliometric analysis has established itself as an essential tool for studying the dynamics of scientific production and the evolution of knowledge (Guerra *et al.*, 2024). However, the exponential growth in the volume of scientific publications has outstripped the capabilities of traditional bibliometric tools, creating a need for more advanced techniques to manage and analyse large volumes of data.

Machine learning, a subfield of artificial intelligence (AI), focuses on developing algorithms that allow computers to learn from data, make predictions and make decisions without specific instructions (Sarker, 2021). These capabilities are beneficial in bibliometric analysis, where the volume of information to be processed requires automation and advanced accuracy.

The integration of machine learning into scientometrics has revolutionized the way researchers manage and explore large databases. This approach has made it possible to automate key processes such as document classification, the identification of emerging trends and the analysis of scientific collaboration networks, improving efficiency and the ability to identify patterns that were previously undetectable.

In recent years, there has been a remarkable increase in the scientific production of bibliometric analyses related to consumer behaviour, driven by the interest in understanding this area from a quantitative perspective. Although several techniques and

computational tools have been used, the application of machine learning has not yet been extensively explored, which represents an opportunity to overcome the current limitations in identifying complex and emerging patterns.

Important studies, such as Saravia *et al.* (2022), have examined consumer behaviour between 2011 and 2021. This analysis is carried out in the Scopus database and is notable for its comprehensiveness, identifying 1,258 relevant documents. They used R Studio software with the Bibliometrix package. They note that the most emerging topics related to the subject matter are sales, decision-making, trade, surveys, purchasing, e-commerce, sustainable development, structural equation modelling and the theory of planned behaviour. Pérez *et al.* (2021) focus specifically on tourism consumer behaviour. They present a bibliometric analysis of articles published between 2009 and 2018 in journals indexed in Scopus, including a sample of 330 articles. They used EndNote X11 bibliographic manager, Excel, VOSviewer r 1.6.8 and NVIVO 12 for data processing. They highlight that the most productive authors according to Lotka's law are Johan Bruwer and Rob Law. The results of the content analysis showed that there are five main nodes into which the research can be grouped: consumer behaviour, travel, marketing, destination and experience, where topics related to the purchase decision process, brand loyalty, use of technology and sustainability are addressed. The intersection between neuroscience and consumer behaviour is also an area of growing interest. Barrera *et al.* (2022) conducted a bibliometric analysis of papers published between 2007 and 2021, identifying and analyzing three research perspectives: advertising and its emotional impact; brands and their quest to persuade consumers; emotions and their influence on the brain and consumer behaviour. In addition, Camarena and Romero's (2023) study on sustainable and responsible consumption, through a bibliometric analysis of articles published between 2000 and 2020, reveals an increase in scientific production on these concepts and highlights a paradigm shift from responsible consumption to sustainable consumption.

Given the growing volume of scientific data and the need to optimize bibliometric analysis, this research is justified in that machine learning offers advanced tools to meet these new demands. While previous studies have applied traditional bibliometric analysis to consumer behaviour, none have explored the potential of machine learning to automate document classification, identify emerging patterns and improve the efficiency of large database analysis. Therefore, this study aims to fill this gap by first reviewing the main applications of machine learning applied to bibliometric analysis of consumer behaviour from 2014 to 2024, based on the Scopus database.

Machine Learning. Review of applications in bibliometrics

Today, we live in the data age, where everything around us is connected to a data source and everything is digitally recorded (Cao, 2020). For example, today's electronic world

has a large number of different types of data, such as Internet of Things (IoT) data, cybersecurity data, smart city data, commercial data, smartphone data, social media data, health data, bibliographic data, and many more. Data can be structured, semi-structured or unstructured.

Data types and their role in bibliometrics

Structured data is data that is organized in rows and columns to facilitate processing, searching and analysis. Typical examples are databases that store records in tables, such as those generated by bibliometric platforms such as Scopus, Dimensions and Web of Science. These data are ideal for processing by supervised learning algorithms such as decision trees, random forest and gradient boosting, thanks to their compatibility with libraries and tools such as Pandas, NumPy, Scikit-learn and TensorFlow.

On the other hand, semi-structured data, such as JSON or XML, which have a flexible structure, and unstructured data, such as free text or PDF documents, present greater analysis challenges due to their variability in format and content. However, artificial intelligence (AI) and machine learning (ML) have evolved to efficiently process this data, allowing applications to organize, analyse and automatically learn from the information (Sarker *et al.*, 2021).

Machine learning and its application in bibliometrics

Machine learning, in particular, has emerged as a powerful tool for bibliometric analysis. Several techniques that can be divided into supervised, unsupervised, semi-supervised and reinforcement learning (Mohammed *et al.*, 2016). In addition, deep learning, a recent and popular technique, has shown great potential in identifying complex patterns in large amounts of data (Janiesch, 2021). Its main applications in bibliometrics include:

(1) document classification, i.e. automatically assigning articles to specific categories or topics. The supervised learning algorithms used, as in Junior *et al.* (2012), are Naïve Bayes, k-nearest neighbours (k-NN) and support vector machine (SVM), which were used in a clustering method to achieve greater accuracy, precision and coverage than traditional methods alone, and to categorize production engineering articles efficiently and automatically. Another such method is random forest, which was used by Garcia (2018) to perform sensitivity analysis in bibliometric models. This approach allows the importance of different variables in the classification of documents to be assessed, using simulated and real data. The research highlights how the algorithm improves model accuracy by avoiding problems such as overfitting, which is crucial when analyzing large amounts of data. Unsupervised learning algorithms used for this purpose are Latent Dirichlet Assignment (LDA) in studies such as Urbizagastegui (2021), which analyses Brazilian bibliometrics. The author notes that, when faced with the situation common to

researchers of finding a huge collection of documents, the first impulse is to organize and group them. To do this, they use this algorithm, which uses complex mathematics and helps in the processes of automatic text classification. With the help of RStudio, they were able to classify the literature into five themes with the most recurrent words and to identify the authors.

(2) Document clustering, which, unlike classification, groups articles according to their thematic similarity without using labels. Algorithms such as K-means and Latent Dirichlet Allocation (LDA) have been used in Fuentealba (2023), in which they obtained 19 clusters on the literature associated with VUCA, validating them with the coherence index and silhouette, respectively. Other techniques include density-based spatial clustering of applications with noise, K-medoids and hierarchical clustering (Vasquez *et al.*, 2019).

(3) Citation prediction is another of the most relevant applications of machine learning, as it allows us to know how many citations an article will receive in the future based on historical data, which is very relevant for researchers to know which high-impact, current and relevant topics to choose to develop their research. Methods such as linear regression (Cabeza, 2019), and the random forest to predict the number of citations that articles in the field of biomedicine will receive, where features such as the number of co-authors, journal quality and keywords were used to train the model (Medina and Ñique, 2017), and others such as neural network models (Neural Network, TabNet) and LightGBM, the latter identified by Hu *et al.* (2023) as having the highest performance.

(4) Co-authorship network analysis to detect groups or communities of collaborating authors. The main algorithms most commonly used in the literature are community detection and the Louvain method, both applied in bibliometric studies (Carratalá *et al.*, 2012; Knop *et al.*, 2019), which also use visualization techniques to map collaborative networks between researchers, revealing patterns of co-authorship and highlighting the most influential communities in the field.

(5) Automatic screening or sifting of the literature using deep or active learning algorithms (Forero and Bennasar, 2022), speeding up reviews to find the most relevant research quickly, without having to spend several days reading. Following this idea, one of the tools that has gained popularity is ASReview Lab, which uses state-of-the-art active learning techniques to solve one of the most interesting challenges in the systematic selection of large amounts of text (ASReview, 2024).

The use of machine learning in bibliometrics has advanced the analysis and organization of large volumes of scientific data and will continue to evolve in the coming years. The processing of structured data offers a significant advantage for the use of supervised techniques, while unstructured data requires more complex techniques, such as deep

Ciencia & Turismo
ISSN: 3028-8665
Vol. 3 No. 4. pp. 23 – 40, octubre – diciembre 2024
Revista multidisciplinar
www.cienciayturismo.org

learning, for efficient analysis. However, the effective use of these techniques requires well-organized datasets, appropriate pre-processing and an optimal choice of ML algorithms depending on the type of data and the research objectives.

2. Methodology

For this article, a literature review was conducted to qualitatively analyse the applications of machine learning in bibliometrics, and a descriptive bibliometric study was conducted. From the point of view of Cohen and Gómez (2022), it consists of identifying, summarizing, systematizing and analyzing all the information generated in relevant information sources to maintain a timely and effective relationship with the subject under analysis. Therefore, to develop the article, a review of the articles published in the Scopus databases was carried out. In addition to this, the study has a quantitative approach, through bibliometric numbers, to make a comparison of the evolution and contribution of the perspectives of different authors on the subject, and the research modality is bibliographic since it uses secondary data as a source of information (Torres, 2016). To proceed with the research, (1) the scope of the research was determined, which includes two aspects: a) as context, full-text publications and b) publications of articles related to studies with the keyword "consumer behaviour"; (2) the research questions were defined, which appear in Table 1.

Table 1

Research questions.

Research questions.
1. What is the trend in article publications and citations per year?
2. What are the topics of the articles?
3. Who are the most representative authors for each topic?
4. What are the main collaborative networks between authors and countries?
5. Which are the most cited articles?

Source: Own elaboration.

Subsequently, (3) inclusion and exclusion criteria were defined. The inclusion criteria determine which articles are selected for the research. The exclusion criteria limit the articles that will not be part of the analysis. The inclusion and exclusion criteria used are described in Table 2 below.

Ciencia & Turismo
ISSN: 3028-8665
Vol. 3 No. 4. pp. 23 – 40, octubre – diciembre 2024
Revista multidisciplinar
www.cienciayturismo.org

Table 2

Inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
1. Journal articles published between 2014 and 2024. 2. Articles containing the keyword: 'consumer behaviour'. 3. Open access articles.	1. Articles in languages other than English. 2. Articles from systematic reviews or bibliometric analysis. 3. Duplicate articles.

Source: Own elaboration.

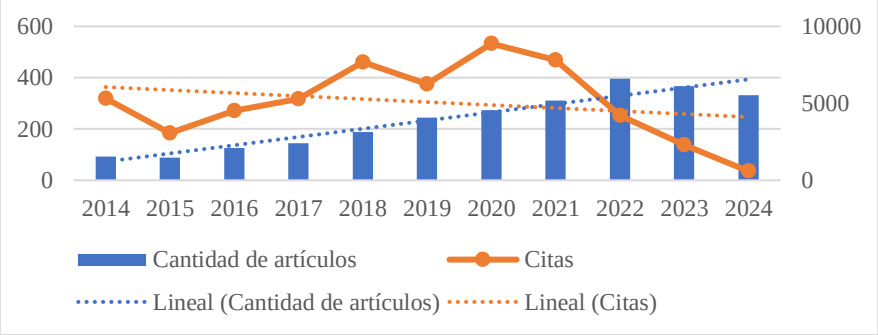
Next, 2563 articles that met the inclusion criteria were processed. The machine learning algorithms were modelled using R Studio software, in this case, Latent Dirichlet Allocation (LDA), and the bibliometrix package (Aria and Cuccurullo, 2017) and ggplot2 were used to visualize the graphs (Wickman, 2016).

3. Results

The analysis shows an increasing trend in scientific output from 2014 to 2024, with a steady increase in the number of published articles, peaking in 2022 with 396 publications. However, the number of citations does not follow a homogeneous pattern over time. Despite the increased production in recent years, citations show a significant decrease from 2020 onwards, which could indicate that more recent articles have not had enough time to be widely cited or that the visibility and impact of research published in recent years on this topic has decreased. This behaviour could also be related to increased competitiveness or the need to disseminate more recent work more widely. The significant drop in citations for 2023 and 2024 is particularly noteworthy, with 2292 and 593 citations respectively, highlighting the importance of continuously monitoring these indicators over time to assess their academic impact.

Figure 1

The Trend in article production and citations.



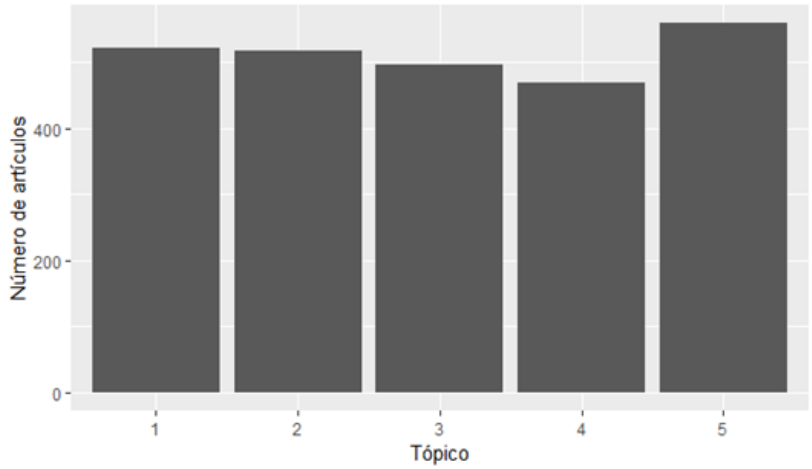
Source: Own elaboration.

Ciencia & Turismo
ISSN: 3028-8665
Vol. 3 No. 4. pp. 23 – 40, octubre – diciembre 2024
Revista multidisciplinar
www.cienciayturismo.org

Figure 2 shows the distribution of the number of articles for each of the five topics identified by the Latent Dirichlet Allocation (LDA) machine learning algorithm. There is considerable variability in the number of publications associated with each category. Topics 1 and 5 concentrate the largest number of articles, significantly outnumbering the other categories. In contrast, themes 3 and 4 have a lower frequency of publications. This uneven distribution suggests a concentration of scientific production in Themes 1 and 5, which could indicate a greater relevance or research interest in these specific areas within the field of study in question.

Figure 2

Distribution of articles by topic.



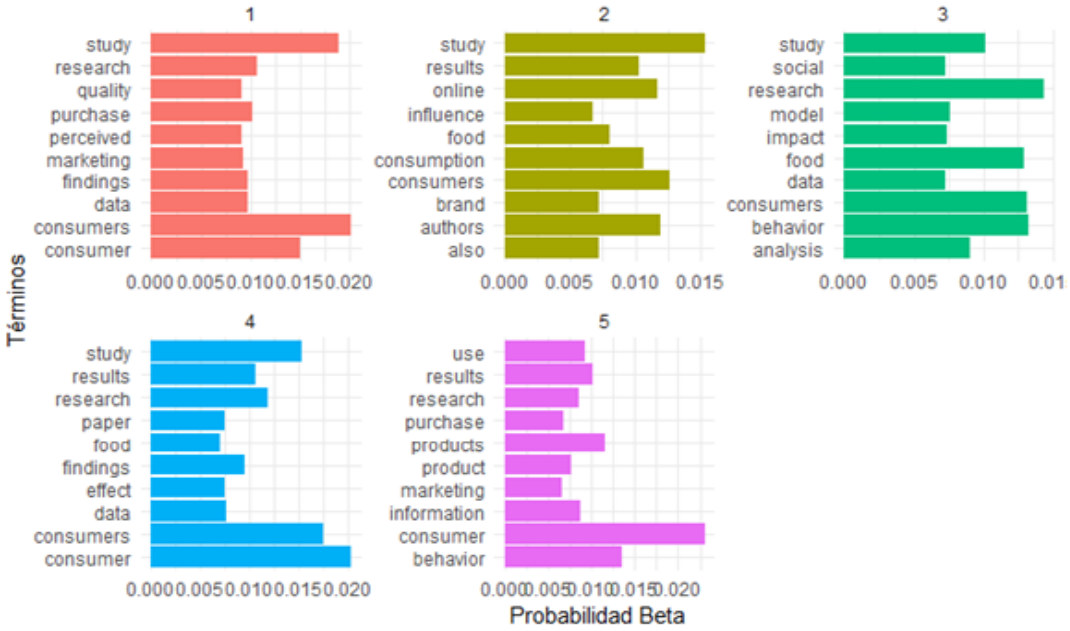
Source: Own elaboration.

Figure 3 shows the probability distribution of the most relevant terms within each of the five topics identified by the LDA model for consumer behaviour studies. Each topic represents an underlying theme in the analyzed documents, and the most likely terms within each topic provide a window into their semantic content. There is a clear thematic distinction between the themes. For example, Topic 1 focuses on market research, which uses data collection to explore variables of perceived product quality that influence consumer purchases. Topic 2 brings together research on the influence of the online environment on food consumption and brand perception. Theme 3 addresses issues related to the social impact of food consumption, as well as predictive modelling and analysis of food consumption data. Topic 4 focuses on the impact of specific foods on health and consumer behaviour. Theme 5 hosts research that examines consumer behaviour and its relationship to marketing and product purchase.

Ciencia & Turismo
ISSN: 3028-8665
Vol. 3 No. 4. pp. 23 – 40, octubre – diciembre 2024
Revista multidisciplinar
www.cienciayturismo.org

Figure 3

Top 10 words by topic.



Source: Own elaboration.

The bibliometric analysis of the relevant authors in the five themes shows a differentiated distribution of citations, reflecting the influence and recognition in their respective fields. In Theme 1, the most prominent authors are Rifkin and Almotairi, with 442 and 435 citations respectively, indicating their pre-eminence in this field. On the other hand, in Topic 2, Gruber leads with 1635 citations, followed by Kunz (1205), indicating a significant impact on the related literature. Topic 3 is dominated by Bonnevie (750 citations) and Buil (366), while in Topic 4 Hieke is the most cited author with 849 citations, followed by Lal (368). In Topic 5, Wang (967 citations) and Carlson (904) stand out as the most influential authors. These data show a strong concentration of citations on certain authors, suggesting academic leadership in their respective topics.

Table 3

Top authors in each field.

Authors	Topic	Cites	Authors	Topic	Cites	Authors	Topic	Cites
Rifkin	1	442	Gruber	2	1635	Bonnevie	3	750
Almotairi	1	435	Kunz	2	1205	Buil	3	366
Petit	1	290	Scarpi	2	580	Ben Ismail	3	344
Dwivedi	1	281	Pantano	2	555	Liao	3	298
Goh	1	279	Dennis	2	476	Ernest	3	190

Ciencia & Turismo
ISSN: 3028-8665
Vol. 3 No. 4. pp. 23 – 40, octubre – diciembre 2024
Revista multidisciplinar
www.cienciayturismo.org

Table 3

Top authors in each field (continuation)

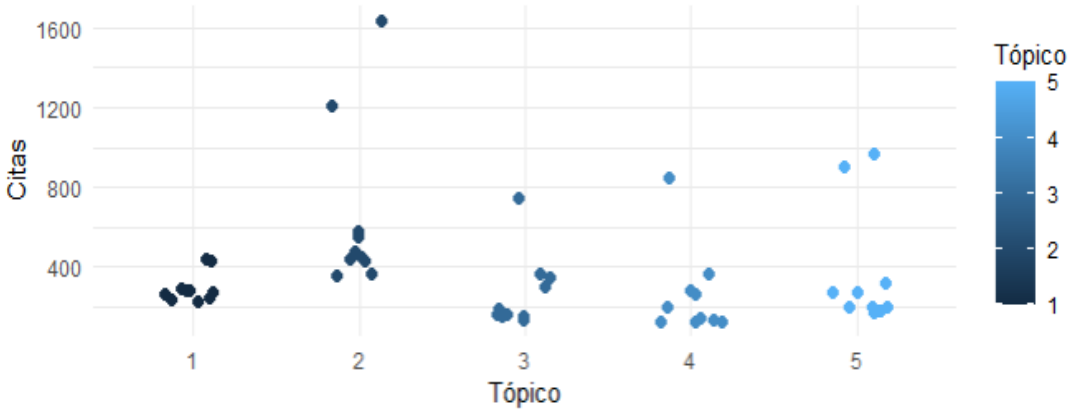
Authors	Topic	Cites	Authors	Topic	Cites	Authors	Topic	Cites
Hieke	4	849	Wang	5	967			
Lal	4	368	Carlson	5	904			
Moeller	4	283	Su	5	321			
Segovia	4	268	Brad	5	277			
Savva	4	200	Hingorani	5	273			

Source: Own elaboration.

Figure 4 shows the dispersion of the number of citations by each of the five themes identified by the LDA model, color-coded. Each dot represents an author and its position on the horizontal axis indicates the dominant topic, while the vertical axis reflects the total number of citations received. There is considerable variation in the number of citations between authors. This variation suggests that the authors assigned to Themes 2 and 5, such as Gruber, Kunz, Wang, Carlson, Hieke and Scarpi, are the most influential and have had the greatest impact on the scientific community in terms of citations received.

Figure 4

Distribution of citations of authors in each field.



Source: Own elaboration.

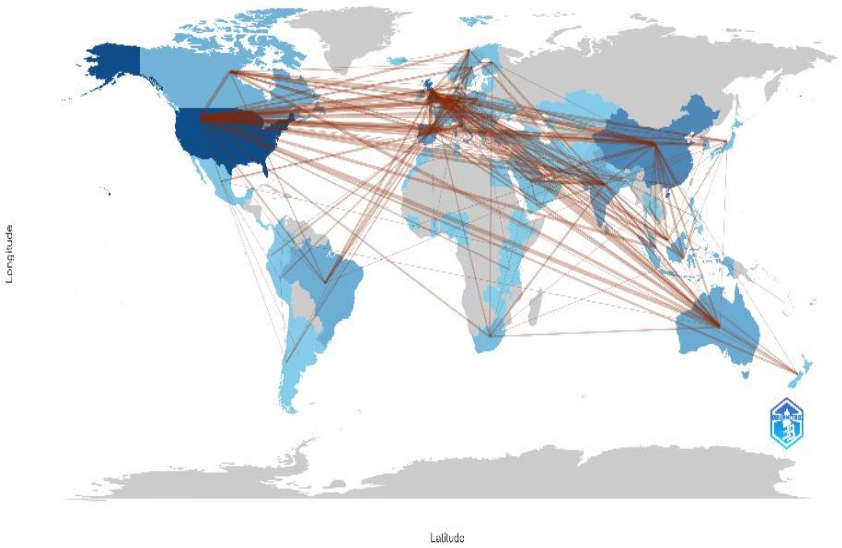
The analysis of scientific collaboration between countries reveals an interconnected network that highlights strong collaboration between some specific geographical regions. In particular, countries such as the United States and the United Kingdom show a high frequency of collaboration and a high scientific output (40 collaborations), suggesting a consolidated scientific relationship. Other countries, such as Australia and the United Kingdom (26 collaborations), the United Kingdom and Italy (24 collaborations), and the

Ciencia & Turismo
ISSN: 3028-8665
Vol. 3 No. 4. pp. 23 – 40, octubre – diciembre 2024
Revista multidisciplinar
www.cienciayturismo.org

United Kingdom and China (23 collaborations), also show significant collaborative links. These data reflect the geopolitical and academic influence in building scientific networks, as well as the importance of consumer behaviour at an international level, with countries with high levels of investment in R&D leading international collaboration. The analysis also suggests that there are opportunities to expand collaborations in regions where the frequency is low or emerging, as in the case of Australia and Brazil (2 collaborations). These results allow us not only to understand the distribution of international scientific output but also to identify potential gaps and strategic areas for strengthening global scientific cooperation.

Figure 5

Collaborations between countries



Países	Fr
USA	1283
UK	669
CHINA	611
ITALY	482
SPAIN	394
BRAZIL	316
AUSTRALIA	305
NETHERLANDS	252
GERMANY	247
INDIA	246

Source: Own elaboration

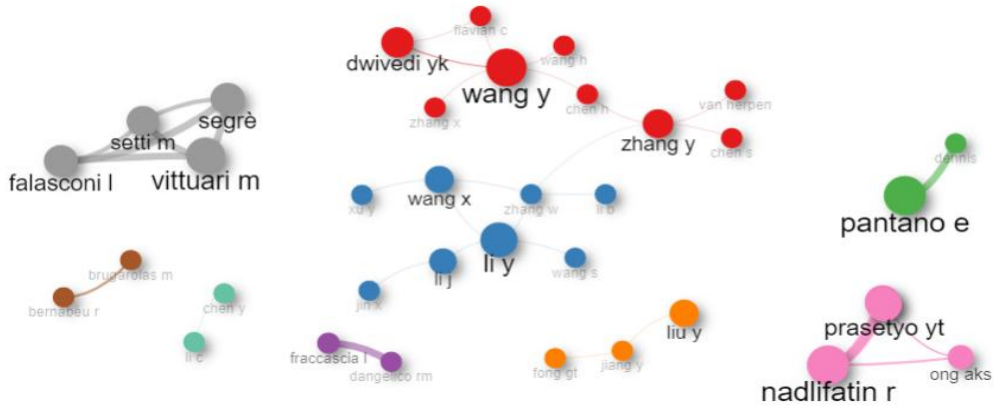
The analysis of the academic collaboration network reveals a clear distribution of authors in different clusters, with key metrics reflecting their importance within the network. The Betweenness metric highlights authors such as Zhang Y, with a value of 77, indicating his crucial role as a facilitator of collaboration between different groups of researchers. In contrast, authors such as Chen S and Dwivedi YK have Betweenness values of 0, suggesting less mediation in the network. The closeness metric, which reflects an author's proximity to the rest of the network, shows that Wang Y has the highest value (0.02), followed by Zhang Y (0.027), which places them in a position accessible to the rest of the researchers. In terms of 'PageRank', which measures overall influence, Wang Y leads with a value of 0.065, while Zhang Y has a value of 0.049. These results show the importance of certain authors, such as Wang Y and Zhang Y, both in terms of mediation and overall

Ciencia & Turismo
ISSN: 3028-8665
Vol. 3 No. 4. pp. 23 – 40, octubre – diciembre 2024
Revista multidisciplinar
www.cienciayturismo.org

influence in the network, positioning them as key players in the collaborative development of the field.

Figure 6

Network of collaboration between authors.



Source: Own elaboration.

The most cited articles are listed in Table 4. The data highlights the impact and relevance of each article in its field. The article with the highest number of citations is Wirtz (2018) with 1205 citations, although it is the second article in terms of citations per year (172.14), indicating a sustained impact, but not necessarily the most recent. The paper by Dwivedi (2021) stands out with the highest citation rate per year (226.00), suggesting rapid recognition in the scientific community. Other articles, such as those by Grunert (2014) and Zhu (2014), although older, maintain a considerable citation rate (77.18 and 68.18), reflecting their relevance in their fields over time.

Table 4

Most cited articles

Artículo	DOI	Cites	C per year
WIRTZ J, 2018, J SERV MANAGE	http://dx.doi.org/10.1007/978-3-031-36589-8_6	1205	172.14
DWIVEDI YK, 2021, INT J INF MANAGE	https://doi.org/10.1016/j.ijinfomgt.2020.102168	904	226.00
GRUNERT KG, 2014, FOOD POLICY	https://doi.org/10.1016/j.foodpol.2013.12.001	849	77.18
ZHU SH, 2014, TOB CONTROL	https://doi.org/10.1136/tobaccocontrol-2014-051670	750	68.18

Ciencia & Turismo
ISSN: 3028-8665
Vol. 3 No. 4. pp. 23 – 40, octubre – diciembre 2024
Revista multidisciplinar
www.cienciayturismo.org

Table 4

Most cited articles (continuation)

Artículo	DOI	Cites	C per year
PANTANO E, 2020, J BUS RES	https://doi.org/10.1016/j.jbusres.2020.05.036	476	95.20
NAPOLI J, 2014, J BUS RES	https://doi.org/10.1016/j.jbusres.2013.06.001	447	40.64
KAMBOJ S, 2018, INT J INF MANAGE	https://doi.org/10.1016/j.ijinfomgt.2017.12.001	444	63.43
KIRK CP, 2020, J BUS RES	https://doi.org/10.1016/j.jbusres.2020.05.028	442	88.40
BENOIT S, 2017, J BUS RES	https://doi.org/10.1016/j.jbusres.2017.05.004	430	53.75
DWIVEDI YK, 2016, GOV INF Q	https://doi.org/10.1016/j.giq.2015.06.003	368	40.89

Source: Own elaboration.

4. Conclusion

- Machine learning is presented as an essential tool for modern bibliometric analysis. The volume of scientific publications has grown exponentially, making it difficult to analyse with traditional tools, and allowing the automation of complex processes that previously required a great deal of effort from researchers.
- There are several machine learning techniques that have been successfully applied to bibliometrics. Supervised, unsupervised and reinforcement machine learning, as well as deep learning, have proven their usefulness. Applications such as document classification, clustering, citation prediction, collaborative networks and automatic literature screening stand out.
- Bibliometric analysis of consumer behaviour benefits from machine learning. The study shows an increasing trend in scientific production on this topic but with a decrease in citations from 2020 onwards. The use of the LDA algorithm allowed the identification of 5 main themes, with key authors such as Rifkin, Almotairi, Gruber, Kunz, Bonnevie, Buil, Hieke, Lal, Wang and Carlson leading research in their respective areas.
- International scientific collaboration in the field of consumer behaviour is robust but with potential for expansion. Analysis of cross-country collaborative networks shows strong links between the US, UK, Australia, Italy and China, reflecting their influence on research. Opportunities to strengthen collaboration with other regions are identified. The network of collaborations between authors shows the importance of figures such as Wang and Zhang as facilitators and influencers. Analysis of the most cited articles highlights the impact of work such as that of Wirtz and Dwivedi.

5. Conflict of interest

The authors declare that there is no conflict of interest.

6. Author contribution statement

All authors contributed significantly to the preparation of the article.

7. Financing costs

This research was funded entirely with the authors' own funds

8. References

Aria, M. y Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

ASReview. (2024, 11 de Octubre). Aprendizaje activo para revisiones sistemáticas. https://asreview.nl/

Barrera Rodríguez, A. M., Duque Hurtado, P. L., y Merchán Villegas, V. L. (2022). Neurociencia y comportamiento del consumidor: análisis estadístico de su evolución y tendencias en su investigación. *Cuadernos Latinoamericanos De Administración*, *18*(35). https://doi.org/10.18270/cuaderlam.v18i35.3855

Cabeza Ramírez, L. J. (2019). *Aproximación bibliométrica a la investigación en emprendimiento* [Tesis Doctoral]. Universidad de Córdoba. https://helvia.uco.es/bitstream/handle/10396/18690/2019000001933.pdf?sequence=1&isAllowed=y

Camarena Gómez, B. O., y Romero Valenzuela, D. M. (2023). El consumo sustentable y responsable: conceptos y análisis desde el comportamiento del consumidor. *Revista Vértice Universitario*, 25(94). https://doi.org/10.36792/rvu.v25i94.75

Cao, L. (2020). Data Science: A Comprehensive Overview. ACM Computing Surveys, *50*(3), 43:1-42. http://dx.doi.org/10.1145/0000000.0000000

Carratalá Munuera, M. C., Orozco Beltrán, D., Gil Guillen, V. F., Navarro Pérez, J., Quirce, F., Merino, J., y Basora, J. (2012). Análisis Bibliométrico de la producción científica internacional sobre atención primaria. *Atención Primaria*, *44*(11), 651-658. https://doi.org/10.1016/j.aprim.2011.12.002

Cohen, N; y Gómez, R. (2022). Pensar el para qué de la Metodología de la investigación. Un recorrido por la producción de los datos y los diseños. *Revista*

Latinoamericana de Metodología de las Ciencias Sociales, 12(2), e122. https://doi.org/10.24215/18537863e122

Forero Corba, W., y Bennasar, F. N. (2024). Técnicas y aplicaciones del Machine Learning e Inteligencia Artificial en educación: Una revisión sistemática. RIED-*Revista Iberoamericana de Educación a Distancia*, 27(1). https://www.redalyc.org/journal/3314/331475280017/html/

Fuentealba, D., Flores-Fernández, C., y Carrasco, R. (2023). Análisis bibliométrico y de contenido sobre VUCA. *Revista Española de Documentación Científica*, *46* (2), e354. https://doi.org/10.3989/redc.2023.2.1968

García, M. G. (2018). Análisis de sensibilidad mediante Ramdom Forest [Tesis de Ingeniería]. Universidad Politécnica de Madrid. https://oa.upm.es/53368/1/TFG_MARTA_GARCIA_RUIZ_DE_LEON.pdf

Guerra Castellón, E. E., Sotero Arteaga, L. K., Vázquez Alfonso, Y., y Colcha Ortiz, A. R. (2024). Diagnóstico de la producción científica registrada sobre la transformación digital en el turismo a partir de un análisis bibliométrico. *AlfaPublicaciones*, *6*(2.1), 46–63. https://doi.org/10.33262/ap.v6i2.1.476

Hu, Z., Cui, J., y Lin, A. (2023). Identifying potentially excellent publications using a citation-based Machine Learning approach. *Information Processing & Management*, *60*(3), 103323. https://doi.org/10.1016/j.ipm.2023.103323

Janiesch, C., Zschech, P., y Heinrich, K. (2021). Machine Learning and deep learning. *Electronic Markets*, *31*(3), 685-695. https://doi.org/10.1007/s12525-021-00475-2

Junior, A., Gomes, F., y Rodrigues, G. (2012). Categorização automática de artigos científicos da engenharia de produção utilizando métodos de aprendizagem de máquina. *Encontro Nacional De Engenharia De Produção: Desenvolvimento Sustentável E Responsabilidade Social: As Contribuições da Engenharia de Produção*, 32. https://abepro.org.br/biblioteca/enegep2012_tn_sto_164_956_20299.pdf

Knop, S., Merchel, R., y Poeppelbuss, J. (2019). Author Collaboration in Ten Years of IPS2: A Bibliometric Analysis. *Procedia CIRP*, *83*, 22-27. https://doi.org/10.1016/j.procir.2019.03.092

Medina Merino, R. F., y Ñique Chacón, C. I. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y

Python. *Interfases*, *10*(010), 165-189. https://doi.org/10.26439/interfases2017.n10.1775

Mohammed, M., Khan, M. B., y Bashier, E. B. M. (2016). *Machine Learning: Algorithms and Applications*. CRC Press. https://doi.org/10.1201/9781315371658

Pérez Ricardo, E. C., Batista Sánchez, E., Medina Labrada, J. R., y Fernández Cueria, N. (2021). Comportamiento del consumidor turístico. Análisis bibliométrico en Scopus. Pensar la Publicidad. *Pensar la publicidad*, *15*(2), 289-298. https://doi.org/10.5209/pepu.77526

Sailema Altamirano, J. C., Vargas Lizano, J. C., y Medina Chicaiza, R. P. (2022). Análisis bibliométrico y contextual de la producción científica del marketing sensorial. *Religación*, 7(31), e210883. https://doi.org/10.46652/rgn.v7i31.883

Saravia Ramos, G. del P., Palomino Flores, J. V., Mercado Hermenegildo, A., y Pichilingue Romero, J. V. U. (2022). Comportamiento del consumidor: Estudio bibliométrico de la base de datos Scopus 2011-2021. *INNOVA Research Journal*, *7*(3), 50-74. https://doi.org/10.33890/innova.v7.n3.2022.2112

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Sarker, I. H., Furhad, M. H., y Nowrozy, R. (2021). AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions. *SN Computer Science*, *2*(3), 173. https://doi.org/10.1007/s42979-021-00557-0

Torres, P. (2016). Acerca de los enfoques cuantitativo y cualitativo en la investigación educativa cubana actual. *Atenas*, 2(34), 1-15. http://atenas.umcc.cu/index.php/atenas/article/view/194

Urbizagastegui, R. (2021). La bibliometría brasileña: minería de textos. *Revista ACB*, *26*(1), 1–18. https://revista.acbsc.org.br/racb/article/view/1768

Vasquez, C. E., y Bolaños, J. M. (2019). *Implementación y evaluación de técnicas de clustering para un sistema de recuperación de documentos jurisprudenciales* [Tesis de Grado]. Fundación Universitaria De Popayán. https://fupvirtual.edu.co/repositorio/files/original/4fa0079ccb70d407eeac136eec3a5c7ec59bae74.pdf

Wickman, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer Verlag. https://ggplot2.tidyverse.org

ISSN: 3028-8665
Vol. 3 No. 4. pp. 23 – 40, octubre – diciembre 2024
Revista multidisciplinar

www.cienciayturismo.org

El artículo que se publica es de exclusiva responsabilidad de los autores y no necesariamente reflejan el pensamiento de la **Revista Ciencia & Turismo.**

El artículo queda en propiedad de la revista y, por tanto, su publicación parcial y/o total en otro medio tiene que ser autorizado por el director de la **Revista Ciencia & Turismo.**