An Acad Bras Cienc (2024) 96(3): e20230789 DOI 10.1590/0001-3765202420230789
Anais da Academia Brasileira de Ciências | *Annals of the Brazilian Academy of Sciences*
Printed ISSN 0001-3765 I Online ISSN 1678-2690
www.scielo.br/aabc | www.fb.com/aabcjournal

PALEONTOLOGY

Overview of computational methods in taphonomy based on the combination of bibliometric analysis and natural language

RONALDO A. LEONI, LAÍS ALVES-SILVA & HERMÍNIO ISMAEL DE ARAÚJO-JÚNIOR

Abstract: Artificial intelligence tools are new in taphonomy and are growing fast. They are being used mainly to investigate bone surface marks. In order to investigate this subject, a bibliometric study was made to understand the growing rate of this intersectional field, the future, and gaps in the field until now. From Scopus and Google Scholar metadata, graphs were made to describe the data, and inferential statistics were made by regression with the Ordinary Least Squares method. Exploratory analysis with word clouds, topic modeling, and natural language processing with Latent Dirichlet Allocation as a method were also made using the entire corpus from the papers. From the first register until 2023, we found eight articles in Scopus and 32 in Google Scholar; the majority of the studies and the most cited were from Spain. The studies are growing fast from 2016 to 2018, and the regression shows that growth can be maintained in the coming years. Exploratory analysis shows the most frequent words are marks, models, data, and bone. Topic modeling shows that the studies are highly concentrated on similar problems and the tools to solve them, revealing that there is much more to explore with computational tools in taphonomy and paleontology as well.

Key words: artificial intelligence, google scholar, machine learning, paleontology, web scraping, scopus.

INTRODUCTION

Artificial intelligence (AI) methods were anticipated to be employed as computational tools in paleontology in the 1990s (Kaesler 1993). Nowadays, they are utilized to study a variety of problems, such as the identification of discrete fossiliferous levels (Martin-Perea et al. 2020), the automatic recognition of palaeobios images (Xu et al. 2020), and the recognition of rare microfossils (Wang et al. 2022). To date, numerous AI computational tools have been employed to investigate taphonomy-related problems as well, such as bone surface marks (BSM). These include cut marks (Byeon et al. 2019, Cifuentes-Alcobendas & Domínguez-Rodrigo 2019, Domínguez-Rodrigo 2018), predation marks (Jiménez-García et al. 2020), and general BSM (Domínguez-Rodrigo 2019). Other issues, such as the diagenetic pathways of assemblages (Pizarro-Monzo et al. 2021), are also studied. All of these works use AI to perform time-consuming tasks, make difficult identifications, and reduce subjectivity in the analysis.

The simplest, rigorous, and most popular approach for exploring scientific data and identifying research gaps is to utilize traditional bibliometric analysis (Donthu et al. 2021). It analyzes the literature on a specific subject by combining statistical methods and data visualization. This approach helps identify the knowledge structure, interactions between publications, current development, and research frontiers. Usually, metadata such as publication year, author(s), language, and keywords are used to conduct statistical studies (Jiang et al. 2016). This approach is valuable not only for understanding the growth of knowledge but also for showcasing how institutions and authors contribute to the advancement of knowledge in a specific field (Wang & Maniruzzaman 2022).

However, analysis based on metadata in a research field with small sample sizes requires extra caution as it can lead to a biased conclusion (Lin 2018). In such a situation, an alternative method that does not rely on a large number of datasets can be useful. Therefore, in this paper, a complementary approach utilizing *topic modeling* for reviewing is suitable. This method leverages the entire corpus to describe the structure of the papers, providing more comprehensive and in-depth content analysis results compared to bibliographic analysis alone (Jiang et al. 2016).

Since the field of computational methods for taphonomy is relatively new and has few papers, the topic modeling approach is thought to be very useful in literature reviews. A broader review of this field can be made by applying topic modeling in conjunction with bibliographic analysis. The bibliometric statistical analysis of the metadata provides a general perspective on the publications and recognizes the research direction. On the other hand, since topic modeling utilizes statistical algorithms, it can extract semantic information and ultimately identify the research gap by assessing substantial textual data, summarizing, and understanding an unstructured collection of texts. And giving a more accurate result of what is discussed inside the field (Jiang 2016, Li & Lei 2019).

Furthermore, despite Scopus having better-advanced search and filtering options that are more suitable for systematic reviews, this database suffers from low coverage, slow indexing, and is not accessible. These characteristics are essential when analyzing a relatively new field. Therefore, despite the challenges of extracting metadata from Google Scholar, this database emerges as a complementary option with more comprehensive coverage than Scopus (Martin-Martin et al. 2021).

No bibliometric study has reviewed how computational methods have been employed until now to address taphonomy problems. Therefore, this paper represents the first comprehensive exploration of this topic. The combination of classical metadata, natural language approaches, and the utilization of both Scopus and Google Scholar is a powerful approach that can potentially reveal the field's overall landscape.

MATERIALS AND METHODS

The collection of articles for bibliometric analysis was conducted on two databases: Scopus and Google Scholar. Data from Scopus were collected using the tools available on the website, accessed through the CAPES (Coordination of Superior Level Staff Improvement) Periódicos portal, which provides access to a wide range of peer-reviewed journals. The data from Google Scholar was collected using automated web scraping techniques.

To identify relevant articles, we utilized the following keywords: "artificial intelligence", "deep learning", "machine learning", and "deep learning, machine learning, and taphonomy", or "fossil diagenesis or biostratinomy". Web scraping, also known as web extraction or web harvesting, is a method used to extract structured data from unstructured data found on the World Wide Web (WWW) and store it (Kumaresan & Ramanujam 2016). This technique is widely recognized as a powerful tool for collecting extensive data (Zhao 2017) and is also employed for gathering data in bibliometric analysis (Santos et al. 2020, Santana & Braga 2020). Web scraping is highly valuable, due to its ability to automatically collect a massive amount of data using a robot or algorithm. As a result, it plays a crucial role in contemporary research focusing on digital phenomena (Farias et al. 2021).

We initially utilized the Beautiful Soup and Selenium libraries to do automated web scraping. These libraries allowed us to interact with the data source and save the information into a file, specifically a comma-separated values (CSV) data frame, for subsequent analysis (Zhao 2017). Subsequently, we clean the data using Pandas, and perform descriptive statistics with Matplotlib, Numpy, Statsmodels and Plotly to visually describe the dataset's characteristics. We used Sklearn to perform regression analysis using the ordinary least squares (OLS) method for inferential statistics. These processes were implemented using the Python programming language, version 3.8.16.

Automated web scraping techniques

We used Selenium to create a Google Chrome browser instance to automate the web browsing process. This library allowed us to interact with the Google Scholar interface effectively. Additionally, we employed Beautiful Soup library to extract data from HTML files, facilitating the reorganization of the data structure for a better comprehension (Richardson 2007).

We configured the Selenium library to conduct article searches across all periods, ordered by date, and in all languages and formats, including citations. Once this configuration was set, we utilized the Beautiful Soup library to systematically parse and structure the unorganized information. To ensure a comprehensive search, we used the same keywords as those in the Scopus database and performed a search in Google Scholar. We captured only articles that contained these predetermined terms, ensuring the completeness of our search.

We utilized Pandas to export the collected data as a CSV file. We performed cleaning procedures during this process for Google Scholar, including removing duplicates, books, theses, and dissertations, as well as undefined numeric values. For Scopus, we focused on removing duplicates and erroneous entries. After the cleaning stage, we manually added certain missing information that was not collected automatically, such as citations, affiliations, correspondence addresses, and databases that were not directly available on the Google Scholar page or were not exported by Scopus.

Natural Language Processing

We employed a specific cleaning rocess for Natural Language Processing (NLP). Initially, we downloaded full-text articles from both the Scopus and Google Scholar databases. We obtained eight texts from Scopus and 32 texts from the Google Scholar database. These texts were transformed into plain text and then merged. Subsequently, we implemented a comprehensive cleaning process, which will be described in detail below.

The cleaning process consisted of two stages: the utilization of Regular Expressions (RegEx) and the Natural Language Toolkit (NLTK). RegEx is a library commonly used for searching, data validation, parsing, finding and replacing, data scraping, and syntax highlighting (Spichak et al. 2012). In this study, we employed RegEx to define a set of rules for eliminating noisy information from the data files. This including eliminating the first page of the articles, which contained information such as authors, addresses, emails, citations and Uniform Resource Locators (URL) embedded throughout the text, and references.

The NLTK Python library, designed for working with human language (Bird et al. 2001), was utilized to perform various text-processing tasks. Firstly, extraneous spaces, punctuation, articles, pronouns, and prepositions were excluded as they provide little meaningful information to the text. Following this process, we obtained the filtered data, consisting of only words. These words were then transformed into tokens, considered the fundamental units of meaningful information in NLP. The tokenization stage is crucial for further analysis and processing of the text data.

The subsequent phase in the NLP workflow typically includes stemming or lemmatization. Stemming involves reducing words to their root form by removing suffixes, to derive the general meaning. On the other hand, lemmatization also reduces words to their root form but considers the context to transform them into their dictionary form without losing meaning. Compared to stemming, lemmatization is generally a more intensive method (Scaccia & Scott 2021). For our analysis, lemmatization was preferred over stemming because it better preserves the meaning of the words, which is a crucial aspect of our study.

Data analysis

An exploratory analysis was conducted using a word cloud to assess the frequency of words, summarize the most relevant terms in the field, and remove words with less meaning. The word cloud figure was generated using a wordcloud library, where the size of each word corresponds to its frequency in the text. This visualization provides insights into the main focus of published articles (Atenstaedt 2012).

We utilized a topic modeling methodology to understand the text structures better, employing the Gensim and pyLDAvis libraries. Gensim is an open-source library designed explicitly for topic modeling and perform tasks such as document indexing and similarity search using unsupervised semantic analysis of plain texts (Řehůřek & Sojka 2010). Additionally, we employed the pyLDAvis library to visualize models generated by Gensim.

A topic model is a powerful machine learning tool that performs unsupervised clustering, categorization, and modeling of objects into latent topics, capturing the underlying significance of a group of documents. It is beneficial because it reveals the semantic structure present in extensive collections of documents (Kherwa & Bansal 2019). A topic is a collection of words frequently appearing together, representing a recurring pattern of co-occurring words. This analysis uncovers hidden structural patterns by linking words that share the same context and differentiating the usage of words in various meanings, thereby connecting documents that exhibit similar patterns (Barde & Bainwad 2017). The Latent Dirichlet (LDA) model was employed for the topic modeling analysis.

LDA is a topic modeling method that treats documents as a combination of different topics. Each document is assigned a distribution of topics, consisting of a small set of frequently used words. This approach enables a more precise assignment of documents to topics, as each document covers only a small set of topics (Barde & Bainwad 2017). To measure topic coherence, we utilized the Umass coherence metric, which compares a word with its preceding and succeeding words (Mohammed & Al-augby 2020).

RESULTS AND DISCUSSION

After completing the web scraping and cleaning procedures, we obtained a CSV file containing the following variables: title, authors, abstract, link, year, citations, affiliations, correspondence address, and database. The final dataset consists of eight rows from Scopus and 32 from Google Scholar, encompassing articles published between 2016 and January 2023.

The oldest publication titled "When felids and hominins ruled at Olduvai Gorge: A machine learning analysis of the skeletal profiles of the non-anthropogenic Bed I sites" was published by Arriaza & Dominguez-Rodrigo (2016) in the Quaternary Science Reviews journal and was identified in Scopus. This article employed classical artificial intelligence tools such as decision trees, random forest, neural networks, and support vector machines. On the other hand, one of the newest articles was conducted at the same palaeontological site, Olduvai Gorge, but utilized deep learning, specifically convolutional neural networks.

Table I presents the top five articles that received the highest citations among the 40 articles identified in the databases. The table includes the article's titles, author names, publication year, the number of citations received, and the journal where they were published.

The most cited articles from the Scopus database are from Palaeogeography, Palaeoclimatology, Palaeoecology; Quaternary Science Reviews; and Scientific Reports. The majority of the retrieved articles were published by Elsevier (38%), followed by Springer Nature Group and Springer (19%), John Wiley & Sons, Inc (12%), and publishers such as the Royal Society and Public Library of Science (6%). In Google Scholar, the order of journals followed a similar pattern, with the majority of the articles belonging to journals of Elsevier (42%), followed by Springer Nature Group (21%), and Springer (17%). Other publishers, such as PeerJ and MDPI, accounted for (8%) of the articles, while the

Table I. Five of the most cited articles.

Title	Authors	Citations	Year	Journal	Database
Use and abuse of cut mark analyses: The Rorschach effect	Domínguez-Rodrigo et al.	65	2017	Journal of Archaeological Science: Reports	Google
The hunted or the scavenged? Australopith accumulation by brown hyenas at Sterkfontein (South Africa)	Arriaza et al.	53	2021	Quaternary Science Reviews	Google
New taphonomic advances in 3D digital microscopy: A morphological characterisation of trampling marks	Courtenay et al.	38	2019a	Quaternary International	Google
Combining machine learning algorithms and geometric morphometrics: A study of carnivore tooth marks	Courtenay et al.	35	2019b	Palaeogeography, Palaeoclimatology, Palaeoecology	Scopus
Successful classification of experimental bone surface modifications (BSM) through machine learning algorithms: a solution to the controversial use of BSM in paleoanthropology?	Domínguez-Rodrigo	33	2019	Archaeological and Anthropological Sciences	Google

Royal Society represented (4%). Notably, the most cited articles were published in journals such as the Journal of Archaeological Science: Reports, Quaternary Science Reviews, and Quaternary International (Table II).

Compared to the Scopus database, Google Scholar has been able to retrieve articles from a more diverse range of regions. While many of the collected papers are from Spain, Google Scholar has also retrieved articles from countries such as India, Iran, the United States, the United Kingdom, Portugal, Indonesia, and Estonia. In this situation, Google Scholar is a suitable supplementary database and can even be considered the primary source. it provides a more comprehensive geographical coverage and includes the most cited articles in the field. This can be attributed to how the database operates, which contributes to its ability to retrieve diverse and highly influential articles.

Google Scholar, by default, collects a broader range of data sources, including scholarly journals, peer-reviewed articles, theses, dissertations, books, and other scientific-related materials, compared to Scopus (Nozuri 2005). Its ranking algorithm is based on more than 200 factors, although the specific weighting of these factors is not fully disclosed. However, citation counts play a significant role in the heuristic algorithm (Beel & Gipp 2009, Rovira et al. 2021). This likely explains why the most cited articles were found in the Google database (Nozuri 2005) and also accounts for its broader coverage (Moed et al. 2016).

In terms of their significance in the field, several journals stand out, including Journal of Archeology Science: Reports, Scientific Reports, Quaternary Science Reviews, Quaternary International, Archeological and Anthropological Sciences (in the case of Google), and Palaegeography, Palaeoclimatology, Palaeoecology, Scientific Reports and Quaternary Science Reviews (in the case of Scopus) (Figure 1). Between the journals, the University of Alcalá has the most scientific works spread out among them, followed by the University of Rovira i Virgili and the University of Bordeaux (Figure 2).

Before 2019, only a few articles were published. However, after this year, the number of publications has significantly increased annually (Figure 3).

The most cited institutions are located in Spain, with the University of Alcalá, Universitat Rovira i Virgili, and University of Burgos each receiving over 20 citations (Figure 4). As a result, Spain is also the most cited country with 448 citations (Figure 5). Additionally, Spain dominates the number of publications in most

Table II. **Three of the most cited journals of each database.**

Journals	Year	Citations	Affiliations	Country	Database
Quaternary Science Reviews	2016	30	University of Alcalá	Spain	Scopus
Palaeogeography, Palaeoclimatology, Palaeoecology	2019	35	Universitat Rovira i Virgili	Spain	Scopus
Scientific Reports	2019	28	University of Alcalá	Spain	Scopus
Journal of Archaeological Science: Reports	2017	65	University of Alcalá	Spain	Google
Quaternary International	2019	38	Universitat Rovira i Virgili	Spain	Google
Quaternary Science Reviews	2021	53	University of Alcalá	Spain	Google

journals (Figure 6). These institutions exhibit strong collaborations among their authors and departments, which is evident in their co-authorship patterns. Despite Spain's prominence in this field, there has been relatively little research on the history of paleontology in the country (Sequeiros et al. 1998). Additionally, although the history of taphonomy in Spain lacks comprehensive documentation, the country's taphonomy school is well known and has significantly contributed over the years to both theoretical (Fernández-López 2000, 2005, Fernández-López & Fernández-Jalvo 2002, Domínguez-Rodrigo et al. 2011) and practical (Domínguez-Rodrigo et al. 2009, Fernández-Jalvo & Andrews 2016) advances in the field. Therefore, this recent surge in remarkable works with innovative techniques in the field from Spain may represent another milestone in the development of the Spanish taphonomy school.

Ordinary least squares (OLS) regression analysis revealed a significant relationship between the increase in the number of published texts on taphonomy, biostratinomy, diagenesis, and artificial intelligence techniques over time ($p < 0.01$) (r-squared = 0.814). The number of texts increased from 1 in 2018 to 8 in 2019 and remained consistently above seven texts per year. (Figure 7). Based on the calculated trend, it is suggested that the number of publications in this field will continue to grow significantly in the coming years.

Natural Language Processing

The most meaningful words (Figure 8) that appeared in these texts are: model, mark, data, bone, sample, method, and algorithm. These words indicate that the texts primarily focused on the problem of identifying BSM using computational approaches. Some frequently appearing words in the word cloud, such as "using", "studying", and "from", were removed as they were less relevant to the analysis. Additionally, the role of samples and data

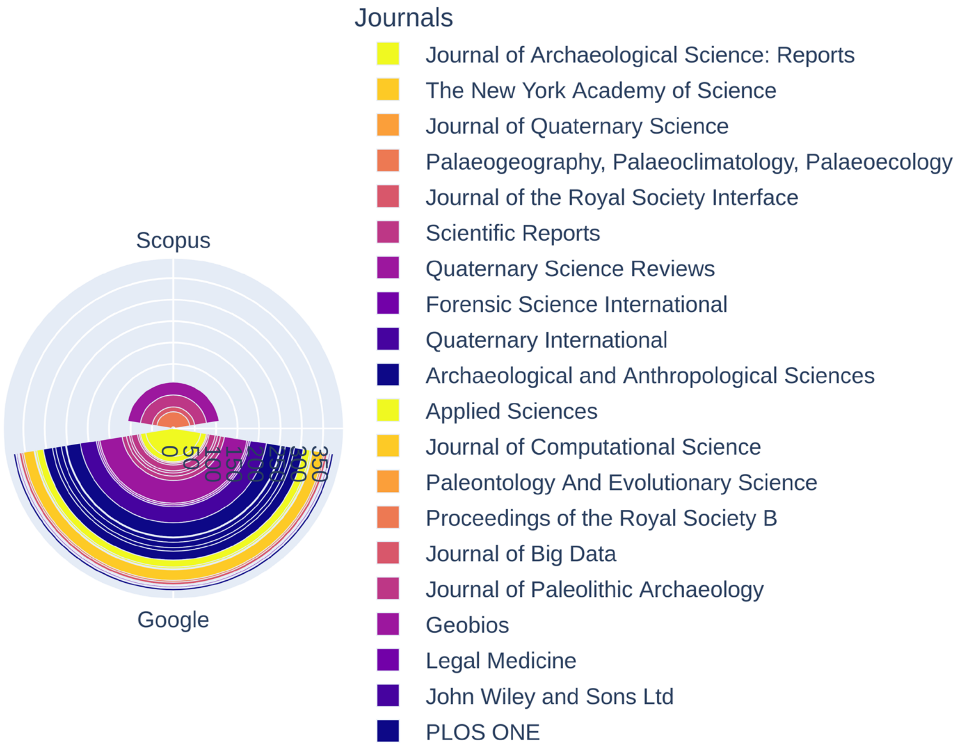

Figure 1. Journals found in the two databases.

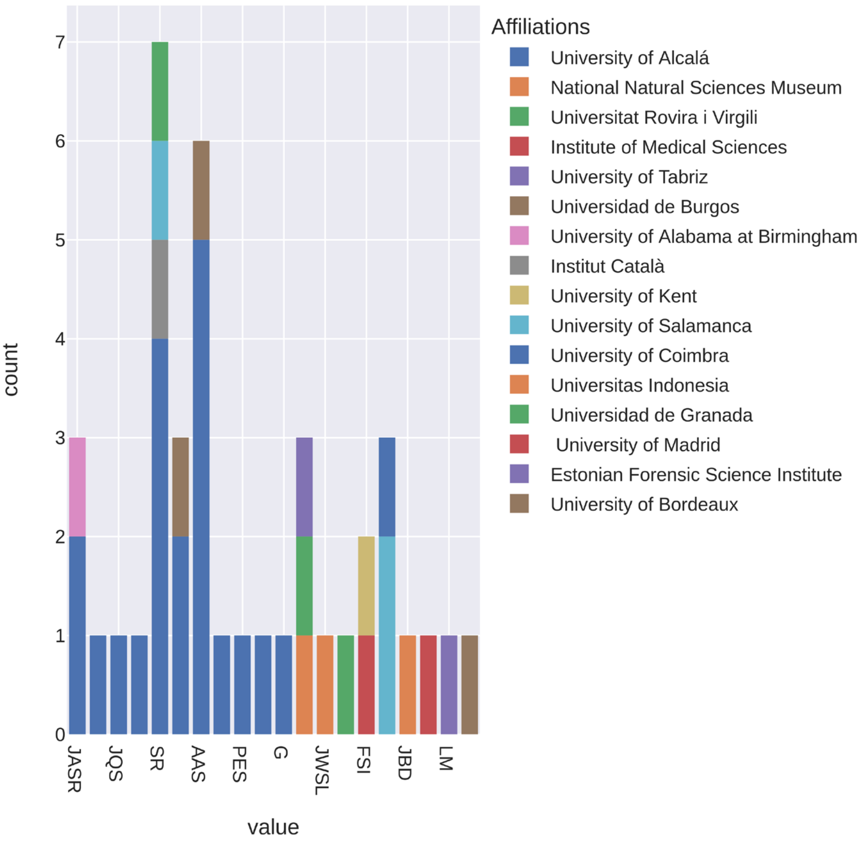

Figure 2. Number of manuscripts per journal by institutions.

is central in these texts, along with models and algorithms. Smaller words like dataset and accuracy further reinforce the central importance of data in these processes and the accuracy achieved through these models.

The LDA with 30 topics revealed that "marks" appear in the majority of the topics at λ = 1 (Figure 9), resulting in relevance given by the probability of the word to these topics, followed by "tooth", "bone", "data", and "mark", "accuracy", "models". However, these other words varied for each topic, leading to a different set of words for each topic when λ was lowered, giving more weight to the relevance of the word to the topic. When we use λ = 0.5 (Figure 10), we observe a more distinct separation of words for each topic. Topics 1 and 2 still have the word "marks" at the top, but in topic 1, it is followed by "bone" and "tooth", while in topic 2, which is the most distant from the others among the first topics, the word "marks" is followed by "marks" and "batallones". Topic 3 shows a word set of "data", "marks", and "sample" in the first three positions.

This significant similarity observed among the words across all topics is remarkable, even where the relevance assignment of a word to a topic shifts from a probabilistic approach to a more specific linkage by giving more weight to word specificity, as shown in the balanced λ (0.5) condition (Figure 10). This finding indicates that the field is highly focused on addressing a similar problem and has utilized the same set of tools to tackle these issues. This outcome is to be expected given the time elapsed since the publication of the first article in this field, signifying the early stages of investigation in the taphonomy field with these tools.

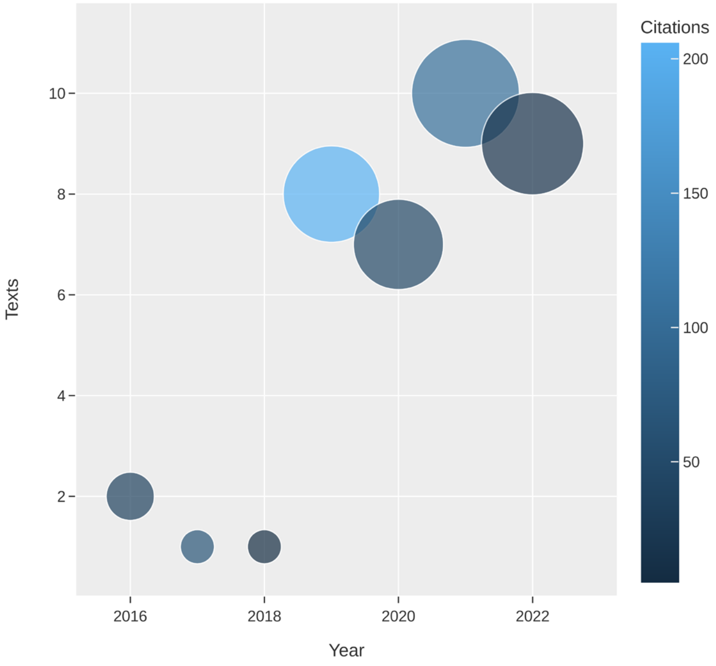

Figure 3. Number of texts per year, bubble size representing the number of texts, bubble color gradient representing number of citations.

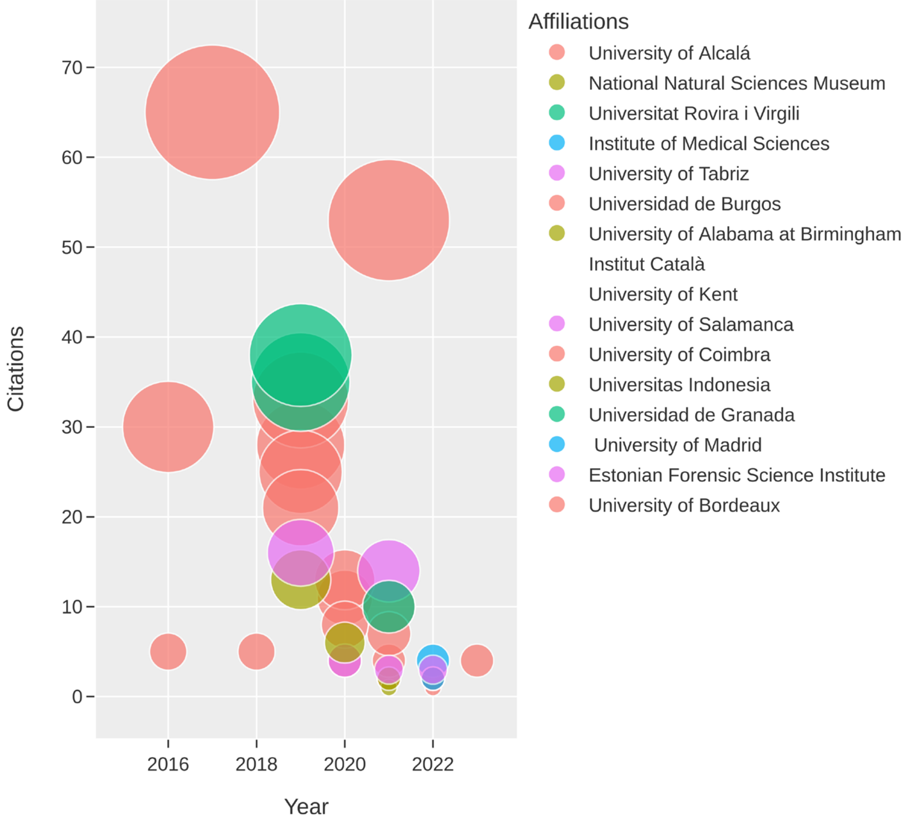

Figure 4. Number of citations by institutions per year.

The optimal number of topics is determined based on their coherence scores. We found values ranging from -18 and -0.28 for 10 to 30 topics, respectively (Table III). As the number of topics increased, the Umass score decreased. This metric assesses the semantic similarity

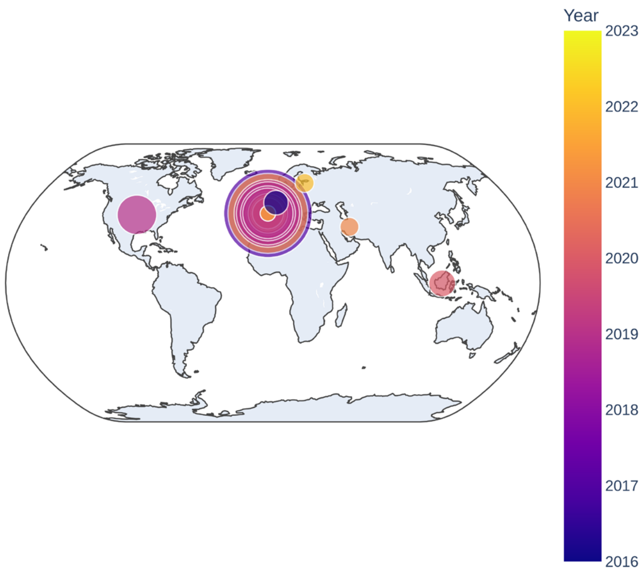

Figure 5. Number of manuscripts per year by country.

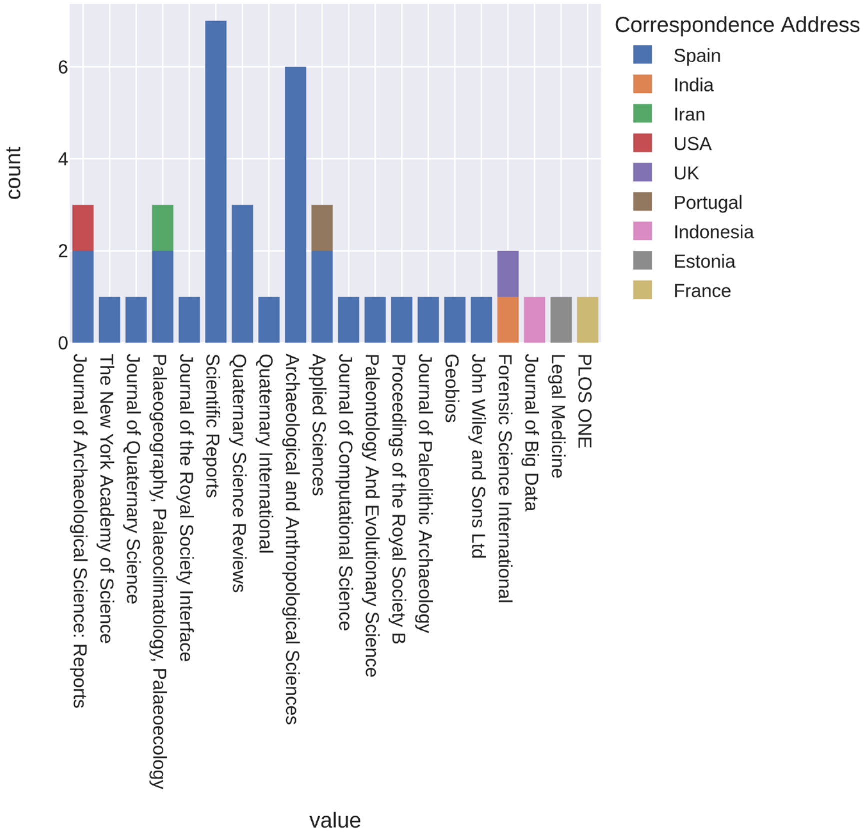

Figure 6. Number of manuscripts per journal by country

among words in a topic (Mohammed & Al-augby 2020). In our data, the high degree of similarity among words resulted in lower coherence scores for the model. This is expected in a novel field like this, where the focus of study is concentrated on a limited number of topics.

The model reveals that in the early stages of the field, there was a focus on addressing urgent and long-standing issues in taphonomy, such as

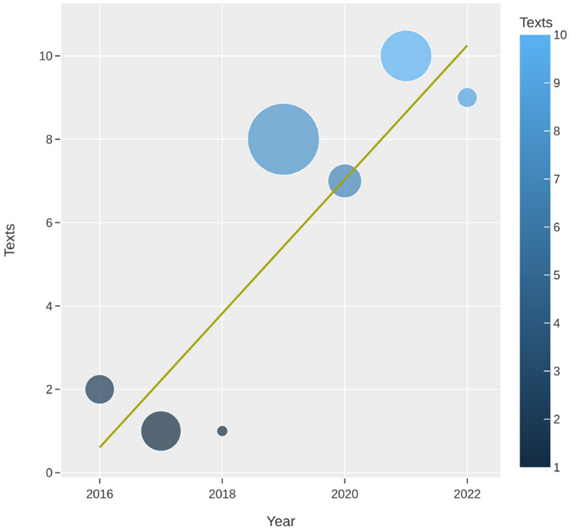

Figure 7. OLS Regression of number of manuscripts per year.

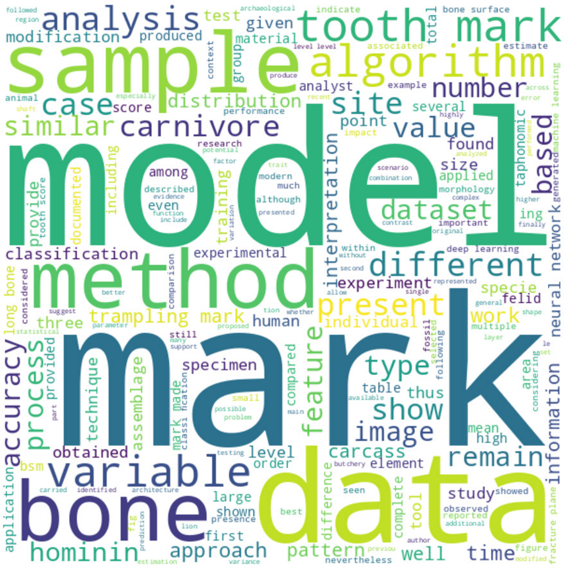

Figure 8. Word cloud of the entire dataset manuscripts.

the classification of cryptic marks like trampling and butchery marks, which have historically caused confusion (Behrensmeyer et al. 1986, Olsen & Shipman 1988) and continue to be challenging (Pineda et al. 2014, Byeon et al. 2019, Pizarro-Monzo & Domínguez-Rodrigo 2020). Predation and tooth marks are also prominent topics in these analyses, and the articles aim to tackle them using computational tools, highlighting the ongoing effort of taphonomits to classify and refine these types of marks. Computational methods allow for a more detailed examination and resolution of these issues (Jiménez-García et al. 2020).

CONCLUSIONS

This initial exploration of the new field through a review reveals a wealth of information. The research production in this field is primarily concentrated in Spain, particularly at the University of Alcalá and the Universitat Rovira i Virgili. These institutions have the highest number of published articles and the most cited papers. The significant increase in studies utilizing computational techniques in taphonomy can mark a pivotal moment for the field's research in the country. Among the databases used, Google Scholar stands out for its ability to retrieve papers from more diverse geographic locations and for including the most cited works. Additionally, we observed that the scientific journals hosting most of the highly cited articles in our analysis prioritize papers with a broad scope, encouraging multidisciplinary work, and promoting innovation. This highlights the utilization of computational tools in addressing taphonomy problems.

The approach utilizing natural language proves helpful even with a small dataset; the coherence index indicates that topic modeling could yield fewer aggregate topic distributions if the number of texts and subjects increases. Additionally, like Scopus, Google Scholar demonstrated its reliability as a source when used with a careful cleaning process, shedding light on a broader diversity of publications.

We have discovered that the field of computational tools in taphonomy is expected to experience an increase in the rate of article publications in the upcoming years. Furthermore,

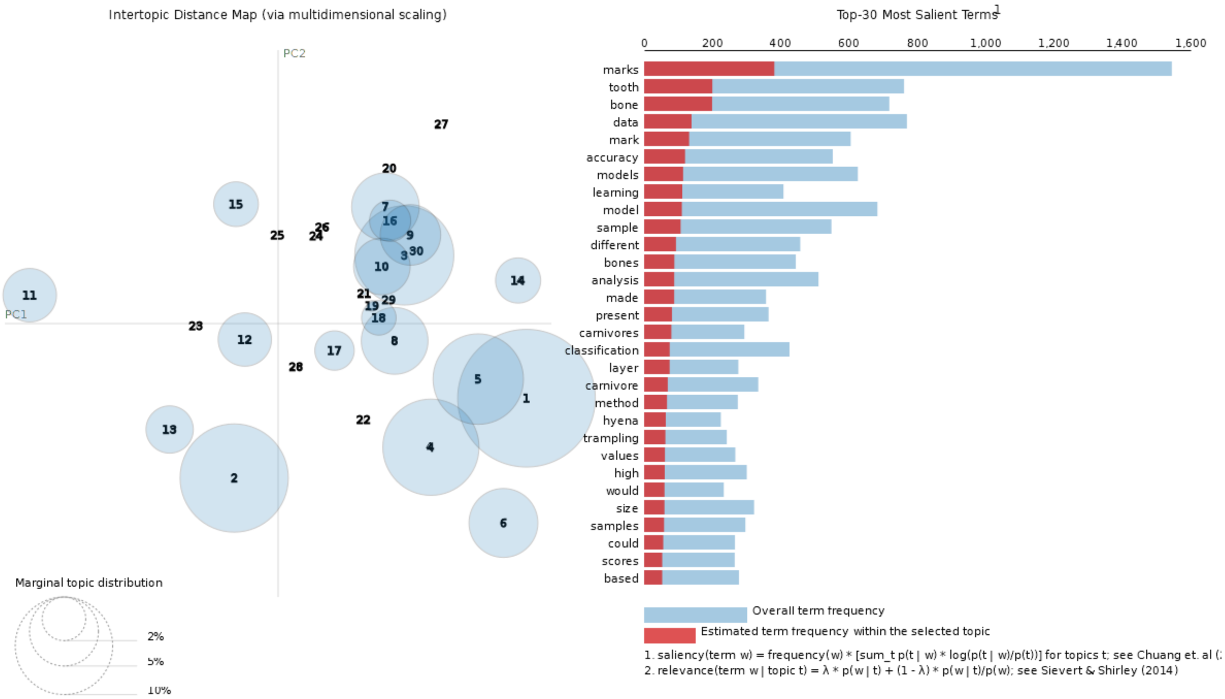

Figure 9. Most frequent terms of the topic modeling analysis using 30 topics, λ = 1.

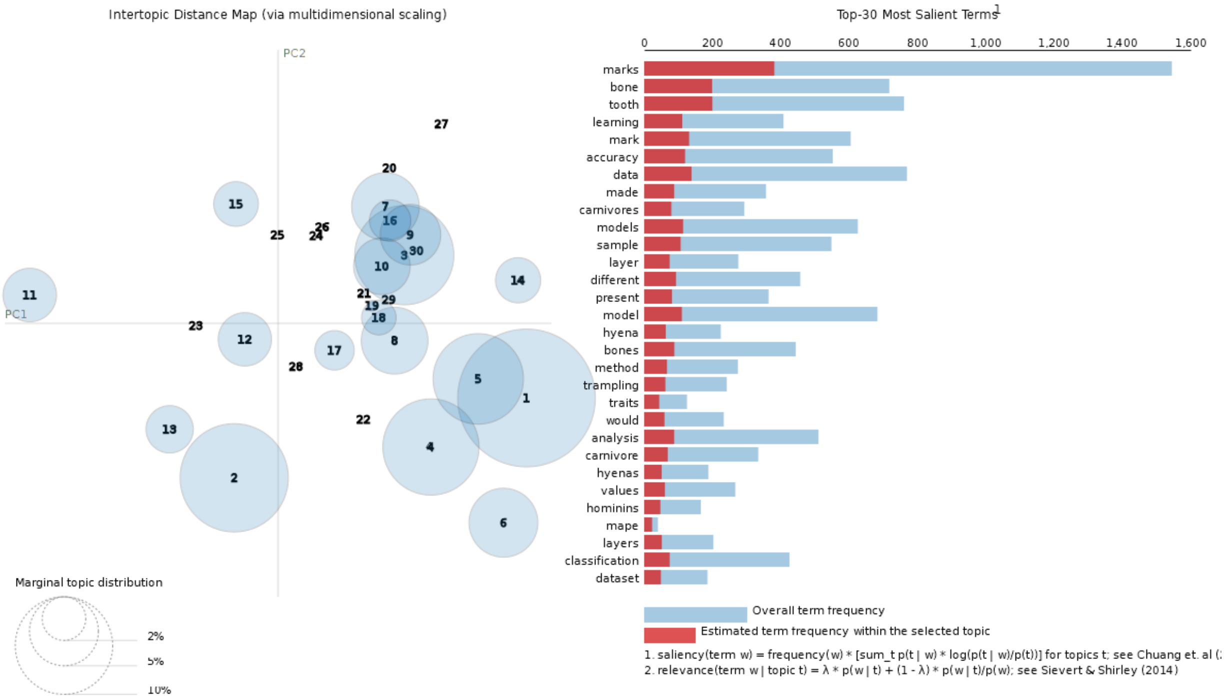

Figure 10. Most frequent terms of the topic modeling analysis using 30 topics, λ = 0.5.

Table III. The coherence ratios change with a shift in the number of topics.

Number of Topic	Coherence Umass Score (LDA)
10	-0.18574806520486858
15	-0.2589698325221563
20	-0.25732291252074746
25	-0.25495115496312887
30	-0.2820409025616922

these computational tools are primarily utilized to identify and address the complexities of BSM, particularly in cryptic scenarios, as evidenced by the exploratory analysis highlighting words such as “marks”, “model”, “bone”, and “data”. Moreover, this focus on studying BSM highlights the potential for researchers in other fields of taphonomy to explore and leverage these computational tools.

Therefore, this study accomplished its aim of gaining a deeper understanding of this emerging field, including its trends, prospects, and potential areas of exploration. While further work remains to be done, it’s crucial first to expand the number of papers available, so that the analyses conducted in this study can be performed with greater precision and robustness.

Acknowledgments

We thank Mr. Eroaldo G. S. Junior for his support on the final development of the web scraping technique and useful discussions in the early stages of this work, and Mr. Zainal Salam for his several valuable suggestions and elucidations to improve the writing during the final stage of writing this manuscript. The authors are grateful to Coordenaçao de Aperfeiçoamento de Pessoal de Nível Superior – CAPES for funding RAL [CNPq; grants: 88887.683402/2022-00] and LAS (process number 88887.602327/2021-00); HIAJR thanks to Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (processes nº E-26/201.371/2021 and E-26/010.002218/2019), Conselho Nacional de Desenvolvimento Científico e Tecnológico -CNPq (process nº 305576/2021-6) and UERJ (Prociencia grant).

REFERENCES

ARRIAZA MC, ARAMENDI J, MATÉ-GONZÁLEZ MÁ, YRAVEDRA J & STRATFORD D. 2021. The hunted or the scavenged? Australopith accumulation by brown hyenas at Sterkfontein (South Africa). Quat Sci Rev 273: 107252.

ARRIAZA MC & DOMINGUEZ-RODRIGO M. 2016. When felids and hominins ruled at Olduvai Gorge: A machine learning analysis of the skeletal profiles of the non-anthropogenic Bed I sites. Quat Sci Rev 139: 43-52.

ATENSTAEDT R. 2012. Word cloud analysis of the BJGP. Br J Gen Pract 62: 148.

BARDE BV & BAINWAD AM. 2017. An Overview of Topic Modeling Methods and Tools. In: International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai: 745-750.

BEEL J & GIPP B. 2009. Google Scholar’s Ranking Algorithm: An Introductory Overview. In: Proceedings of ISSI, Rio de Janeiro: 230-241.

BEHRENSMEYER AK, GORDON KD & YANAGI GT. 1986. Trampling as a cause of bone surface damage and pseudo-cutmarks. Nature 319: 768-771.

BIRD S, LOPER E & KLEIN E. 2001. NLTK Project. Natural Language Toolkit Documentation. Nltk. https://www.nltk.org/. Accessed on November 30, 2022.

BYEON W, DOMÍNGUEZ-RODRIGO M, ARAMPATZIS G, BAQUEDANO E, YRAVEDRA J, MATÉ-GONZÁLEZ MA & KOUMOUTSAKOS P. 2019. Automated identification and deep classification of cut marks on bones and its paleoanthropological implications. J Comput Sci 32: 36-43.

CIFUENTES-ALCOBENDAS G & DOMÍNGUEZ-RODRIGO M. 2019. Deep learning and taphonomy: high accuracy in the classification of cut marks made on feshed and defeshed bones using convolutional neural networks. Sci Rep 9: 18933.

COURTENAY LA, YRAVEDRA J, HUGUET R, ARAMENDI J, MATÉ-GONZÁLEZ MÁ, GONZÁLEZ-AGUILERA D & ARRIAZA MC. 2019b. Combining machine learning algorithms and geometric morphometrics: a study of carnivore tooth marks. Palaeogeogr Palaeoclimatol Palaeoecol 522: 28-39.

COURTENAY LA, YRAVEDRA J, HUGUET R, OLIÉ A, ARAMENDI J, MATÉ-GONZÁLEZ MÁ & GONZÁLEZ-AGUILERA D. 2019a. New taphonomic advances in 3D digital microscopy: A morphological characterisation of trampling marks. Quat Int 517: 55-66.

DOMÍNGUEZ-RODRIGO M. 2018. Distinguishing butchery cut marks from crocodile bite marks through machine learning methods. Sci Rep 8: 5786.

DOMÍNGUEZ-RODRIGO M. 2019. Successful classification of experimental bone surface modifications (BSM) through machine learning algorithms: a solution to the controversial use of BSM in paleoanthropology? Archaeol Anthropol Sci 11: 2711-2725.

DOMÍNGUEZ-RODRIGO M & FERNÁNDEZ-LÓPEZ S, ALCALÁ L. 2011. How can taphonomy be defined in the XXI Century. Journal of Taphonomy 9: 1-13.

DOMÍNGUEZ-RODRIGO M, JUANA S, GALÁN AB & RODRIGUEZ M. 2009. A new protocol to differentiate trampling marks from butchery cut marks. J Archaeol Sci 36: 2643-2654.

DOMÍNGUEZ-RODRIGO M, SALADIÉ P, CÁCERES I, HUGUET R, YRAVEDRA J, RODRÍGUEZ-HIDALGO A, MARTIN P, PINEDA A, MARÍN J, GENÉ C, ARAMENDI J & COBO-SANCHEZ, L. 2017. Use and abuse of cut mark analyses: The Rorschach effect. J Archaeol Sci 86: 14-23.

DONTHU N, KUMAR S, MUKHERJEE D, PANDEY N & LIM WM. 2021. How to conduct a bibliometric analysis: An overview and guidelines. J Bus Res 133: 285-296.

FARIAS MT, ANGELUCI ACB & PASSARELLI B. 2021. Web scraping and data science in applied research in communication: A study on online reviews. Revista Observatório 7: a1en.

FERNÁNDEZ-JALVO Y & ANDREWS P. 2016. Atlas of Taphonomic Identifications 1001+ Images of Fossil and Recent Mammal Bone Modification (Vertebrate Paleobiology and Paleoanthropology Series). Dordrecht: Springer.

FERNÁNDEZ-LÓPEZ SR. 2000. Temas de Tafonomia. Universidad Complutensed e Madrid, 167 p.

FERNÁNDEZ-LÓPES SR. 2005. Alteración tafonómica y tafonomía evolutiva. Bol R Soc Esp Hist Nat 100: 149-175.

FERNÁNDEZ-LÓPEZ SR & FERNÁNDEZ-JALVO Y. 2002. The limit between biostratinomy and fossildiagenesis. Current Topics on Taphonomy and Fossilization: 27-37.

JIANG H, QIANG M & LIN P. 2016. A topic modeling based bibliometric exploration of hydropower research. Renew Sustain Energy Rev 57:226-237.

JIMÉNEZ-GARCÍA B, AZNARTE J, ABELLÁN N, BAQUEDANO E & DOMÍNGUEZ-RODRIGO M. 2020. Deep learning improves taphonomic resolution: high accuracy in differentiating tooth marks made by lions and jaguars. J R Soc Interface 17: 20200446.

KAESLER RL. 1993. A Window of Opportunity: Peering Into a New Century of Paleontology. J Paleontol 67: 329-333.

KHERWA P & BANSAL P. 2019. Topic Modeling: A Comprehensive Review. EAI Endorsed Scal Inf Syst 7: e2.

KUMARESAN U & RAMANUJAM K. 2016. A framework for extraction of journal information from scientific publishers web site. In: 10th International Conference on Intelligent Systems and Control (ISCO), Coimbatore: 1-5.

LI X & LEI L. 2019. A bibliometric analysis of topic modelling studies (2000–2017). J Inf Sci 47: 161-175.

LIN L. 2018. Bias caused by sampling error in meta-analysis with small sample sizes. PLoS ONE 13: e0204056.

MARTIN-MARTIN A, THELWAL M, ORDUNA-MALEA E & LÓPEZ-CÓZAR ED. 2021. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. Scientometrics 126: 871-906.

MARTIN-PEREA DM, COURTENAY LA, DOMINGO MS & MORALES J. 2020. Application of artificially intelligent systems for the identification of discrete fossiliferous levels. PeerJ 8: e8767.

MOED HF, BAR-ILAN J & HALEVI G. 2016. A new methodology for comparing Google Scholar and Scopus. JOI 10: 533-551.

MOHAMMED SH & AL-AUGBY S. 2020. LSA & LDA Topic Modeling Classification: Comparison study on E-books. Indones J Electr Eng Comput Sci 19: 353-362.

NOZURI A. 2005. Google Scholar: The New Generation of Citation Indexes. Libri 55: 170-180.

OLSEN SL & SHIPMAN P. 1988. Surface Modification on Bone: Trampling versus Butchery. J Archaeol Sci 15: 535-553.

PINEDA A, SALADIÉ P, VERGES JM, HUGUET R, CÁCERES I & VALLVERDÚ J. 2014. Trampling versus cut marks on chemically altered surfaces: an experimental approach and archaeological application at the Barranc de la Boella site (la Canonja, Tarragona, Spain). J Archaeol Sci 50: 84-93.

PIZARRO-MONZO M & DOMÍNGUEZ-RODRIGO M. 2020. Dynamic modification of cut marks by trampling: temporal assessment through the use of mixed-effect regressions and deep learning methods. Archaeol Anthropol Sci 12: 1-13.

PIZARRO-MONZO M, ORGANISTA E, COBO-SÁNCHEZ L, BAQUEDANO E & DOMÍNGUEZ-RODRIGO M. 2021. Determining the diagenetic paths of archaeofaunal assemblages and their palaeoecology through artificial intelligence: an application to Oldowan sites from Olduvai Gorge (Tanzania). J Quat Sci 37: 543-557.

ŘEHŮŘEK R & SOJKA P. 2010. Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: University of Malta, 46-50.

RICHARDSON L. 2007. Beautiful soup documentation. April.

ROVIRA C, CODINA L & LOPEZOSA C. 2021. Language Bias in the Google Scholar Ranking Algorithm. Future internet 13: 31.

SANTANA TS & BRAGA AH. 2020. Uma Análise Cienciométrica das Publicações do Congresso da Sociedade Brasileira de Computação na Perspectiva das Mulheres na Computação. In: Anais do XIV Women in Information Technology, Cuiabá, Brasil. Porto Alegre: SBC, 279-283.

SANTOS BS, SILVA I, RIBEIRO-DANTAS MC, ALVES G, ENDO PT & LIMA L. 2020. COVID-19: A scholarly production dataset report for research analysis. Data in Brief 32: 106178.

SCACCIA JP & SCOTT VC. 2021. 5335 days of Implementation Science: using natural language processing to examine publication trends and topics. Implement Sci 16: 47.

SAN ROMÁN LS, LÓPEZ SRF, GOY AG, BASTARRECHE AM, BENITEZ JMT, HEVIA GM, SANDOVAL J, JIMÉNEZ CD, RUIZ PB, SÁEZ FO & RIVAS AL. 1998. Historia del conocimiento de los ammonites (moluscos fósiles) del Jurásico de España. LLULL 21: 517-546.

SPICHAK E, DIETL W & ERNST MD. 2012. A type system for regular expressions. In: Proceedings of the 14th Workshop on Formal Techniques for Java-like Programs. Association for Computing Machinery: 20-16.

WANG B, SUN R, YANG X, NIU B, ZHANG T, ZHAO Y, ZHANG Y, ZHANG Y & HAN J. 2022. Recognition of Rare Microfossils Using Transfer Learning and Deep Residual Networks. Biology 12: 16.

WANG J & MANIRUZZAMAN M. 2022. A global bibliometric and visualized analysis of bacteria-mediated cancer therapy. Drug Discovery Today 27: 103297.

XU Y, DAI Z, WANG J, LI Y & WANG H. 2020. Automatic Recognition of Palaeobios Images Under Microscope Based on Machine Learning. IEEE Access 8: 172972-172981.

ZHAO B. 2017. Web Scraping. In: SCHINTLER LA & MCNEELY CL. (Eds), Encyclopedia of Big Data. Springer International Publishing: 1-3.

How to cite

LEONI RA, ALVES-SILVA L & ARAÚJO-JÚNIOR HI. 2024. Overview of computational methods in taphonomy based on the combination of bibliometric analysis and natural language. An Acad Bras Cienc 96: e20230784. DOI 10.1590/0001-3765202420230789.

Manuscript received on July 12, 2023;
accepted for publication on November 6, 2023

RONALDO A. LEONI
https://orcid.org/0000-0002-4169-8922

LAÍS ALVES-SILVA
https://orcid.org/0000-0001-9692-9989

HERMÍNIO ISMAEL DE ARAÚJO-JÚNIOR
https://orcid.org/0000-0003-4371-0611

Programa de Pós-Graduação em Geociências, Universidade
do Estado do Rio de Janeiro, Rua São Francisco Xavier,
524, Maracanã, 20950-000, Rio de Janeiro, RJ, Brazil

Correspondence to: **Ronaldo Araujo Leoni**
E-mail: ronaldoaleoni@gmail.com

Author contributions

Conceived and designed the analysis LEONI RA, ALVES-SILVA L, Collected the data, Performed the analysis, Wrote the paper LEONI RA, Supported the manuscript writing ALVES-SILVA L, ARAUJO-JUNIOR HIS, Supervised the project ARAUJO-JUNIOR HIS