CHAPTER 7

A FRAMEWORK FOR COLLECTING, PROCESSING, AND ANALYZING SCIENTIFIC DATA ON SOCIAL MEDIA

Thiago Magela Rodrigues Dias
Department Computer Science, CEFET-MG, *Brazil.*
ORCID: *https://orcid.org/0000-0001-5057-9936*
Email: *thiagomagela@cefetmg.br*

Rafael Gonçalo Ribeiro
Department Computer Science, CEFET-MG, *Brazil.*
ORCID: *https://orcid.org/0000-0003-0646-6605*

Patrícia Mascarenhas Dias
Department Computer Science, UEMG, *Brazil.*
ORCID: *https://orcid.org/0000-0002-8448-6874*

ABSTRACT

Given the increasing use of social media, it became imperative to understand the dissemination and discussion of scientific publications on these online platforms. The analysis of these data on interaction and circulation of scientific research was investigated in altmetrics studies and provided valuable information on how science was perceived and shared by the general public. The objective of this study was to propose a platform for the collection and analysis of social data related to scientific publications, with a focus on the video-sharing platform YouTube. By collecting data from YouTube, the platform sought to understand how scientific publications were disseminated and discussed on social

media. The Social4Science solution facilitated the acquisition of social data from YouTube and its correlation with scientific data from publications, enabling the analysis of multiple metrics. This methodological approach facilitated the identification of trends and patterns in discourse concerning scientific publications on social media. The findings indicated that the proposed platform held considerable promise in fostering a more profound comprehension of the interaction between science and the public. Furthermore, it had the potential to generate new avenues for future research in this domain. It was imperative to comprehend the manner in which scientific publications were received and discussed on social media platforms. This understanding was crucial for effective scientific communication and for fostering connections between the scientific community and the general public. The proposed platform contributed to this understanding, allowing researchers and professionals in the field to identify opportunities for engagement and develop effective strategies for scientific dissemination.

KEYWORDS: scientific production, social media, altmetrics, open data, bibliometrics

HOW TO CITE: Dias, T. M. R., Gonçalo Ribeiro, R., & Dias, P. M. (2025). A framework for collecting, processing, and analyzing scientific data on social media. In A. Semeler (Ed.), Artificial Intelligence and Data Science Practices in Scientific Development, *Advanced Notes in Information Science, volume 8* (pp. 191-207). ProMetrics: Tallinn, Estonia. DOI: 10.47909/978-9916-9331-4-5.115.

1 INTRODUCTION

The dissemination of scientific discoveries and the processes that facilitate it play a foundational role in the advancement of society and culture. The establishment of effective communication between the academic community and society is imperative, given that knowledge and research are intended to benefit society as a whole. Consequently, it is imperative that the method by which results are disseminated is congruent with the needs and expectations of the public. This is essential to establish a substantial and pertinent relationship between science and society in general (Neto, 2018). In this context, YouTube has demonstrated its significance as a primary platform for scientific dissemination on the Internet. As the most prominent global video-sharing platform, it encompasses a broad spectrum of content, encompassing numerous subjects and themes. In the context of Brazil, YouTube boasts a substantial and engaged audience, thereby establishing itself as a conducive platform for the dissemination of scientific knowledge (Da Fonseca & Bueno, 2021). The dissemination of scientific knowledge via YouTube has facilitated the creation of educational videos, debates, interviews, and practical demonstrations, thereby promoting interaction and dialogue between researchers and interested audiences. According to Reale and Martyniuk (2016), the dissemination of scientific knowledge via YouTube is an effective medium for democratizing scientific knowledge.

The analysis of scientific articles mentioned in YouTube videos offers the opportunity to collect a wide range of relevant data. This data may include the title of the scientific article, the names of the authors, the name of the journal in which the article was published, the year of publication, and the number of citations received, among other aspects of interest. This information is crucial for comprehending the interaction between the digital platform and scientific production, as well as for examining the impact and dissemination of scientific research on social media. The extraction of these data can facilitate the acquisition of insights regarding citation trends, the most frequently cited areas of research, and the subjects most prevalent in scientific videos on YouTube. This analysis also allows for the exploration of the connection between scientific dissemination and the academic framework, with the identification of the relevance and

influence of the scientific articles mentioned. A close examination of the citations in the videos reveals potential discrepancies between scientific research and its public dissemination, underscoring areas that merit heightened attention in the realm of scientific communication. Consequently, the extraction of data from scientific articles cited in YouTube videos signifies a pivotal approach to comprehending the nexus between scientific production and its dissemination in the digital domain. This contributes to a more comprehensive understanding of the propagation of scientific knowledge and its interactions with the general public. For instance, it is possible to identify emerging trends mentioned in the videos, thereby highlighting the most prominent and relevant topics in the realm of online scientific dissemination. Furthermore, it is possible to assess the influence of authors and journals, identifying those that are most mentioned and recognized on the platform.

This analysis facilitates a more profound comprehension of the dynamics underlying the dissemination of scientific research in the digital environment. A further critical component of this investigation entails the analysis of the relationship between the popularity of videos on YouTube and the number of citations received by the scientific articles mentioned in these videos. This correlation may reveal the influence of online videos on the dissemination and recognition of academic research. Comprehension of this relationship is essential for obtaining a comprehensive understanding of the interaction between scientific dissemination and the impact of research. A thorough analysis of the collected data may reveal gaps in scientific communication, indicating areas where there is a disconnect between scientific production and its online dissemination. These discrepancies may result in initiatives aimed at enhancing communication and public engagement, thereby fostering a more profound comprehension and esteem for scientific endeavors. In view of the aforementioned points, the objective of this study is to propose an innovative computational platform for the collection, processing, and analysis of scientific data on social media. It is imperative to underscore the utilization of the term "social media" in this context, as opposed to the term "digital social networks." The former term is more comprehensive, encompassing a broader array of online platforms that facilitate the creation and dissemination

of content, in addition to fostering interactions and connections between users.

The proposed platform, designated Social4Science, aims to address a significant gap in scientific research by providing an efficient tool to explore the vast universe of social media and understand how scientific information is disseminated, discussed, and perceived by the general public. The collection and analysis of data from this platform has enabled the acquisition of significant insights into science-related trends, patterns, and interactions. Moreover, the platform encompasses a wide range of functionalities that facilitate the identification of influencers, the analysis of the impact and relevance of scientific publications, the detection of emerging themes, and other significant analyses. In light of the aforementioned data, researchers will be in a position to make informed decisions, develop more effective dissemination strategies, and improve communication between the academic community and society. Consequently, Social4Science signifies a substantial methodology for investigating the possibilities of social media within the domain of scientific research. It provides a thorough and detailed perspective on the interactions between science and society, enhancing scientific communication, fostering inclusive dialogue, and establishing a robust connection between academia and the general public. The objective of the tool is to collect and analyze data from social media platforms, such as YouTube, with the aim of understanding how scientific publications are disseminated and discussed on these digital forums. Specifically, the objective of this study is to investigate the characteristics of videos published on YouTube that reference a Digital Object Identifier (DOI), with the aim of identifying relevant trends and patterns.

The objective of this study is to obtain results on how science is communicated and discussed in the online environment of YouTube. To this end, data will be collected from YouTube and analyzed using various techniques. By examining the characteristics of videos that contain DOIs, understanding of the manner in which scientific information is disseminated, the subjects that are addressed, and the manner in which the public engages with this content can be enhanced. Indicators of online attention have been a subject of discussion in the context of altmetric studies, which focus on understanding the social impact of research results on the social web (Araújo, 2020). These analyses can be

useful for researchers, journal editors, and other professionals involved in scientific dissemination, as they can help understand better how science is perceived and shared by the general public and to identify opportunities to increase the visibility of publications. Research that has been developed with these more contextual approaches is increasing in the extant literature, and it is indicative of the concern in the altmetric field to contribute to the deepening of the analysis and investigation of where and how articles are used by different communities that interact with them online (Araújo, 2020).

2 METHODOLOGY

This study employed the Altmetric portal, accessible via the Altmetric Explorer platform, as a tool to search for scientific publications that were cited in videos published on YouTube. The relationship between videos and scientific articles is established when a video mentions an article using the DOI, which is usually included in the video description. The utilization of the DOI as a unique identifier facilitates the precise linkage of a particular video to a corresponding scientific article. A search of the Altmetric portal for YouTube videos that mention DOIs revealed a dataset for analysis and study of the interactions between social media and scientific research. This approach of searching for references to scientific articles in YouTube videos using the DOI is an effective way to identify the presence and reach of science on this platform. The Altmetric Explorer platform offers resources that facilitate the collection and processing of data, enabling detailed analyses to be carried out on the characteristics of videos and citations of scientific articles. The entire data extraction process is initiated from a relationship extracted from Altmetric, containing a file with the video identifier and the DOI of a publication. The Social4Science platform receives this relationship as input and begins the entire data collection and analysis process, divided into two segments:

1. Social analysis: Data collection from YouTube videos.

2. Bibliometric analysis: Data collection from scientific articles.

The architectural design of the proposed platform is illustrated in Figure 1.

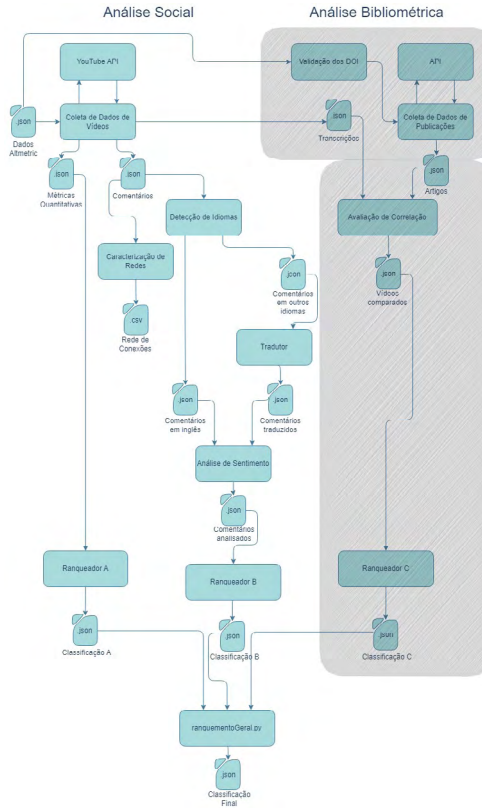

Figure 1. General architecture of the Social4Science platform.

The process of data collection and processing carried out by the platform commences with the input of a file provided by Altmetric, containing the video identifiers and the DOIs of the scientific articles. These DOIs are employed in the "bibliometric analysis" stage, while the video identifiers are employed in the "social analysis" stage. The subsequent "bibliometric analysis" stage entails the utilization of DOIs to procure pertinent information regarding the publications, including title, authors, journal

of publication, year of publication, and the number of citations received. These bibliometric data are essential for understanding the relevance and impact of the scientific articles mentioned in the YouTube videos. Consequently, the "social analysis" stage employs the video identifiers to investigate the social and interaction aspects associated with the videos that reference the scientific articles. This social analysis encompasses the identification of trends; the assessment of video popularity; the evaluation of user interactions, such as likes, shares, and comments; and the identification of relevant influencers or channels in scientific dissemination. The platform enables a comprehensive and in-depth approach to understanding the impact and dissemination of science on social media, especially on YouTube, by separating the bibliometric analysis and social analysis stages. The integration of bibliometric information and social data facilitates the acquisition of significant insights regarding the reception, discussion, and dissemination of scientific publications on this platform. This integration contributes to the advancement of scientific dissemination and the cultivation of a deeper understanding of the interactions between science and society.

In the context of "social analysis," video data are collected through the utilization of a publicly accessible YouTube Application Programming Interface (API), coinciding with the generation of specific data extracts. These metrics can be calculated and exported to other analysis and visualization tools, enabling further in-depth analysis. As a case in point, the sets comprising quantitative data from the videos, including the number of views, comments, and likes for each video, are emphasized. In addition, sets containing data from the channels in which the videos were published, the interaction networks identified from the comments on each video, extracts from the video descriptions, the transcriptions of each audio, and a set of standardized data in English from all the comments extracted are highlighted. In "bibliometric analysis," the set of DOIs is examined via API to ensure their validity. In the event that a DOI is found to be valid, the associated data are directed to the OpenAlex API, thereby facilitating the retrieval of information concerning the article in question. This includes details such as the article's title, authorship, year of publication, abstract, keywords, and the journal in which it was published, among other pertinent information. To

complement the data, a new request for the same DOI is sent to the OpenCitations API, retrieving the article's citations.

This comprehensive array of data is stored in data extracts that are also subject to analysis using various metrics implicit in the platform itself. These data extracts are made available in formats that can be imported by other analysis and visualization tools. Quantitative data play a fundamental role in the platform, allowing for different types of ranking and the analysis of correlations between social analytics and bibliometric analyses. These quantitative metrics offer valuable insights into the popularity, engagement, and reach of the videos and scientific publications referenced therein. Conversely, the datasets comprising textual information from videos, including titles, comments, descriptions, and transcripts, are correlated with the textual data from scientific publications, such as titles, abstracts, and keywords. In this context, correlation measures, such as the Levenshtein distance or the calculation of cosine similarity, are adopted to explore the relationships between the texts. The Levenshtein distance is a metric that calculates the difference between two sequences of characters, such as video titles and scientific publication titles. This measure enables the assessment of the thematic affinity or dissimilarity between the texts, thereby providing insights into the thematic proximity between the videos and the publications. The cosine similarity is a measure that quantifies the similarity between two-word vectors, such as the terms present in video comments and the keywords of scientific publications. This allows for the identification of semantic associations and relationships between the texts. The Social4Science platform employs correlation measures to reveal connections between the content of videos and scientific publications, identify thematic patterns, and explore how information is transmitted and discussed on social media.

Therefore, the integration of quantitative and textual data furnishes a thorough and enlightening analysis, enabling comprehension of the quantitative and textual dimensions implicated in the propagation and discourse of scientific publications on YouTube. To initiate the case study, a set containing 65,534 DOIs that had YouTube video citations at the time was collected from the Altmetric platform in March 2022. A series of verifications was conducted on a set of DOIs to ascertain the characteristics of scientific publications. A subsequent analysis of the publication

type revealed that the majority of the publications were classified as articles (94.9%), followed by a smaller proportion of books (3%) and book chapters (1%). It is also noteworthy that a total of 45 datasets were referenced.

3 RESULTS

Social analysis entails the examination of data derived from videos, including metrics such as the number of likes, views, and comments. Conversely, bibliometric analysis encompasses quantitative data derived from articles, including DOI validation, the number of citations received by other articles, and the number of videos that mention the article in question. The analysis of these data points enables the discernment of trends and patterns in discussions concerning scientific publications on social media platforms. For instance, it is possible to ascertain the most popular publications on these platforms, identify the topics that generate the most discussions, and determine the primary influencers in this context. Through bibliometric analysis, taking into account the date of data collection, the publication period of the articles mentioned in the videos was presented in chronological order (Figure 2).

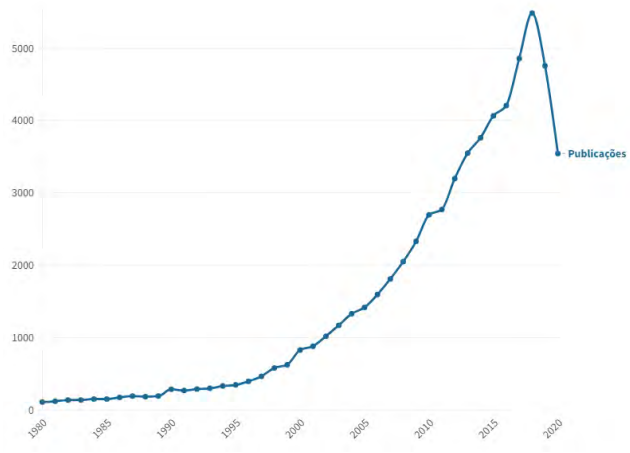

Figure 2. Publication period of the mentioned articles.

A subsequent analysis of the dataset revealed that the oldest article identified was published in 1980. A notable increase in the number of scientific articles is evident over time, with an even more pronounced trend from the year 2000 onwards. Concurrently, the utilization of DOI in scientific articles experienced a marked increase. A substantial surge in the number of scientific articles mentioned on YouTube was observed beginning in 2006, reaching its zenith in 2018. This growth can be attributed to a series of factors, such as advances in technology and research tools, broader access to scientific information, and increased collaboration between researchers on a global scale. The use of social media as a mechanism for disseminating research results has also played a significant role in this development. Additionally, the representativeness of the primary journals in which the articles were published could be ascertained. The objective of this analysis was to quantify the articles published in each journal, with a focus on identifying those that were most frequently mentioned in YouTube videos during the specified period (Figure 3).

Figure 3. Representation of journals in articles referenced in videos.

The following prestigious journals have been observed: *Nature*, *The American Journal of Clinical Nutrition*, *PLoS ONE*, *Nutrients*, *The Journal of Strength and Conditioning Research*, and *Science*, among others. These journals have gained international recognition for their editorial quality and the scientific rigor of their publications. It is noteworthy that specific domains of knowledge exhibit a higher prevalence of YouTube's utilization as a medium for the dissemination of scientific articles. This phenomenon can be attributed to several factors, including the nature of these areas, which are more readily transmitted through videos. It is important to note that certain regions may exhibit a heightened demand for direct and accessible communication, particularly in cases involving topics of public interest (Figure 4).

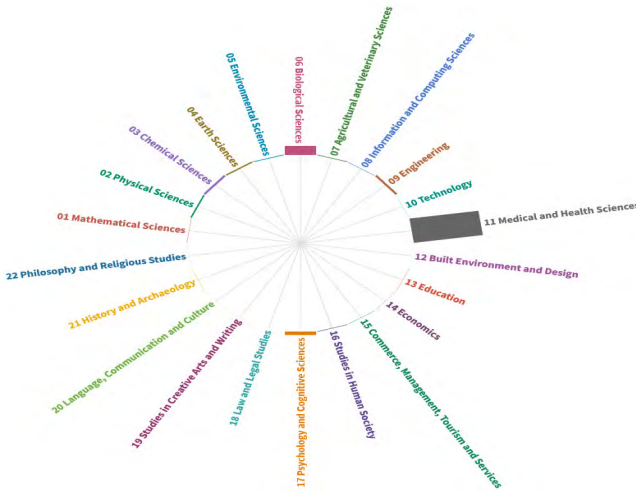

Figure 4. Predominant areas of the mentioned articles.

The utilization of YouTube as a scientific dissemination platform in these domains facilitates more dynamic and interactive communication, thereby providing a more engaging learning experience. Such videos may include a variety of content, such as practical demonstrations, interviews with experts, debates, and analyses of scientific articles, among other materials designed to

stimulate the interest and curiosity of the public. A close examination of the classification of the knowledge areas of the articles reveals a substantial concentration in two primary domains, as evidenced by the data collected. The analysis revealed that 69% of the articles were from the field of "Medical and Health Sciences," while 11.5% were from "Biological Sciences." These two areas encompass approximately 80% of the entire set of articles studied. This concentration in the domains of "medical and health sciences" is unsurprising, as these disciplines are inherently associated with human health and exert a substantial influence on individuals' lives. Scientific dissemination in these areas is of particular pertinence, as it facilitates the dissemination of crucial information regarding medical treatments, advancements in research, and disease prevention to the general public. The field of "biological sciences" also has a significant concentration of articles mentioned on YouTube. This phenomenon can be elucidated by the paramount significance of these studies in comprehending life and its biological entities. Topics related to the biological sciences, such as genetics, evolution, ecology, and biotechnology, have the potential to arouse the interest and curiosity of a wide audience. This, in turn, can contribute to the dissemination of content related to these subjects on YouTube. It is imperative to acknowledge that, despite the predominance of these two primary domains, other fields of knowledge are also represented in the articles that have been disseminated on YouTube, albeit to a more limited extent.

A number of additional bibliometric analyses were also conducted, including the validation of the DOIs cited, with the objective of verifying the authenticity of the publication. Subsequent to this stage, a collection and analysis of the data contained within the titles, abstracts, and keywords was undertaken, along with information regarding the number of citations of these articles by other publications. Additionally, data from the journals in which the publications were disseminated should be considered, including the impact factor and the Qualis. These data are systematically collected by public APIs from a variety of sources. The social analyses are predicated on information extracted from videos published on YouTube that include a DOI. This approach emphasizes the utilization of YouTube's public API for data retrieval, enabling the acquisition of information directly from the platform. The process of extracting data from YouTube is entirely

automated, commencing with the initial list provided, in which the video identifiers are extracted and the requests are made to the YouTube API. A subsequent analysis of the collected data revealed that the videos were classified into categories. These categories refer to the channels in which these videos are published (Figure 5).

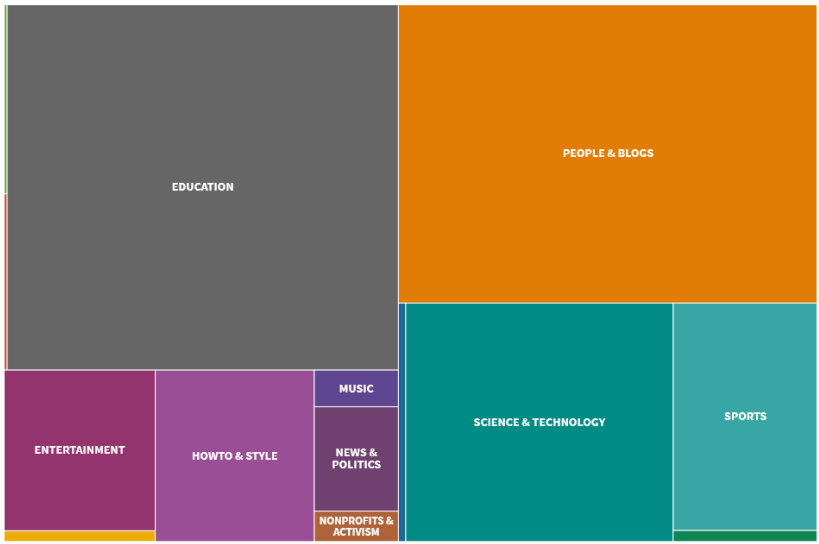

Figure 5. Category of channels where videos are published.

A survey of video content reveals that the majority of channels that present videos with DOIs primarily fall into the categories of "education," "people and blogs," and "science and technology." One hypothesis for the greater representation of the "education" category may be related to classes or dissemination of study results. To achieve a more profound comprehension of this categorization and its repercussions, it is imperative to undertake a thorough examination of the representation of these channels, taking into account the number of subscribers, the quantity of videos published, and the date the channel was registered on YouTube. This information can yield additional results and further enrich the analyses. In addition, the data from "social analysis" indicate

that comment networks are established through the developed platform. A subsequent analysis of the comments associated with each video is then conducted, with the objective of identifying the connections between the various channels. Therefore, a set of comments can be utilized to facilitate an analysis of the interactions between channels, with consideration given to the comments that are made or received by them. This network analysis approach facilitates the acquisition of significant findings regarding the dynamics of interactions between channels in the context of the comments.

In addition to the characterized networks, several other quantitative analyses are performed that aggregate relevant information. A comprehensive set of data, including metrics such as the number of views, comments, likes, duration, and language of the videos, is considered. These metrics offer valuable insights into the reach, engagement, and characteristics of the videos, thereby facilitating a more comprehensive understanding of their relevance and impact on the platform. The content of the videos is also the object of study, covering elements such as the title, description, and audio transcription. These elements are of paramount importance, as they facilitate numerous analyses aimed at correlating the content of the videos with the data from the scientific articles, which are also collected, such as title, abstract, and keywords. These analyses facilitate a more profound comprehension of the subjects addressed in the videos and the identification of relationships between the content of the videos and the content of the scientific articles mentioned. Consequently, they contribute to a comprehensive and contextualized analysis of scientific dissemination.

4 CONCLUSION

The Social4Science platform, as delineated in this study, facilitates the aggregation and examination of scientific data derived from social media, yielding significant insights concerning the propagation of scientific content. Through the analysis of these data, it is possible to identify trends, patterns, and knowledge gaps in discussions about scientific publications on social media. This instrument offers researchers and professionals in the scientific field crucial information, enabling them to adjust

communication strategies and promote scientific knowledge, thereby establishing a more effective connection with the general public. The data collected by the proposed platform can be used to establish significant correlations between different variables. These correlations provide a more profound understanding of the relationship between the popularity of a video on YouTube and the characteristics of the scientific article it references, taking into account factors such as research area, publication type, and country of origin. These analyses facilitate a comprehensive examination of the impact and repercussions of scientific publications on social media, thereby contributing to the understanding of the process of dissemination and the reach of scientific knowledge. The complete tool, developed with the source code of all the framework modules, will be made available in a GitHub repository for any community of interest.

Conflict of interest

The authors declare no potential conflicts of interest.

Contribution statement

Conceptualization: Thiago Magela Rodrigues Dias, Rafael Gonçalo Ribeiro, Patrícia Mascarenhas Dias
Data curation: Thiago Magela Rodrigues Dias, Rafael Gonçalo Ribeiro
Formal Analysis: Thiago Magela Rodrigues Dias, Rafael Gonçalo Ribeiro, Patrícia Mascarenhas Dias
Methodology: Thiago Magela Rodrigues Dias, Rafael Gonçalo Ribeiro, Patrícia Mascarenhas Dias
Writing – Original Draft: Rafael Gonçalo Ribeiro
Writing – Review and Editing: Thiago Magela Rodrigues Dias, Rafael Gonçalo Ribeiro, Patrícia Mascarenhas Dias

Statement of data consent

All codes developed to build the platform can be noted in the following repository: *https://github.com/RafaelGoncalo/social4ScienceCLI*.

REFERENCES

Araújo, R. F. (2020). Communities of attention networks: Introducing qualitative and conversational perspectives for altmetrics. *Scientometrics*, *124*(3), 1793–1809. *https://doi.org/10.1007/s11192-020-03566-7*

Da Fonseca, A. A., & Bueno, L. M. (2021). Breve panorama da divulgação científica brasileira no YouTube e nos podcasts. *Cadernos de Comunicação*, *25*(2). *https://doi.org/10.5902/2316882X63121*

Neto, J. R. S. (2018). Alcance da divulgação científica por meio do YouTube: estudo de caso no canal Meteoro Brasil. *Múltiplos Olhares em Ciência da Informação*, *8*(2).

Reale, M. V., & Martyniuk, V. L. (2016). Divulgação Científica no Youtube: a construção de sentido de pesquisadores nerds comunicando ciência. In *Congresso brasileiro de ciências da comunicação* (vol. 39, pp. 1–15).