Social4Science: Mapping Trends and Patterns in Discussions of Scientific Publications on Social Media

Rafael G. P. Ribeiro[1], Thiago M. R. Dias[1], Patrícia M. Dias[2], Emerson de Sousa Costa[1]

[1]*Federal Center for Technological Education of Minas Gerais*
Rua Álvares de Azevedo 400 - Bela Vista – Zip-Code: 35503-822 - Divinópolis - MG - Brazil
rafael.goncalo.ribeiro@gmail.com, thiagomagela@cefetmg.br
[2]*State University of Minas Gerais*
Av. Paraná, 3001 - Jardim Belvedere – Zip-Code: 35501-170 - Divinópolis – MG - Brazil.
patriciamdias@gmail.com

Abstract. With the growing use of social media, it has become increasingly important to understand how scientific publications are disseminated and discussed on these online platforms. Analyzing this data on the interaction and circulation of scientific research has been investigated in altmetrics studies and can provide valuable information on how science is perceived and shared by the general public. This work aims to propose a platform for collecting and analyzing social data related to scientific publications, with a focus on the video-sharing platform YouTube. By collecting data from YouTube, the platform seeks to understand how scientific publications are disseminated and discussed on social media. In the solution developed, called Social4Science, it is possible to obtain social data from YouTube and correlate it with scientific data from publications. This approach makes it possible to identify trends and patterns in discussions about scientific publications on social media. The results obtained show that the proposed platform is very promising in providing a deeper understanding of the interaction between science and the public, and also opens up possibilities for future studies on this subject.

Keywords: Altmetrics, Data Analysis, Scientific Communication.

1 Introduction

The sharing of discoveries and the process of scientific dissemination play a fundamental role in social and cultural advancement. It is essential to establish effective communication between the academic community and society, since knowledge and research are intended to benefit society as a whole. Therefore, it is crucial that the way in which results are transmitted is aligned with the needs and expectations of the public, in order to establish a solid and relevant relationship between science and society in general [1]. In this context, YouTube has proven to be a highly relevant platform for scientific dissemination on the Internet. As the largest video sharing site in the world, it hosts a wide variety of content, covering several areas and themes. In Brazil, in particular, YouTube has a large and engaged audience, making it a favorable environment for scientific dissemination to reach a wide audience [2]. This platform offers a unique opportunity for scientists and science communicators to reach a large number of people and share knowledge in an accessible and engaging way. The dissemination of science through

YouTube enables the creation of educational videos, debates, interviews and practical demonstrations, promoting interaction and dialogue between researchers and interested audiences. In addition, the audiovisual nature of YouTube allows for more dynamic and visually appealing communication, contributing to arousing people's interest and curiosity in relation to science.

For authors Reale and Martyniuk [3], scientific dissemination through YouTube is an excellent tool 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 valuable for understanding 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 this data can provide insights into citation trends, the most mentioned research areas and the topics most covered in scientific videos on YouTube. This analysis also allows us to explore the connection between scientific dissemination and the academic framework, identifying the relevance and influence of the scientific articles mentioned through the use of rankers. Furthermore, by examining the citations in the videos, it is possible to identify potential gaps between scientific research and its public dissemination, highlighting areas that deserve greater attention in scientific communication. Therefore, extracting data from scientific articles mentioned in YouTube videos represents a promising approach to understanding the relationship between scientific production and its reach in the digital sphere, contributing to a more comprehensive understanding of the dissemination of scientific knowledge and its interactions with the general public.

For example, it is possible to identify emerging trends mentioned in the videos, highlighting the most prominent and relevant topics in online scientific dissemination. In addition, it is possible to assess the influence of authors and journals, identifying those that are most mentioned and recognized on the platform. This analysis allows us to better understand the dynamics of the dissemination of scientific research in the digital environment.

Another important aspect is the analysis of the relationship between the popularity of videos on YouTube and the number of citations received by scientific articles mentioned in these videos. This correlation can reveal the influence of online videos on the dissemination and recognition of academic research. Understanding this relationship contributes to a more complete view of the interaction between scientific dissemination and the impact of research.

In addition, the analysis of the data collected can reveal gaps in scientific communication, indicating areas where there is a disconnect between scientific production and its online dissemination. These gaps can direct efforts to improve communication and public engagement, promoting greater understanding and appreciation of science.

In view of the above, this study has the general objective of proposing an innovative computational platform for the collection, processing, and analysis of scientific data on social media. It is important to emphasize that the term “social media” is used here instead of “social networks,” as it broadly encompasses online platforms that enable the creation and sharing of content, as well as interactions and connections between users.

The proposed platform, called Social4Science, aims to fill a significant gap in scientific research by providing an efficient tool for exploring the vast universe of social media and understanding how scientific information is disseminated, discussed, and perceived by the general public. By collecting and analyzing data from these platforms, it is possible to obtain valuable insights into trends, patterns, and interactions related to science.

In addition, the platform encompasses a wide range of functionalities that allow the identification of influencers, the analysis of the impact and relevance of scientific publications, the detection of emerging themes, among other important analyses. Based on this data, researchers will be able to make informed decisions, develop more effective dissemination strategies, and improve communication between the academic community and society.

Therefore, Social4Science represents an innovative and promising approach to exploring the potential of social media in the context of scientific research. It offers a comprehensive and in-depth view of the interactions between science and society, boosting scientific communication, fostering more inclusive dialogue, and establishing a solid bridge between academia and the general public.

The purpose of the tool is to collect and analyze data from social media, such as YouTube, with the aim of

understanding how scientific publications are disseminated and discussed on these platforms. Specifically, we seek to investigate the characteristics of videos published on YouTube that reference a DOI (Digital Object Identifier), identifying relevant trends and patterns.

By collecting data from YouTube and applying analysis techniques, this work aims to obtain results on how science is communicated and discussed in this online environment. By analyzing the characteristics of videos that mention DOIs, we can better understand how scientific information is transmitted, what topics are covered, and how the public interacts with this content.

Indicators of online attention have been discussed in the context of altmetric studies, which focus on understanding the social impact of research results on the social web [4]. These analyses can be useful for researchers, journal editors, and other professionals involved in scientific communication, as they can help to better understand how science is perceived and shared by the general public and to identify opportunities to increase the visibility of publications.

Studies developed with these more contextual approaches are growing in the literature and signal 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 [4].

2 METHODOLOGY

This study used the Altmetric portal through the Altmetric Explorer platform as a tool to search for scientific publications that were cited in videos published on YouTube. This relationship between videos and scientific articles is established when a video mentions an article using the DOI (Digital Object Identifier), which is usually included in the video description.

Using the DOI as a unique identifier allows a specific video to be precisely linked to a corresponding scientific article. By searching the Altmetric portal for YouTube videos that mention DOIs, it was possible to identify and collect data for the analyses and study of 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. In addition, 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. From this relationship extracted from Altmetric, containing a file with the video identifier and the DOI of a publication, the entire data extraction and analysis process is initiated using public APIs and the Python programming language. The Social4Science platform receives this relationship as input and begins the entire data collection and analysis process, divided into two segments:

1) Social Analysis: collection and analysis of data from YouTube videos.
2) Bibliometric Analysis: collection and analysis of data from scientific articles.

The architecture of the Proposed Platform can be seen in Figure 1.

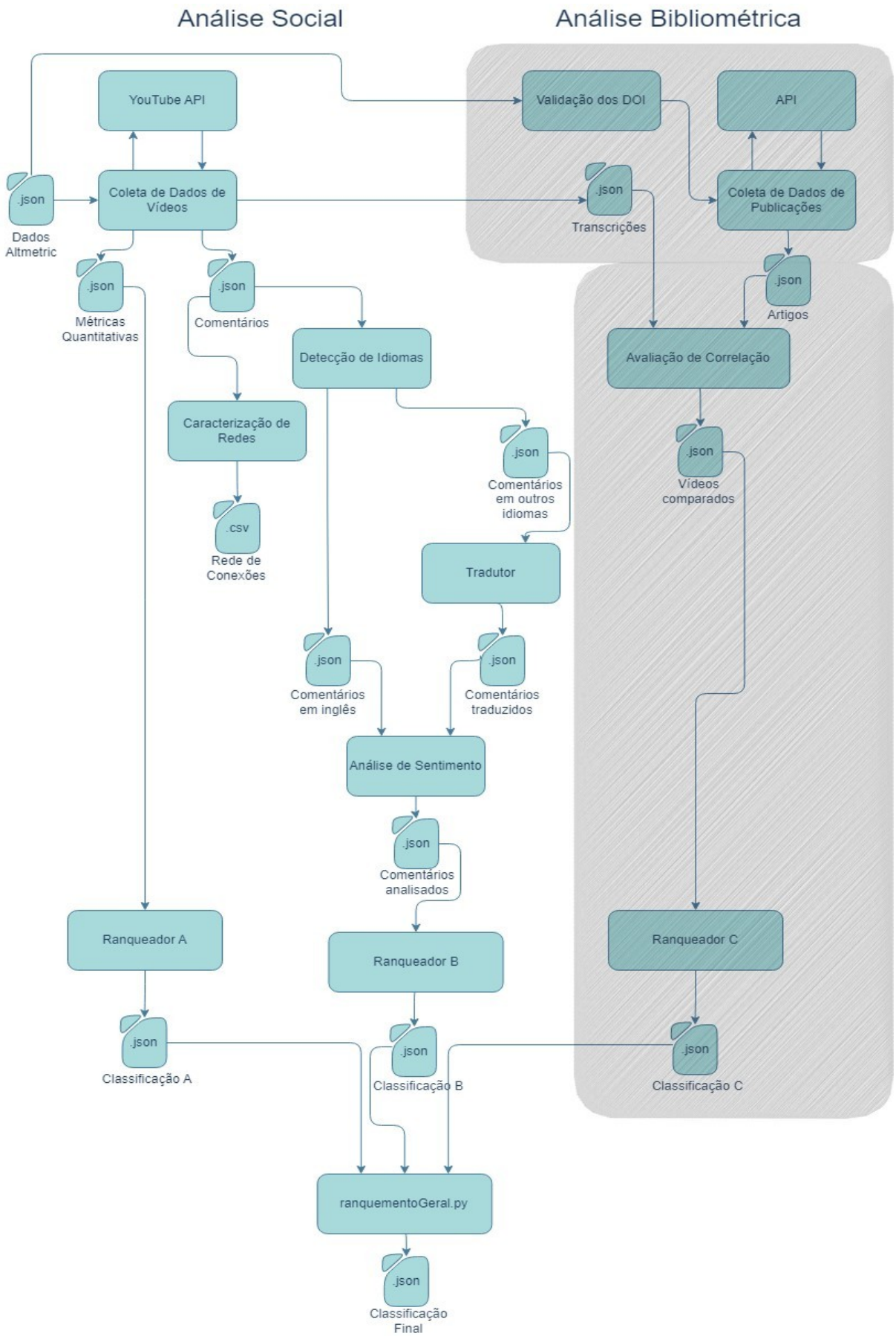

Figure 1: General architecture of the Social4Science Platform.

The data collection and processing process carried out by the platform begins with the input of a file provided by Altmetric, containing the video identifiers and the DOIs of the scientific articles. These DOIs are used in the "Bibliometric Analysis" stage, while the video identifiers are used in the "Social Analysis" stage.

The "Bibliometric Analysis" stage consists of using the DOIs of the scientific articles to obtain relevant information about the publications, such as title, authors, journal of publication, year of publication and number of citations received. This bibliometric data is essential for understanding the relevance and impact of the scientific articles mentioned in the YouTube videos.

In turn, the "Social Analysis" stage uses the video identifiers to explore the social aspects and interactions related to the videos that cite the scientific articles. This social analysis can include identifying trends, analyzing the popularity of the videos, evaluating user interactions, such as likes, shares and comments, and identifying relevant influencers or channels in scientific dissemination.

By separating the bibliometric and social analysis stages, the platform allows for a comprehensive and in-depth approach to understanding the impact and dissemination of science on social media, especially on YouTube. By combining bibliometric information and social data, it is possible to obtain important information about how scientific publications are received, discussed and shared on this platform, contributing to the advancement of scientific dissemination and to the understanding of interactions between science and society.

In Social Analysis (1), video data is collected via a public YouTube Application Programming Interface (API), when some data extracts are generated. They are used to calculate various metrics and can be exported to other analysis and visualization tools, allowing for other forms of in-depth analysis. Examples of these extracts include sets containing quantitative data from the videos, such as the number of views, comments and likes for each video, as well as extracts containing data from the channels on 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, finally, a set of standardized data in English from all the comments extracted.

In Bibliometric Analysis (2), the set of DOIs is checked via API in order to validate them. If it is a valid DOI, its data is sent to the OpenAlex API, thus retrieving information about the article in question, such as its title, authors, year of publication, abstract, keywords, journal of publication, among others. In addition, in order to complement the data, a new request for the same DOI is sent to the OpenCitations API, retrieving the article's citations. This entire data set is stored in data extracts that are also subject to analysis using various metrics implicit in the platform itself and are made available in formats that can be imported by other analysis and visualization tools.

Quantitative data plays a fundamental role in the platform, allowing different types of ranking and the analysis of correlations between Social Analysis and Bibliometric Analysis. These quantitative measures provide valuable input on the popularity, engagement and reach of videos and scientific publications mentioned in them.

On the other hand, the data sets that contain textual information from videos, such as titles, comments, description and transcription, are correlated with the textual data of scientific publications, such as titles, abstracts and keywords. In this context, correlation measures, such as Levenshtein distance or the calculation of cosine similarity, are adopted to explore the relationships between texts.

Levenshtein distance is a metric that calculates the difference between two sequences of characters, such as video titles and scientific publication titles. This measure allows us to assess the similarity or dissimilarity between texts, providing information about the thematic proximity between videos and publications.

Cosine similarity, in turn, is a measure that quantifies the similarity between two vectors of words, such as terms present in video comments and keywords in scientific publications, allowing us to identify semantic associations and relationships between texts.

By using these correlation measures, the Social4Science platform can reveal connections between the content of videos and scientific publications, identify thematic patterns, and explore how information is transmitted and discussed on social media. In this way, the combination of quantitative and textual data provides a comprehensive and enriching analysis, allowing us to understand both the quantitative and textual aspects involved in the dissemination and discussion of scientific publications on YouTube.

As an initial case study, a set containing 65,534 DOIs that at the time had citations of YouTube videos was collected from the Altmetric Platform in March 2022. From this set of DOIs, several characteristics of scientific publications were verified, considering the analysis of the type of publication, it was found that the majority were Articles (94.9%), followed in smaller quantities by Books (3%) and Book Chapters (1%). It is also worth mentioning a total of 45 Data Sets that were also referenced.

3 Results

Social analysis involves analyzing data from videos, such as the number of likes, views, and comments. On the other hand, bibliometric analysis includes quantitative data from articles, such as DOI validation, the number of citations received by other articles, and the number of videos that mention the article in question. From this data, it is possible to identify trends and patterns in discussions about scientific publications on social media. For example, we can determine which publications are most popular on these platforms, which topics generate the most discussions, and who the main influencers are in this context.

Through bibliometric analysis, and taking into account the date of data collection, it was possible to present in chronological order the publication period of the articles that were mentioned in the videos (Figure 2).

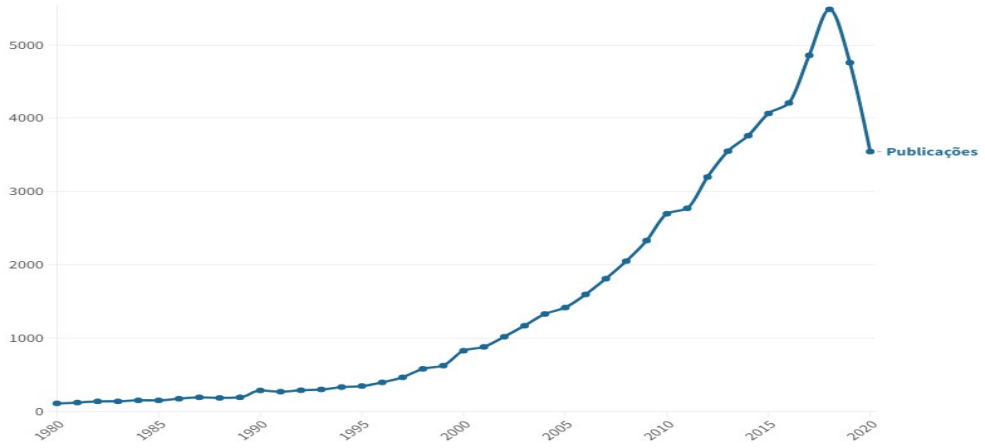

Figure 2: Publication period of the mentioned articles.

After analyzing the data, it was found that the oldest article identified was published in 1980. From this reference, it is possible to observe a significant growth in the number of scientific articles over time, with an even more pronounced trend starting in the year 2000. It was during this period that the use of DOI in scientific articles became more common.

It was also possible to determine the representativeness of the main journals in which the articles were published. This analysis aimed to quantify the articles published in each journal, highlighting those that were most mentioned in YouTube videos during the period analyzed (Figure 3).

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

It is possible to observe the representation of some prestigious journals, such as Nature, the American Journal of Clinical Nutrition, Plos One, Nutrients, the Journal of Strength & Conditioning Research and Science, among others. These journals are internationally recognized for their editorial quality and the scientific rigor of their publications.

In addition, it is interesting to note that certain areas of knowledge have a greater use of YouTube as a tool for disseminating scientific articles. This can be attributed to several factors, such as the nature of these areas, which can be more easily transmitted through videos. In addition, some areas may have a greater demand for direct and accessible communication, especially when it comes to topics of public interest (Figure 4).

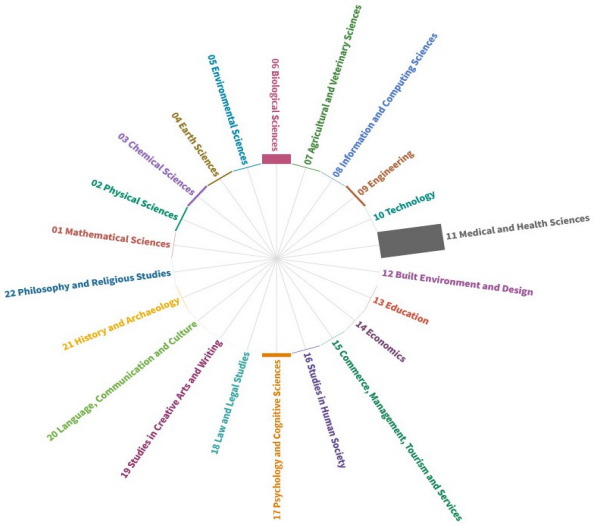

Figure 4: Predominant areas of the mentioned articles.

The use of YouTube as a scientific dissemination platform in these areas allows for more dynamic and interactive communication, providing a more engaging learning experience. Videos can include practical demonstrations, interviews with experts, debates and analyses of scientific articles, among other content that arouses the interest and curiosity of the public.

4 Conclusion

The Social4Science platform proposed in this work allows the collection and analysis of scientific data on social media, providing valuable results on the dissemination and dissemination of scientific content. Through the analysis of this data, it is possible to identify trends, patterns and knowledge gaps in discussions about scientific publications on social media. This tool offers valuable information to researchers and professionals in the scientific field, allowing adjustments in communication strategies and promotion of scientific knowledge, establishing a more effective connection with the general public.

All the tools developed with the source code of the framework modules are available in a command line interface and can be accessed through GitHub (https://github.com/RafaelGoncalo/social4scienceCLI) by the entire community of interest. A graphical interface for the tools has also been developed and is available through the Heroku platform (https://social4science-9d06f052d2f9.herokuapp.com/).

The upcoming features of the framework will be disclosed through the same means mentioned above and can be easily accessed and modified.

References

[1] NETO, José Ricardo Silva. Alcance da divulgação científica por meio do YouTube: estudo de caso no canal Meteoro Brasil. Múltiplos Olhares em Ciência da Informação, v. 8, n. 2, 2018.

[2] DA FONSECA, André Azevedo; BUENO, Leonardo Mendes. Breve panorama da divulgação científica brasileira no YouTube e nos podcasts. Cadernos De Comunicação, v. 25, n. 2, 2021

[3] REALE, Manuella Vieira; MARTYNIUK, Valdenise Leziér. Divulgação Científica no Youtube: a construção de sentido de pesquisadores nerds comunicando ciência. In: CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO. 2016. p. 1-15.

[4] ARAUJO, Ronaldo Ferreira. Communities of attention networks: introducing qualitative and conversational perspectives for altmetrics. Scientometrics, v.124, 1793-1809, 2020.