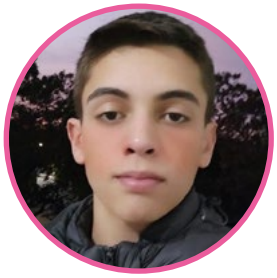
UM FRAMEWORK PARA COLETA E ANÁLISE DE DADOS SOCIAIS DE PUBLICAÇÕES CIENTÍFICAS

A Framework for Collecting and Analyzing Social Data from Scientific Publications
Un Estructura para recopilar y analizar datos sociales de publicaciones científicas

Thiago Magela Rodrigues Dias
Doutor em Modelagem Matemática e Computacional, Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte, MG, Brasil.
Professor, Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte, MG, Brasil.
Lattes: http://lattes.cnpq.br/4687858846001290
ORCID: https://orcid.org/0000-0001-5057-9936



Rafael Gonçalo Pereira Ribeiro
Bolsista de Iniciação Científica Jr., Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Divinópolis, MG, Brasil.
Aluno do Curso Técnico em Informática, Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Divinópolis, MG, Brasil.
Lattes: http://lattes.cnpq.br/6995985875605301

Patrícia Mascarenhas Dias
Doutora em Modelagem Matemática e Computacional, Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte, MG, Brasil.
Professor, Universidade do Estado de Minas Gerais (UEMG), Divinópolis, MG, Brasil.
Lattes: http://lattes.cnpq.br/6871965805554986
ORCID: https://orcid.org/0000-0002-8448-6874

Ronaldo Ferreira de Araujo
Doutor em Ciência da Informação, Universidade Federal de Minas Gerais (CEFET-MG), Belo Horizonte, MG, Brasil.
Professor, Universidade Federal de Alagoas (UFAL), Belo Horizonte, MG, Brasil.
Lattes: http://lattes.cnpq.br/3328212638040851
ORCID: https://orcid.org/0000-0003-0778-9561

Resumo

Introdução: Com o crescente uso das mídias sociais, tornou-se cada vez mais importante entender como as publicações científicas são disseminadas e discutidas nestas plataformas online. A análise desses dados de interação e circulação da pesquisa científica tem sido investigada nos estudos de altmetria e pode fornecer informações valiosas sobre como a ciência é percebida e compartilhada pelo público em geral. **Metodologia**: O trabalho tem como objetivo propor uma plataforma para a coleta e análise de dados sociais de publicações científicas. Através

da coleta de dados do YouTube, a plataforma busca entender como as publicações científicas são divulgadas e discutidas nas mídias sociais. **Resultados**: Com o framework proposto foi possível obter um ferramental que viabiliza a coleta de dados sociais, especificamente do Youtube e correlacionar com dados científicos de publicações, verificando diversas métricas de análises. **Conclusão**: A solução é bastante promissora e por meio de sua utilização foi possível identificar tendências e padrões nas discussões sobre as publicações científicas nas mídias sociais, e ainda viabilizar diversos outros estudos na temática.

Palavras-chave: produção científica; mídias sociais; Dados Abertos; altmetria; bibliometria.

Abstract

***Introduction**: With the increasing 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 can provide valuable insights into how science is perceived and shared by the general public. **Methodology**: This work aims to propose a platform for the collection and analysis of social data from scientific publications. By collecting data from YouTube, the platform seeks to understand how scientific publications are disseminated and discussed on social media. **Results**: As a result of this work, it was possible to obtain a tool that enables the collection of social data, specifically from Youtube and correlate with scientific data from publications, verifying various analysis metrics. **Conclusion**: With the analysis of the collected data, it is possible to identify trends and patterns in the discussions about scientific publications in social media, and also enable several other studies on the subject.*

***Keywords**: scientific production; social media; Open Data; altmetrics; bibliometrics.*

Resumen

***Introducción**: Con el uso creciente de las redes sociales, se ha vuelto cada vez más importante comprender cómo se difunden y discuten las publicaciones científicas en estas plataformas. El análisis de estos datos puede proporcionar información valiosa sobre cómo el público en general percibe y comparte la ciencia. **Metodología**: Este trabajo tiene como objetivo proponer una plataforma para la recolección y análisis de datos sociales de publicaciones científicas. Al recopilar datos de YouTube, la plataforma busca comprender cómo se difunden y discuten las publicaciones científicas en las redes sociales. **Resultados**: Como resultado de este trabajo se logró obtener una herramienta que permite la recolección de datos sociales, específicamente de Youtube y correlacionarlos con datos científicos de publicaciones, verificando diversas métricas de análisis. **Conclusión**: Con el análisis de los datos recopilados, es posible identificar tendencias y patrones en las discusiones sobre publicaciones científicas en las redes sociales, y también posibilitar varios otros estudios sobre el tema.*

***Palabras clave:** producción científica; redes sociales; información abierta; altmetría; bibliometría.*

DOI 10.22477/vi.widat.68

1. Introdução

O repasse das descobertas e processo de iniciação científica é extremamente importante para o desenvolvimento tanto social quanto cultural. A comunicação entre o meio acadêmico e a sociedade é crucial, uma vez que todo conhecimento e pesquisa desenvolvida tem como objetivo garantir um retorno à sociedade. Logo, a forma de repassar os resultados precisa estar alinhada à sociedade e às necessidades sociais para que o público perceba a sólida relação entre sociedade e ciência (NETO, 2018).

Neste sentido, o YouTube tem oferecido uma oportunidade interessante ao movimento de divulgação científica na Internet. A plataforma é o maior site de compartilhamento de vídeos do mundo, comportando conteúdos de diversos assuntos e temáticas. O Brasil é um dos principais consumidores dessa plataforma. Graças a isso, a divulgação científica encontrou um espaço de grande reverberação, com relativa demanda de público (FONSECA; BUENO, 2021).

Para os autores Reale e Martyniuk (2016), a divulgação científica por meio do YouTube é uma excelente ferramenta para democratizar o conhecimento científico.

Ao extrair informações de artigos científicos citados em vídeos do YouTube, é possível coletar uma série de dados, incluindo título do artigo científico, autores do artigo científico, nome do periódico em que o artigo foi publicado, ano de publicação do artigo, número de citações recebidas pelo artigo, dentre outros.

Com esses dados, é possível realizar diversas análises, incluindo a identificação de tendências mencionadas nos vídeos, a identificação de autores ou revistas de maior destaque mencionadas nos vídeos do YouTube, a análise da relação entre a popularidade dos vídeos do YouTube e o número de citações recebidas pelos artigos científicos mencionados nos vídeos, dentre outros.

Diante do exposto, este trabalho tem como objetivo a proposição de uma plataforma computacional para a coleta, tratamento e análise de dados científicos em mídias sociais. O trabalho assume o conceito de "mídias sociais", em detrimento de "redes sociais" por entender que o primeiro se refere às plataformas online que permitem a criação e compartilhamento de conteúdo, enquanto as "redes sociais" se refere aos relacionamentos e conexões estabelecidas entre pessoas ou entidades em uma plataforma específica, seja ela online ou offline.

O intuito do trabalho é coletar dados de mídias sociais como o Youtube e avaliar como as publicações científicas são disseminadas e discutidas nestas plataformas. Como resultado das etapas desenvolvidas se torna possível a análise sobre as características dos vídeos publicados no Youtube e que referenciam algum DOI (*Digital Object Identifier*).

Indicadores de atenção online têm sido debatidos no contexto de estudos altmétricos, que focam na compreensão do impacto social de resultados de pesquisas na web social (ARAÚJO, 2020). Essas análises podem ser úteis para pesquisadores, editores de periódicos e outros profissionais envolvidos na divulgação científica, pois

podem ajudar a compreender melhor como a ciência é percebida e compartilhada pelo público em geral e a identificar oportunidades para aumentar a visibilidade das publicações.

Estudos desenvolvidos com essas abordagens mais contextuais estão crescendo na literatura e sinalizam a preocupação no campo altmétrico em contribuir para o aprofundamento da análise e investigação de onde e como os artigos são utilizados por diversas comunidades que interagem com os artigos online (ARAÚJO, 2020).

2. Procedimentos Metodológicos

Para a realização deste trabalho, foi utilizada a busca no portal da Altmetric[1] por meio da plataforma *Altmetric Explorer* pelo conjunto de publicações científicas que possuem citações em vídeos publicados no YouTube. Tal relação é caracterizada quando algum vídeo cita algum artigo com uso do DOI, geralmente na descrição do vídeo. Logo, é possível vincular um vídeo, pelo seu identificador, a um artigo, pelo seu DOI.

A partir desta relação extraída da Altmetric, contendo um arquivo com Identificador do vídeo e o DOI de uma publicação, todo o processo de extração dos dados é iniciado. A plataforma recebe esta relação como entrada e dá início a todo o processo de coleta e análise dos dados, dividido em dois segmentos:

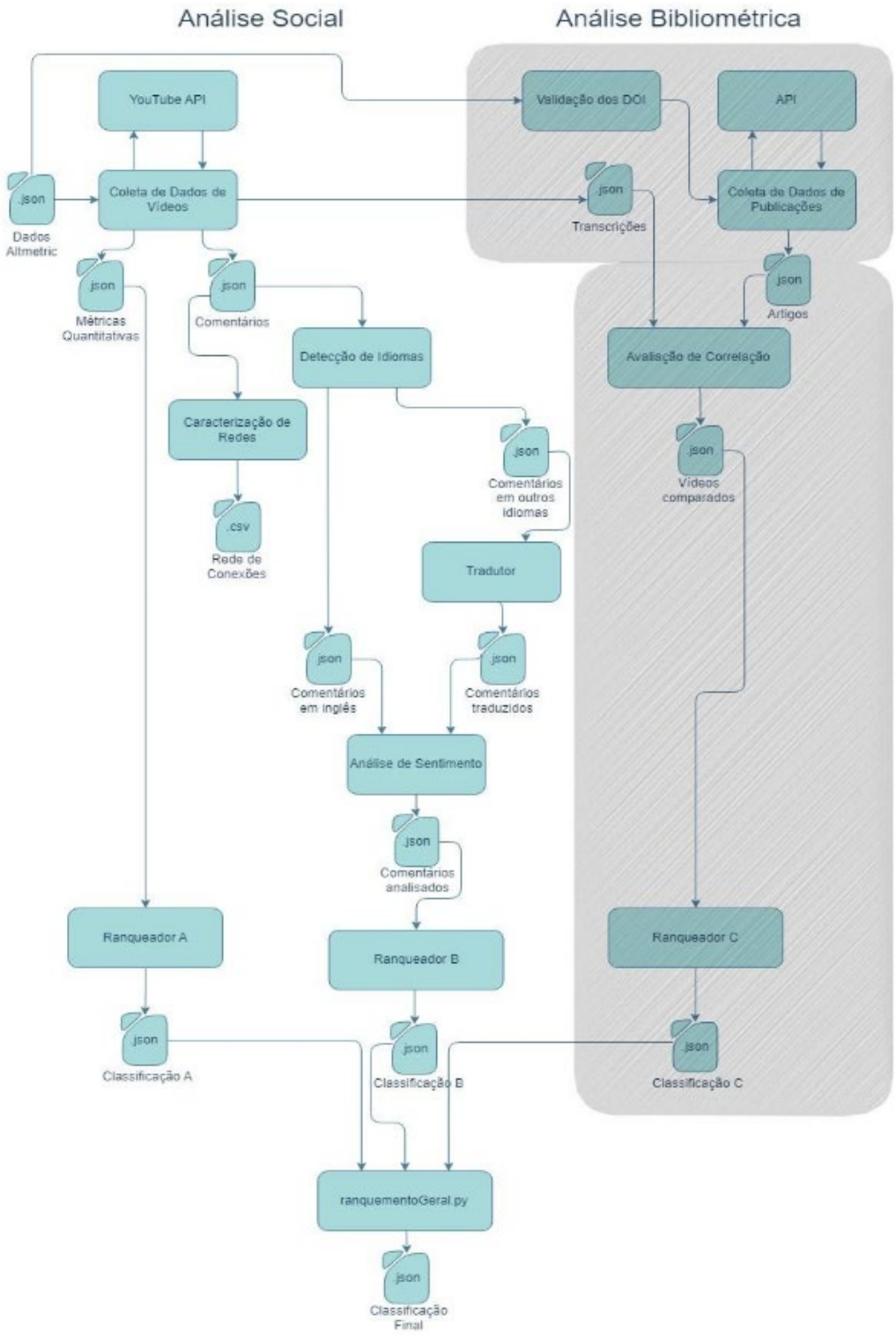
1. Análise Social: coleta de dados dos vídeos do YouTube.

2. Análise Bibliométrica: coleta de dados dos artigos científicos.

A arquitetura da Plataforma Proposta pode ser observada na Figura 1.

1 https://www.altmetric.com/

Figura 1 – Arquitetura geral do Framework desenvolvido.



Fonte: Os autores (2023)

Como pode ser observado, todo o processo de coleta e tratamento dos dados a ser realizado pela plataforma se inicia com a entrada do arquivo da Altmetric contendo o Identificadores dos vídeos e os DOI dos artigos. Logo, os DOI são utilizados para a etapa de "Análise Bibliométrica" e os Identificadores dos vídeos são utilizados para a etapa de "Análise Social".

Na Análise Social (1) os dados dos vídeos são coletados via uma *Application Programming Interface (*API) pública do Youtube, quando são gerados alguns extratos de dados. Eles são utilizados para o cálculo de diversas métricas e podem ser exportados para outras ferramentas de análises e visualizações, permitindo outras formas de aprofundamento.

Como exemplo destes extratos, destacam-se os conjuntos contendo dados quantitativos dos vídeos, como por exemplo, quantidade de visualizações, quantidade de comentários e curtidas de cada vídeo, bem como, extratos contendo dados dos canais em que os vídeos foram publicados, das redes de interações identificadas a partir dos comentários de cada vídeo, extratos das descrições dos vídeos, das transcrições de cada áudio e, por fim, um conjunto de dados padronizados no idioma inglês de todos os comentários extraídos.

Já na Análise Bibliométrica (2), o conjunto dos DOI é verificado via API, a fim de validá-los. Caso seja um DOI válido, seus dados são enviados para a API da OpenAlex[2], recuperando, desta forma, informações sobre o artigo em questão, como por exemplo, seu título, autores, ano de publicação, resumo, palavras-chave, periódico de publicação, dentre outros. Além disso, visando complementar os dados uma nova requisição do mesmo DOI é enviada para a API da OpenCitations[3], recuperando as citações do artigo.

Todo este conjunto de dados são armazenados em extratos de dados que também são objeto de análises por diversas métricas implícitas na própria plataforma e são disponibilizados em formatos que possam ser importados por outras ferramentas de análise e visualização.

Os dados quantitativos podem ser utilizados para diversos tipos de ranqueamento e também para correlações entre as Análises Sociais e Análises Bibliométricas. Já os conjuntos contendo informações textuais dos vídeos, como título, comentários, descrição e transcrição, são correlacionados com os dados textuais das publicações, como títulos, resumos e palavras-chave, sendo adotado para estas análises algumas medidas de correlação como a distância de Levenshtein ou à similaridade do cosseno.

Como um estudo de caso inicial foram coletados da Plataforma Altmetric em março de 2022 um conjunto contendo 65.534 DOI´s que possuíam na época citações de vídeos do Youtube. Deste conjunto de DOI´s foram verificadas diversas características das publicações científicas, considerando a análise do tipo de publicação, verificou-se que a maioria era de Artigos (94,9%), seguido em menor quantidade por Livros (3%) e Capítulos de Livros (1%). Ressalta-se ainda um total de 45 Conjuntos de Dados que também eram referenciados.

2 https://openalex.org/

3 https://opencitations.net/

3. Resultados

Na Análise Social todos os dados dos vídeos são analisados como quantidade de curtidas, quantidade de visualizações e quantidade de comentários. Por outro lado, na análise bibliométrica, todos os dados quantitativos dos artigos são analisados, como a validação do DOI, recuperação da quantidade de citações recebidas por outros artigos e a quantidade de vídeos que citam o referido artigo. A partir desses dados, é possível identificar tendências e padrões nas discussões sobre as publicações científicas nas mídias sociais. Por exemplo, pode-se identificar quais publicações são mais populares nas mídias sociais, quais assuntos geram mais discussões e quais são os principais influenciadores.

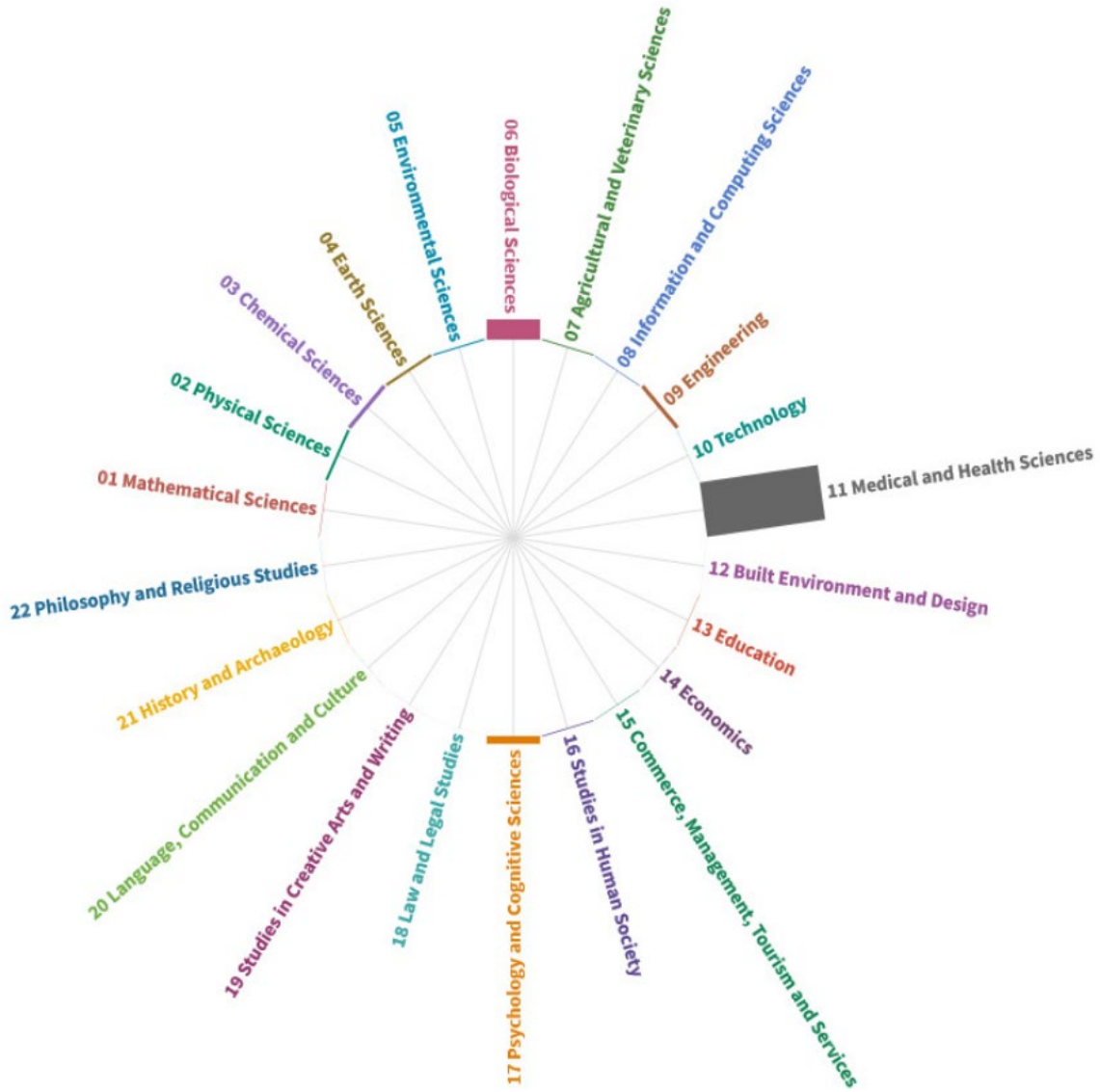
Considerando ainda as análises bibliométricas, foi possível verificar a representatividade dos principais periódicos em que os artigos haviam sido publicados (Figura 2).

Figura 2 – Representatividade dos periódicos dos artigos referenciados nos vídeos.

Fonte: Os autores (2023)

É possível observar a representatividade de alguns periódicos de prestígio como Nature, American Journal of Clinical Nutrition, Plos One, Nutrients, Journal of Strength & Conditioning Research e Science. Outra observação está relacionada às áreas do conhecimento dos artigos analisados e que tendem a ter uma maior utilização do Youtube como ferramenta de divulgação de artigos (Figura 3).

Figura 3 – Principais áreas dos artigos referenciados.



Fonte: Os autores (2023)

Utilizando a classificação de áreas do conhecimento dos artigos, conforme dados coletados, percebe-se uma grande concentração em duas áreas principais, Ciências Médicas e da Saúde (69%) e Ciências Biológicas (11,5%), sendo apenas estas duas áreas detentoras de aproximadamente 80% de todo o conjunto.

Diversas outras análises bibliométricas também foram realizadas, como por exemplo a validação dos DOI que são citados, objetivando verificar a real existência da publicação e, após esta etapa, a coleta e análise dos dados contidos nos títulos, resumos, palavras-chave, bem como, informações do número de citações desses artigos por outras publicações. E ainda, dados dos periódicos em que foram publicados, como fator de impacto e Qualis. Todos estes dados são automaticamente coletados por API públicas de diversas fontes e de forma automática.

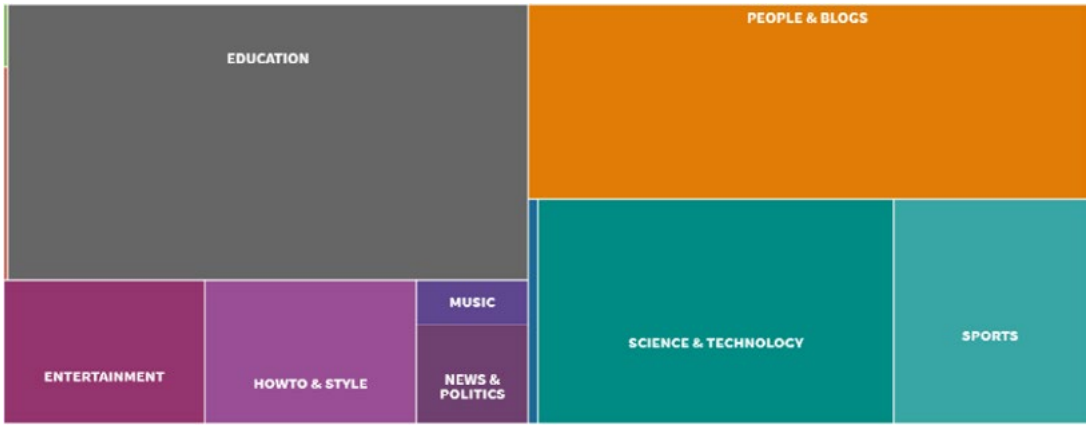
Já as Análises Sociais levam em conta informações dos vídeos que são publicados no Youtube e que fazem referência a algum DOI. Logo, as análises sociais se concentram em informações que são obtidas diretamente do Youtube, utilizando-se para isso a sua API pública de requisição dos dados. Todo este processo de extração

dos dados do Youtube também ocorre de forma automática a partir da listagem inicial informada. A partir dos dados de entrada os identificadores dos vídeos são extraídos e as requisições são realizadas na API do Youtube.

Foi possível observar nos dados coletados como os vídeos estão classificados em categorias. Tais categorias são referentes aos canais em que estes vídeos são publicados (Figura 4).

Percebe-se que os canais que possuem vídeos que citam publicações com DOI estão em sua maioria categorizados como Educação, Pessoas e Blogs, e ainda, Ciência e Tecnologia. Para melhor compreensão desta categorização e seu impacto, é importante analisar qual a representatividade destes canais, uma vez que a quantidade de inscritos no canal, a quantidade de vídeos publicados ou a data de registro do canal no Youtube podem agregar diversas novas informações.

Figura 4 – Categorias dos canais dos vídeos.



Fonte: Os autores (2023)

Considerando ainda os dados da Análise Social, são caracterizados pela plataforma desenvolvida redes de comentários. Os comentários de cada um dos vídeos são analisados e as ligações entre os canais são identificadas. Logo, a partir de um conjunto de comentários, é possível identificar com análise de redes como os canais interagem, tendo em vista os comentários que realizam ou recebem.

Além das redes caracterizadas, diversas outras análises quantitativas são realizadas, como por exemplo a quantidade de visualizações que um determinado vídeo recebe, a quantidade de comentários, a quantidade de curtidas, a duração e idioma dos vídeos, dentre outros dados.

Considerando ainda a Análise Social, o conteúdo dos vídeos também são objeto de análises, como por exemplo o título dos vídeos, a descrição dos vídeos, bem como a transcrição do áudio. Tais elementos são importantes pois viabilizam diversas outras análises que visam correlacionar os conteúdos dos vídeos com dados do conteúdo dos artigos, que também são coletados, como o título, resumo e palavras-chave.

4. Considerações Finais

A plataforma proposta neste trabalho permite a coleta e análise de dados científicos em mídias sociais, possibilitando uma melhor compreensão da divulgação e propagação do conteúdo científico nas mídias sociais. Com a análise dos dados coletados, é possível identificar tendências e padrões nas discussões sobre as publicações científicas nas mídias sociais, o que pode ser útil para pesquisadores, editores de periódicos e outros profissionais envolvidos.

Com os dados coletados pela plataforma proposta, é possível realizar diversas correlações entre os dados. Essas correlações podem ajudar a compreender melhor como a popularidade de um vídeo no YouTube está relacionada às características do artigo científico que ele menciona e como isso pode variar de acordo com a área, o tipo de publicação ou o país de origem.

Todo o ferramental desenvolvido com o código fonte de todos os módulos do *framework* será disponibilizado em um repositório do GitHub para toda comunidade de interesse.

Agradecimentos

Os autores expressam agradecimento à Altmetric.com por fornecer os dados altmétricos deste estudo gratuitamente para fins de pesquisa.

Referências

ARAUJO, Ronaldo Ferreira. Communities of attention networks: introducing qualitative and conversational perspectives for altmetrics. **Scientometrics**, v.124, 1793-1809, 2020. https://doi.org/10.1007/s11192-020-03566-7

FONSECA, André Azevedo da.; BUENO, Leonardo Mendes. Breve panorama da divulgação científica brasileira no YouTube e nos podcasts. **Cadernos De Comunicação**, v. 25, n. 2, 2021. https://scholar.archive.org/work/kci5gaqbyvbxhonfsiyjvf3fai/access/wayback/https://periodicos.ufsm.br/ccomunicacao/article/download/63121/pdf_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. Disponível em: https://periodicos.ufmg.br/index.php/moci/article/download/16885/13644.

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, 39., 2016, São Paulo. **Anais eletrônicos [...]**. São Paulo: Sociedade Brasileira de Estudos Interdisciplinaes da Comunicação, 2016, p. 1-15. Disponível em: https://www.portalintercom.org.br/anais/nacional2016/resumos/R11-0897-1.pdf.