



IDENTIFICAÇÃO DE ÁREAS DO CONHECIMENTO DO CNPQ USANDO O GPT

Jesús P. Mena-Chalco

<https://orcid.org/0000-0001-7509-5532>.

jesus.mena@ufabc.edu.br.

Universidade Federal do ABC (UFABC) |

<https://ror.org/028kg9j04> | Santo André, Brasil.

Luís Paulo Faina Garcia

<https://orcid.org/0000-0003-0679-9143>.

luis.garcia@unb.br.

Universidade de Brasília (UnB) |

<https://ror.org/02xfp8v59> | Brasília, Brasil.

Rafaela Lora Grando

<https://orcid.org/0000-0003-4239-9622>.

rafaela.grando@fiocruz.br.

Fundação Oswaldo Cruz (Fiocruz) |

<https://ror.org/04jhsww08> | Rio de Janeiro, Brasil.

Waldeyr Mendes Cordeiro da Silva

<https://orcid.org/0000-0002-8660-6331>.

waldeyr.mendes@fiocruz.br.

Fundação Oswaldo Cruz (Fiocruz) |

<https://ror.org/04jhsww08> | Rio de Janeiro, Brasil.

Rodrigo Gomes Ferrari Cesar

<https://orcid.org/0000-0001-8470-2728>.

rodrigo.ferrari@fiocruz.br.

Fundação Oswaldo Cruz (Fiocruz) |

<https://ror.org/04jhsww08> | Rio de Janeiro, Brasil.

Vanessa de Arruda Jorge

<https://orcid.org/0000-0002-5298-9311>.

vanessa.jorge@fiocruz.br.

Fundação Oswaldo Cruz (Fiocruz) |

<https://ror.org/04jhsww08> | Rio de Janeiro, Brasil.

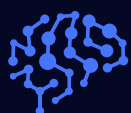
Eixo temático: Técnicas, Ferramentas e Infraestruturas

Modalidade: Resumo expandido

DOI: 10.22477/ix.ebbc.356

Resumo: A compreensão da ciência é facilitada pela classificação em áreas do conhecimento. Ter um panorama da produção científica é importante para otimizar recursos, identificar colaborações e promover a interdisciplinaridade. Este artigo mostra a utilização do *Generative Pre-trained Transformer* (GPT) para classificar artigos científicos. Foram considerados como estudo de caso 8.753 artigos produzidos por pesquisadores da Fundação Oswaldo Cruz, e as áreas do conhecimento do CNPq. Os resultados mostram desempenho comparável ou superior à avaliação humana. A maior acurácia foi alcançada na área de Ciências da Saúde (85,35%), seguida por Ciências Biológicas (63,19%) e Ciências Sociais Aplicadas (55,26%).

Palavras-Chave: Bibliometria. Processamento de Linguagem Natural. Classificação. GPT.



1 INTRODUÇÃO

A ciência, em sua vastidão e complexidade, encontra na classificação por áreas e subáreas do conhecimento um instrumento fundamental para sua organização. Compreender a produção científica de um ator acadêmico em sua totalidade é crucial para diversos fins, desde a caracterização de sua trajetória até a tomada de decisões estratégicas. Instituições públicas com viés acadêmico, como é o caso da Fundação Oswaldo Cruz (Fiocruz), utilizam a compreensão da produção científica em áreas e linhas de pesquisa também como um componente de prestação de contas perante a sociedade, pois permitem o acompanhamento do seu alinhamento científico com as estratégias do Sistema Único de Saúde (SUS). Essa categorização, que delinea fronteiras conceituais e metodológicas, facilita o mapeamento da ciência (Leydesdorff, 1987), tornando-o mais acessível a cientistas, educadores, gestores de recursos e demais interessados.

Identificar as áreas do conhecimento (Mendoza *et al.* 2022) nas quais investe seus esforços revela-se fundamental para uma série de objetivos, como: (i) otimizar recursos e financiamento; (ii) identificar colaborações e redes de pesquisa; (iii) avaliar a produção científica; (iv) promover a interdisciplinaridade; (v) promover a gestão do conhecimento e a recuperação da informação e (vi) facilitar a compreensão do impacto científico.

O uso de técnicas Inteligência Artificial, mais especificamente IA Generativa como GPT, oferece uma abordagem inovadora e eficaz para lidar com grandes volumes de dados da produção científica. A capacidade do GPT de compreender e gerar texto de forma contextual e semântica possibilita uma classificação mais precisa e detalhada, considerando não apenas palavras-chave, mas também o contexto e a relevância das publicações. Além disso, sua capacidade de aprender com conjuntos de dados vastos e diversificados torna-o altamente adaptável a diferentes campos e disciplinas. Ao utilizar o GPT para a classificação de produção científica, os pesquisadores podem economizar tempo e recursos e ainda explorar novos olhares sobre a temática pesquisada.

Este artigo explora o desafio da classificação de artigos científicos em suas respectivas áreas de conhecimento. Para tanto, o foco da pesquisa reside na aplicação do GPT (Brown *et al.* 2020), como Modelo de Linguagem para a tarefa de classificação das áreas de conhecimento de uma amostra de artigos, com autoria e/ou coautoria de pesquisadores da Fiocruz, de acordo com a hierarquia proposta pelo CNPq.

2 O USO DO GPT PARA TAREFAS DE CLASSIFICAÇÃO

Modelos de Linguagem Grandes (do inglês: Large Language Models ou LLMs), como o GPT, representam um avanço significativo no campo da IA, com potencial para revolucionar a maneira como abordamos diversas tarefas, inclusive no âmbito acadêmico (Rahaman *et al.*, 2023). Essas ferramentas têm se mostrado promissoras para tarefas de classificação de texto devido à sua capacidade de Processamento de Linguagem Natural. Os modelos de linguagem são treinados com grande volume de dados textuais, o que lhes permite capturar padrões complexos e nuances na linguagem humana (Alizadeh *et al.*, 2022; Huang, Kwak; An, 2023). Sua capacidade de processar uma ampla gama de informações

textuais torna-o uma ferramenta versátil para diferentes tipos de análises de texto. Isso pode ser especialmente útil em cenários onde a classificação manual de dados seria demorada e custosa (Gilardi *et al.*, 2023).

A classificação temática de publicações científicas em áreas de conhecimento é um problema complexo na cientometria e bibliometria (Hjørland, 2002), especialmente em face de dados incompletos e da crescente interdisciplinaridade da pesquisa (Leydesdorff; Rafolds, 2011). Este desafio se intensifica quando as publicações se situam em áreas fronteiriças (Wagner *et al.*, 2011), desafiando os métodos tradicionais de classificação.

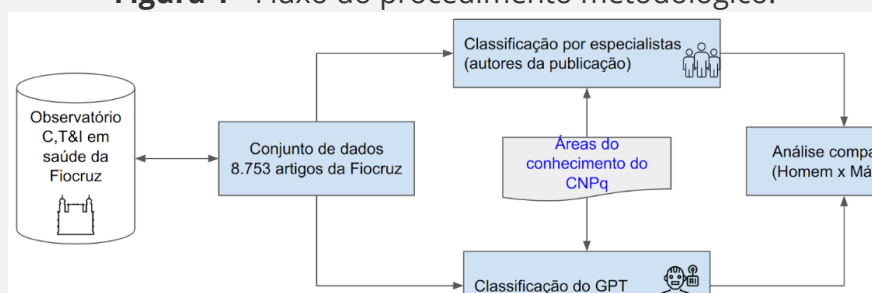
A classificação de publicações científicas, em áreas/tópicos/termos/temas/assuntos, é uma demanda de longa data da comunidade científica, com diversas soluções propostas ao longo dos anos. Essa necessidade se justifica principalmente pela importância da categorização para a recuperação eficiente de informações em bases de dados científicas (Carage; Bulgarov; Mihalcea, 2015). De acordo com, Mendoza *et al.* (2022) algumas características justificam a escolha de uma abordagem computacional em substituição aos métodos manuais sendo as principais: menor tempo de análise; equanimidade e consistência no tratamento dos dados; assim como, a transparência e responsabilidade.

3 PROCEDIMENTO METODOLÓGICO

Este trabalho propõe a aplicação do GPT como classificador para identificação de áreas do conhecimento do CNPq em publicações do tipo artigo científico. Este processo foi feito através de uma comparação da classificação das áreas do conhecimento categorizadas manualmente por especialistas humanos em contraposição a uma classificação de máquina feita pelo GPT. A investigação se baseia em um conjunto de publicações da Fiocruz selecionadas do Observatório em Ciência, Tecnologia & Inovação em Saúde (Observatório CT&I em Saúde) da Fiocruz (Observatório, 2024) e utiliza o sistema de classificação em áreas do conhecimento propostas pelo CNPq (CNPq, 2024) como referência. A metodologia empregada neste estudo é detalhada e organizada visualmente na Figura 1.

Para classificar as publicações, utilizamos o sistema de classificação do CNPq, que é referência nacional na avaliação da produção científica. O sistema possui quatro níveis hierárquicos que permitem uma análise detalhada da área de conhecimento de cada publicação. Esses níveis, neste trabalho, são entendidos como: Grande área, Área, Subárea e Especialidade (CNPq, 2024). Foram consideradas 8.753 publicações científicas como amostra para este estudo de caso.

Figura 1 - Fluxo do procedimento metodológico.



Fonte: Elaborado pelos autores (2024).

A amostra de publicações selecionadas representa aproximadamente 20% do total de artigos publicados pela instituição no período (40.050 artigos) e conta com categorização manual, realizada pelas Unidades da Fiocruz. Este processo demandou uma ampla organização de tal forma que todas as classificações tenham o mesmo padrão e formato.

No entanto, é importante destacar que algumas dessas amostras apresentavam hierarquias incompletas, com ausência de categorizações em níveis específicos. A distribuição dos valores ausentes revela que, dentre as amostras analisadas por especialista, 21 não possuíam classificação de área, 576 não tinham categorização de subárea e a maioria significativa, totalizando 6.416 publicações, não continha classificação de especialidade.

O GPT utilizado foi a versão GPT-3.5-Turbo. Ele apresenta uma melhor relação custo-desempenho comparado com a versão GPT-4.0 ou GPT-4.0-Turbo. Além disso, ele foi empregado de forma hierárquica, utilizando o Classificador Local por Nó Pai - uma estratégia padrão para classificação hierárquica. Para implementar essa estratégia, foram desenvolvidas strings dinâmicas que consideravam somente os possíveis rótulos do próximo nível hierárquico. Isso significa que, cada string foi ajustada, de acordo com a categoria atribuída ao nó pai na hierarquia. As classificações foram realizadas de maneira sistemática até o último nível da hierarquia, garantindo uma análise completa e detalhada de todas as amostras. A Figura 2 apresenta um modelo da consulta realizada.

Por fim, a avaliação do desempenho do modelo foi realizada utilizando a acurácia como métrica. Essa métrica foi calculada de forma global, ou seja, para uma classificação ser 100% efetiva é necessária a mesma resposta do especialista e do modelo do GPT. Foi tomado como padrão de referência a resposta do especialista, esta foi comparada a resposta do GPT, analisando a convergência ou a divergência da categorização de cada nível. Além disso, para uma análise mais detalhada, o desempenho do modelo foi avaliado por grande área, proporcionando reflexões sobre sua eficácia em diferentes contextos e categorias de classificação.

Figura 2 - Exemplo de consulta ao GPT, via API, usada para identificar a Grande área do conhecimento de uma publicação.

CNPq areas: 'Ciências da Saúde, Ciências Biológicas, Ciências Exatas e da Terra, Ciências Agrárias, Ciências Humanas, Engenharias, Ciências Sociais Aplicadas'
Instructions: Classify the title, abstract, keywords and journal into one of the above CNPq areas.
Title: '...'
Abstract: '...'
Keywords: '...'
Journal: '...'

Fonte: Elaborado pelos autores (2024).

4 RESULTADOS E ANÁLISES

Nesta seção, são expostos os resultados referentes à acurácia (ou precisão) do procedimento adotado. Uma classificação feita pelo computador é considerada correta quando coincide com a classificação realizada por um humano.

A Figura 3(A) apresenta os resultados de desempenho do GPT em classificar os diferentes níveis

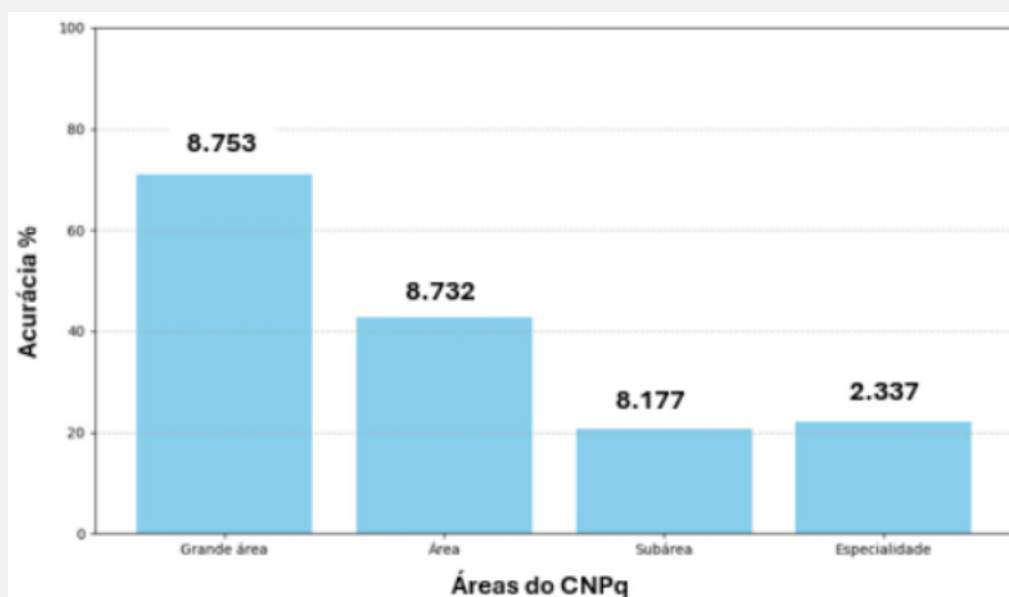
do CNPq: grande área, área, subárea e especialidade. O eixo x apresenta os níveis do CNPq e o eixo y, a acurácia alcançada pelo GPT. Além disso, cada barra apresenta em seu topo o valor do suporte - número total de amostras. Os resultados mostram uma alta acurácia (71,05%) na classificação das grandes áreas, porém, houve uma queda significativa nos níveis mais específicos, com acurácias de 42,72% para área, 20,63% para subárea e 22,17% para especialidade. Essa variação sugere que o modelo enfrentou desafios ao classificar amostras em níveis hierárquicos mais detalhados.

A Figura 3(B) contempla os resultados de desempenho do GPT em classificar as diferentes grandes áreas: Ciências da Saúde, Ciências Biológicas, Ciências Sociais Aplicadas, Ciências Exatas e da Terra, Ciências Agrárias, Ciências Humanas e Engenharias. O eixo x apresenta a acurácia e o eixo y as diferentes áreas do CNPq. Além disso, cada barra apresenta a sua frente o valor do suporte. Por suporte, entende-se a quantidade de amostras fornecidas para categorização. O maior desempenho foi alcançado na área de Ciências da Saúde, com uma acurácia de (85,35%), seguida por Ciências Biológicas (63,19%) e Ciências Sociais Aplicadas (55,26%). No entanto, observou-se uma queda acentuada na acurácia em áreas como Ciências Agrárias (7,74%), Ciências Humanas (3,91%) e Engenharias (2,56%). Essa variação pode ser atribuída não apenas à capacidade do modelo de classificação, mas também pela pouca quantidade de amostras disponíveis. Verificou-se também uma maior confusão entre classes como Ciências Agrárias e Ciências Biológicas, Ciências Sociais Aplicadas e Ciências da Saúde e Ciências Sociais Aplicadas e Engenharias.

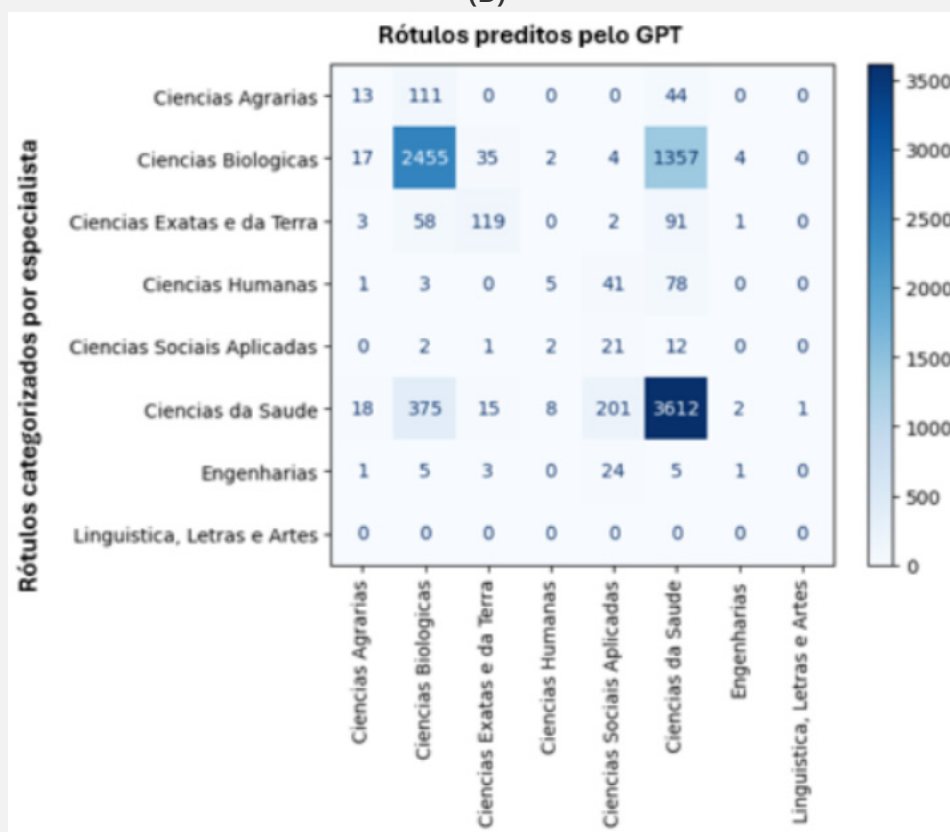
Os resultados apresentados indicam que o GPT, apesar de uma taxa de erro elevada para níveis como subárea e especialidade, tem capacidade de ser utilizado como um classificador de grande área (aproximadamente 71%) do CNPq. No nível de grande área, percebe-se que rótulos como Ciências Biológicas e Ciência das Saúde são os mais frequentes (com mais amostras rotuladas pelo especialista) e com melhor desempenho. Em nível de área, rótulos como Medicina, Saúde Coletiva e Parasitologia apresentam maior quantidade de amostras, e também melhor desempenho.

Figura 3 - Resultados da classificação usando o GPT: (A) Acurácia por nível. (B) Matriz de classificação.

(A)



(B)



Fonte: Elaborado pelos autores (2024).

Em casos onde apresenta-se maior dificuldade de classificação, outras estratégias de classificação hierárquica, fazendo a identificação do nível anterior por inferência através da hierarquia do CNPq, ou aumento de amostras com as informações das publicações pode ser utilizado para a classificação, dando assim mais suporte ao GPT para classificar corretamente.

Ainda houveram alguns casos onde a máquina apresentou uma precisão maior na identificação do rótulo da publicação. A avaliação da maior acurácia foi conduzida examinando amostras onde a classificação do computador divergiu da classificação humana (ver Figura 4). Embora a avaliação seja subjetiva, ela fornece evidências de que, em muitos casos, a classificação feita pelo GPT é mais precisa do que a realizada pelo humano.

Figura 4 - Artigo categorizado com mais precisão pela máquina.

Título: The genome of the zoonotic malaria parasite *Plasmodium simium* reveals adaptations to host switching.

Resumo: Background: *Plasmodium simium*, a malaria parasite of non-human primates (NHP), was recently shown to cause zoonotic infections in humans in Brazil. We sequenced the *P. simium* genome to investigate its evolutionary history and to identify any genetic adaptations that may underlie the ability of this parasite to switch between host species. Results: Phylogenetic analyses based on whole genome sequences of *P. simium* from humans and NHPs reveals that *P. simium* is monophyletic within the broader diversity of South American *Plasmodium vivax*, suggesting *P. simium* first infected NHPs as a result of a host switch of *P. vivax* from humans. The *P. simium* isolates show the closest relationship to Mexican *P. vivax* isolates. Analysis of erythrocyte invasion genes reveals differences between *P. vivax* and *P. simium*, including large deletions in the Duffy-binding protein 1 (DBP1) and reticulocyte-binding protein 2a genes of *P. simium*. Analysis of *P. simium* isolated from NHPs and humans revealed a deletion of 38 amino acids in DBP1 present in all human-derived isolates, whereas NHP isolates were multi-allelic. Conclusions: Analysis of the *P. simium* genome confirmed a close phylogenetic relationship between *P. simium* and *P. vivax*, and suggests a very recent American origin for *P. simium*. The presence of the DBP1 deletion in all human-derived isolates tested suggests that this deletion, in combination with other genetic changes in *P. simium*, may facilitate the invasion of human red blood cells and may explain, at least in part, the basis of the recent zoonotic infections.

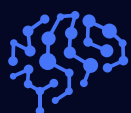
Palavras-chave: Comparative genomics, Malaria, *Plasmodium simium* *Plasmodium vivax*, Zoonosis.

Journal: BMC Biology.

Classificação original: Ciências da Saúde, Medicina, Clínica Médica, Doenças Infecciosas e Parasitárias.

Classificação GPT: Ciências Biológicas, Genética, Genética Molecular e de Microorganismos.

Fonte: Elaborado pelos autores (2024).

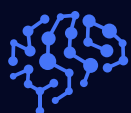


5 CONCLUSÕES

Os experimentos realizados evidenciam que o uso do GPT para a classificação de publicações em diferentes áreas do conhecimento pode aprimorar a precisão da classificação textual. Observou-se que essa ferramenta é eficaz para analisar padrões e sutilezas linguísticas, sendo ainda capaz de discernir nuances e detalhes que podem passar despercebidos na classificação manual por especialistas. O treinamento em bases com grande volume de informação permite que o GPT compreenda jargões técnicos e realize uma classificação comparável à realizada por humanos (Chiang; Lee, 2023). O GPT abre para novas possibilidades na área de bibliometria, como a automatização da classificação de grandes conjuntos de dados e a análise de padrões bibliométricos em diferentes áreas do conhecimento.

REFERÊNCIAS

- ALIZADEH, Meysam *et al.* Content moderation as a political issue: the twitter discourse around Trump's ban. **Journal of Quantitative Description: Digital Media**, v. 2, 4 out. 2022. DOI: <https://doi.org/10.51685/jqd.2022.023>. Disponível em: <https://journalqd.org/article/view/3424> Acesso em: 6 mar. 2024.
- BROWN, Tom B. *et al.* Language models are few-shot learners. **Advances in neural information processing systems**, v. 33, p. 1877-1901, 2020. DOI: <https://doi.org/10.48550/arXiv.2005.14165>. Disponível em: <https://arxiv.org/abs/2005.14165>. Acesso em: 8 mai. 2024.
- CARAGEA, Cornelia; BULGAROV, Florin; MIHALCEA, Rada. Co-training for topic classification of scholarly data. In: Proceedings of the 2015 Conference on empirical methods in natural language processing, Lisbon, Portugal. **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg**, PA, USA: Association for Computational Linguistics, 2015. DOI: <https://doi.org/10.18653/v1/d15-1283>. Disponível em: <https://aclanthology.org/D15-1283>. Acesso em: 6 mar. 2024.
- CHIANG, Cheng-Han.; LEE, Hung-yi. Can large language models be an alternative to human evaluations? **ArXiv preprint**, arXiv:2305.01937, 2023. DOI: <https://doi.org/10.48550/arXiv.2305.01937>. Disponível em: <https://arxiv.org/abs/2305.01937>. Acesso em: 8 mai. 2024.
- CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO. **Áreas de conhecimento do CNPq**. 2024. Disponível em: <https://lattes.cnpq.br/documents/11871/24930/TabeladeAreasdoConhecimento.pdf>. Acesso em: 6 mar. 2024.
- GILARDI, Fabrizio; ALIZADEH, Meysam; KUBLI, Maël. ChatGPT outperforms crowd workers for text-annotation tasks. **Proceedings of the national academy of sciences**, v. 120, n. 30, 18 jul. 2023. DOI: <https://doi.org/10.1073/pnas.2305016120>. Disponível em: <https://www.pnas.org/doi/10.1073/pnas.2305016120>. Acesso em: 6 mar. 2024.
- HJORLAND, Birger. Domain analysis in information science. **Journal of documentation**, v. 58, n. 4, p. 422-462, ago. 2002. DOI: <https://doi.org/10.1108/00220410210431136>. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/00220410210431136/full/html>. Acesso em: 6 mar. 2024.



HUANG, Fan; KWAK, Haewoon; AN, Jisun. Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. **ArXiv preprint**, arXiv:2302.07736, 2023. Disponível em: <https://arxiv.org/abs/2302.07736>. Acesso em: 8 mai. 2024.

LEYDESDORFF, Loet. Various methods for the mapping of science. **Scientometrics**, v. 11, n. 5-6, p. 295-324, maio 1987. DOI: <https://doi.org/10.1007/bf02279351>. Disponível em: <https://link.springer.com/article/10.1007/BF02279351>. Acesso em: 6 mar. 2024.

LEYDESDORFF, Loet.; Rafols, I. Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. **Journal of informetrics**, v. 5, n. 1, p. 87-100, jan. 2011. DOI: <https://doi.org/10.1016/j.joi.2010.09.002>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1751157710000854?via%3Dihub>. Acesso em: 6 mar. 2024.

MENDOZA, Óscar E. *et al.* Benchmark for research theme classification of scholarly documents. In: WORKSHOP ON SCHOLARLY DOCUMENT PROCESSING, 3., 2022. **Anais [...]**. Gyeongju: Association for computational linguistics, 2022. p. 253-262. Disponível em: <https://aclanthology.org/2022.sdp-1.31>. Acesso em: 8 mai. 2024.

OBSERVATÓRIO DA FIOCRUZ EM C,T&I EM SAÚDE, 2024. Disponível em: <https://observatorio.fiocruz.br>. Acesso em: 6 mar. 2024.

RAHMAN, Mizanur *et al.* ChatGPT and academic research: a review and recommendations based on practical examples. **Journal of education, management and development studies**, v. 3, n. 1, p. 1-12, 27 mar. 2023. DOI: <https://doi.org/10.52631/jemds.v3i1.175>. Disponível em: <https://journals.cspc.edu.ph/index.php/jemds/article/view/175>. Acesso em: 6 mar. 2024.

WAGNER, Caroline S. *et al.* Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. **Journal of informetrics**, v. 5, n. 1, p. 14-26, jan. 2011. DOI: <https://doi.org/10.1016/j.joi.2010.06.004>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1751157710000581?via%3Dihub>. Acesso em: 6 mar. 2024.