em questão
E-ISSN 1808-5245

Primeiras impressões na ciência: prevendo o impacto de artigos a partir de títulos e metadados básicos

Yonara Costa Magalhães [I]
https://orcid.org/0000-0001-5502-9634

João Pedro Cavalcanti Azevedo [II]
https://orcid.org/0009-0003-2035-6872

Antônio de Abreu Batista Junior [II]
https://orcid.org/0000-0002-6013-9704

Jesús Pascual Mena-Chalco [III]
https://orcid.org/0000-0001-7509-5532

[I] Universidade Ceuma, São Luís, MA, Brasil
[II] Universidade Federal do Maranhão, São Luís, MA, Brasil
[III] Universidade Federal do ABC, Santo André, SP, Brasil

Resumo: A avaliação do impacto científico depende, majoritariamente, de indicadores baseados em citações, que apenas se consolidam após longos períodos, limitando sua utilidade em decisões editoriais, avaliativas e de gestão da informação. Este estudo investiga em que medida o título de artigos científicos, isoladamente ou enriquecido com metadados bibliográficos simples disponíveis antes da publicação, pode antecipar o impacto futuro em termos de citações. O problema é formulado como classificação binária de artigos de alto e baixo impacto. Os experimentos utilizam 78.707 artigos da revista Physical Review A, da American Physical Society, com impacto definido por agrupamento por quartis de citações. Emprega-se um modelo baseado em representações linguísticas contextuais, avaliado por validação cruzada estratificada (5-fold), comparando o uso exclusivo do título e sua combinação com metadados simples, ano de publicação, número de autores e pontuação. Os resultados, avaliados pelo Coeficiente de Correlação de Matthews, métrica normalizada cujo valor máximo é 1 e valores próximos de 0 indicam desempenho aleatório, indicam desempenho limitado para títulos isolados (MCC ≈ 0,283) e ganho expressivo com metadados (MCC ≈ 0,509), estatisticamente significativo (Wilcoxon, $p < 0,05$). Os achados indicam o potencial de abordagens preditivas como apoio à avaliação científica, políticas editoriais e comunicação da ciência.

Palavras-chave: predição de citações; impacto científico; BERT; títulos de artigos; cientometria

1 Introdução

As práticas de citação são fundamentais para estruturar o conhecimento científico (Algaba *et al*., 2025), servindo originalmente para fundamentar argumentos e reconhecer trabalhos anteriores. Contudo, para além do seu propósito informativo, o número de citações tornou-se uma métrica central na ciência contemporânea para avaliar o impacto de pesquisas e a relevância de autores (SILVA *et al*., 2024; Nørgaard; Lie; Lund, 2026). O desafio reside no fato de que esse indicador só se torna disponível após um período significativo, geralmente anos após a publicação, o que limita sua utilidade para avaliações imediatas e decisões estratégicas (Wang; Song; Barabási, 2013).

Diante dessa lacuna temporal, surge a necessidade de ferramentas preditivas que permitam antecipar o potencial de impacto de um manuscrito ainda na fase de redação. Essa tarefa, embora desafiadora, pode contribuir para acelerar a identificação de trabalhos promissores e otimizar processos de seleção, divulgação, recomendação e financiamento de pesquisas.

Vital Jr. *et al*. (2025) demonstraram que é possível alcançar 80% de precisão na classificação de artigos pertencentes aos 20% mais citados utilizando apenas o resumo, o que justifica o desenvolvimento de ferramentas baseadas em texto.

Neste trabalho, em consonância com pesquisas anteriores, avaliamos a capacidade preditiva de modelos baseados em Transformers (Devlin *et al*., 2019) para estimar a classe a que pertencerão artigos científicos, utilizando apenas o título enriquecido com metadados muito básicos. O objetivo é identificar se essa configuração pode ser suficiente para esta tarefa de predição. Especificamente, este trabalho avalia a eficácia de modelos de classificação de texto do estado da arte sob condições de restrição de dados, examinando como a limitação informacional afeta sua robustez e precisão preditiva.

Os nossos resultados confirmam a eficácia de um classificador baseado na arquitetura Transformers para estimar o impacto de manuscritos, utilizando

exclusivamente metadados de pré-publicação (título, ano de publicação, número de autores e presença de sinais de pontuação no título). Os resultados demonstram que a concatenação dos metadados de pré-publicação com o título, formando uma única sequência textual, potencializa significativamente a capacidade preditiva do modelo em comparação ao uso isolado do título.

2 Trabalhos relacionados

Nas últimas décadas, a produção científica tem crescido exponencialmente, transformando a avaliação do impacto da pesquisa em um desafio complexo. Instituições de pesquisa, agências de financiamento e os próprios pesquisadores enfrentam a difícil tarefa de medir a relevância e a influência de trabalhos acadêmicos em meio a um volume massivo de publicações.

Neste contexto, as citações consolidaram-se como uma das métricas mais utilizadas para avaliar o impacto acadêmico. A importância e a influência das citações foram discutidas pela primeira vez por Garfield (1955), que lançou as bases da Cientometria moderna. No entanto, o uso de citações como métrica de impacto só é possível meses ou anos após a publicação, quando se podem observar os seus rastos na literatura. Neste contexto, a capacidade de prever o impacto de uma publicação científica utilizando apenas informações antecedentes à sua disponibilização tornou-se uma tarefa de grande relevância para a comunidade científica, pois permite uma avaliação precoce do potencial de influência de novos estudos.

Dada a relevância e a atualidade da predição do impacto de publicações (Zhang; Wu, 2024), diversas abordagens têm sido propostas sob diferentes perspectivas metodológicas. No âmbito das análises temporais e da dinâmica de citações, Ji *et al*. (2024) propuseram um novo modelo de aprendizagem profunda (Deep Learning) sequência a sequência (Seq2Seq) com um mecanismo inovador de atenção dinâmica hierárquica, o DMA-Nets, para a predição de citações a longo prazo. Os resultados experimentais demonstraram que o modelo é mais eficaz do que métodos anteriores tradicionais, alcançando maior precisão. Sob uma ótica diferente, Newman (2014) focou na vantagem temporal, propondo um modelo baseado na premissa de que os primeiros artigos a explorar um novo

campo tendem a receber mais citações. Validado cinco anos após a publicação, o estudo demonstrou que os artigos preditos como "muito citados" chegaram a receber até 23 vezes mais citações do que a média.

Complementando estas abordagens baseadas em métricas, Stegehuis, Litvak e Waltman (2015) utilizaram modelos de regressão para prever o total de citações futuras, considerando fatores como o fator de impacto da revista e o número de citações iniciais. Este trabalho evidenciou que, embora exista uma forte correlação com as métricas iniciais, a predição de impacto a longo prazo exige informações adicionais.

Com o avanço das técnicas computacionais, surgiram abordagens focadas no conteúdo textual. Recentemente, Hirako, Sasano e Takeda (2024) propuseram o modelo CiMaTe, baseado em modelos de linguagem como o BERT. A principal inovação do estudo foi o uso do texto completo do artigo dividido em seções estruturadas, o que permitiu melhorar o desempenho preditivo em diferentes áreas científicas.

A busca por eficiência através de metadados é evidenciada por Ma *et al*. (2021), que demonstraram como a semântica do título e do resumo pode prever citações. Esse potencial preditivo é corroborado por Baba, Baba e Ikeda (2019), cujos resultados validam o uso de características textuais do resumo, e por Galke *et al*. (2017), que analisaram o desempenho relativo de títulos e textos completos e concluíram que o ganho ao usar texto completo é pequeno.

Contudo, ainda existe pouco entendimento sobre o aumento de precisão obtido ao concatenar metadados bibliográficos pré-publicação num formato de texto contínuo. Especificamente, a integração de variáveis como o quantitativo de autores e o ano de publicação ao corpo do título – para processamento em modelos Transformers, como o BERT – configura uma abordagem ainda pouco explorada na literatura. A literatura carece de uma validação estatística robusta sobre se essa informação adicional justifica o custo computacional ou o título sozinho é suficiente.

Por fim, enquanto a nossa proposta foca na predição do potencial de impacto via metadados básicos, outras abordagens automatizam a alocação precisa de citações no corpo do texto (Jeong *et al*., 2020). Ambas as frentes

convergem para a consolidação de sistemas de suporte à decisão que elevam o rigor e a legibilidade da produção científica.

3 Procedimento metodológico

Para avaliar empiricamente a viabilidade de prever, com base apenas no título de um artigo científico, se este será muito citado, o problema foi definido como uma tarefa de classificação binária, na qual cada artigo é rotulado como muito citado ou pouco citado, a partir da comparação entre o número de citações recebidas por cada artigo e medidas estatísticas (como média, quartil ou mediana) observadas no conjunto de dados. A hipótese testada é que o enriquecimento semântico dos títulos, via metadados bibliográficos básicos disponíveis antes da publicação, proporciona ganho informacional suficiente para superar a capacidade preditiva obtida apenas com os títulos.

3.1 Coleta e pré-processamento dos dados

Os dados utilizados neste trabalho foram obtidos do repositório disponibilizado pela *American Physical Society* (APS). Este repositório reúne metadados de artigos científicos publicados nas revistas da APS, incluindo informações como título, autores, instituição de afiliação, ano de publicação, revista, referências e número de citações recebidas. A partir dos metadados recolhidos, foi realizada uma etapa de pré-processamento para organizar e padronizar as informações relevantes. Como parte desse processo, calculou-se a média do número de citações recebidas pelos artigos do conjunto de dados; especificamente para este estudo, consideramos os artigos publicados na revista *Physical Review A* (PRA).

Utilizou-se a média observada (aproximadamente oito citações) como limiar para a definição da variável alvo binária (Tabela 1). Seguindo abordagens anteriores (Vergoulis *et al*., 2021), estabeleceu-se que artigos com citações acima da média pertencem à Classe 1 (Alto Impacto), ao passo que artigos com citações iguais ou inferiores à média foram atribuídos à Classe 0 (Baixo Impacto).

No entanto, como o critério adotado para distinguir artigos “De Alto Impacto” e “Baixo Impacto” com base apenas na média de citações é metodologicamente frágil, realizamos experimentos adicionais usando outras

alternativas, como a mediana e agrupamentos por quartis para proporcionar uma segmentação mais robusta e representativa. Os resultados apresentados na seção Resultados referem-se a este último critério, o agrupamento por quartis, considerado mais correto.

A Tabela 1 apresenta o detalhamento dos seis conjuntos de dados experimentais. Observa-se que o critério 'Agrupamento pela Média' (Conjuntos de Dados 1 e 2) gera o cenário de maior desbalanceamento (71% vs 29%), refletindo a típica natureza de cauda longa das redes de citação, enquanto o critério da Mediana (Conjuntos de Dados 3 e 4) e o agrupamento por quartis (Conjuntos de Dados 5 e 6) proporcionam uma distribuição mais equilibrada. Por isso, optamos preferencialmente por esta última.

Tabela 1 - Resumo descritivo dos seis conjuntos de dados gerados a partir de diferentes limiares de citação

Critério adotado	Conjunto de Dados	N Total	Classe 0*	Classe 1	% Classe 0	% Classe 1
Agrupamento pela Média	1 e 2	78.707	55.989	22.718	71,14%	28,86%
Agrupamento pela Mediana	3 e 4	78.707	42.277	36.430	53.71%	46.29%
Agrupamento por quartis	5 e 6	43618	23153	20465	53.08%	46.92%

Nota: * Classe de artigos pouco citados. Os datasets com numeração ímpar contêm apenas o título e o rótulo; os datasets com numeração par contêm título mais metadados simples e o rótulo.
Fonte: Elaborado pelos autores.

A Tabela 2 apresenta amostras representativas das instâncias que compõem os conjuntos de dados. As entradas fornecidas aos modelos são de natureza textual e foram estruturadas de duas formas distintas: em uma configuração, a entrada consiste exclusivamente no título; na outra, os metadados foram concatenados ao título para constituir uma única sequência de entrada (*input string*).

Tabela 2 - Exemplos de instâncias do conjunto de dados com o título enriquecido por metadados simples

Título Enriquecido com Metadados	Sucesso (Rótulo)
Theory of Binary Boson Solutions, in the year of 1970, with 3 authors, without punctuation in the title	Sim (1)
Analytical Equation for Rigid Spheres: Application to the Model of Longuet-Higgins and Widom for the Melting of Argon, in the year of 1970, with 2 authors, with punctuation in the title	Não (0)

Fonte: Elaborado pelos autores.

Desta forma, foram obtidos dois conjuntos de dados: um apenas com o título e outro com o título concatenado com os metadados simples. Esta estratégia permitiu avaliar o desempenho dos modelos em diferentes contextos, desde situações com uma entrada textual mínima (somente o título) até configurações mais completas com mais informações. O principal objetivo foi investigar até que ponto as informações disponíveis antes da disponibilização da publicação, especialmente os metadados básicos, são suficientes para prever a classe a que o artigo pertencerá: muito citado ou pouco citado.

3.2 Modelo de classificação utilizado e avaliação do modelo

A evolução recente do Processamento de Linguagem Natural (PLN) foi marcada pela transição de representações estáticas para modelos fundamentados na arquitetura Transformer (Vaswani *et al*., 2017). Dentro deste contexto, embora modelos generativos como a família GPT (Brown *et al*., 2020) tenham ganhado notoriedade, optou-se neste trabalho pela utilização do BERT (Devlin *et al*., 2019). Ao contrário de modelos que processam o texto de forma linear, o BERT analisa cada termo em relação ao contexto completo da frase simultaneamente. Esta característica é determinante para o tratamento de enunciados de elevada densidade informativa, como os títulos de produções científicas, pois garante que o sentido de cada palavra seja interpretado à luz dos restantes componentes da frase, captando relações semânticas essenciais.

Adotamos a validação cruzada estratificada para garantir que cada partição refletisse a distribuição real das classes. Baseamo-nos em Kohavi (1995) para justificar esta escolha como um meio de obter estimativas de desempenho mais consistentes, tratando-a como um protocolo de validação rigoroso e não como uma técnica de reamostragem.

Um conjunto de métricas foi calculado de acordo com Koyejo *et al*., 2014, com especial atenção ao Matthews Correlation Coefficient (MCC) (Chicco; Jurman, 2020) e ao F1-Score, por serem métricas mais sensíveis à detecção da minoria (artigos de alto impacto). O Coeficiente de Correlação de Matthews (MCC) e o F1-Score quantificam a qualidade da classificação binária, sendo que uma pontuação de 1 representa uma predição perfeita e uma pontuação de 0 indica um desempenho não melhor que o acaso (ou ausência de precisão/recall, no caso do F1).

3.3 Análise estatística

Dadas as restrições computacionais que limitaram a amostragem a uma validação cruzada de cinco *folds* (k = 5), a análise estatística foi estruturada sob uma hipótese direcional. Formulou-se a hipótese alternativa de que o modelo enriquecido com metadados simples é superior ao modelo treinado apenas com títulos, mantendo-se a hipótese nula (H_0) de inexistência de ganho de desempenho. A verificação de significância foi conduzida através do teste de postos sinalizados de Wilcoxon unicaudal (one-tailed Wilcoxon Signed-Rank Test), adequado para detectar melhorias em amostras emparelhadas reduzidas (Demšar, 2006). Adotou-se um nível de significância de $\alpha = 0,05$; portanto, um p-valor > 0,05 implica a não rejeição de H_0, confirmando a ausência de evidência estatística de que a adição de metadados simples incrementa a eficácia do classificador.

4 Resultados e discussão

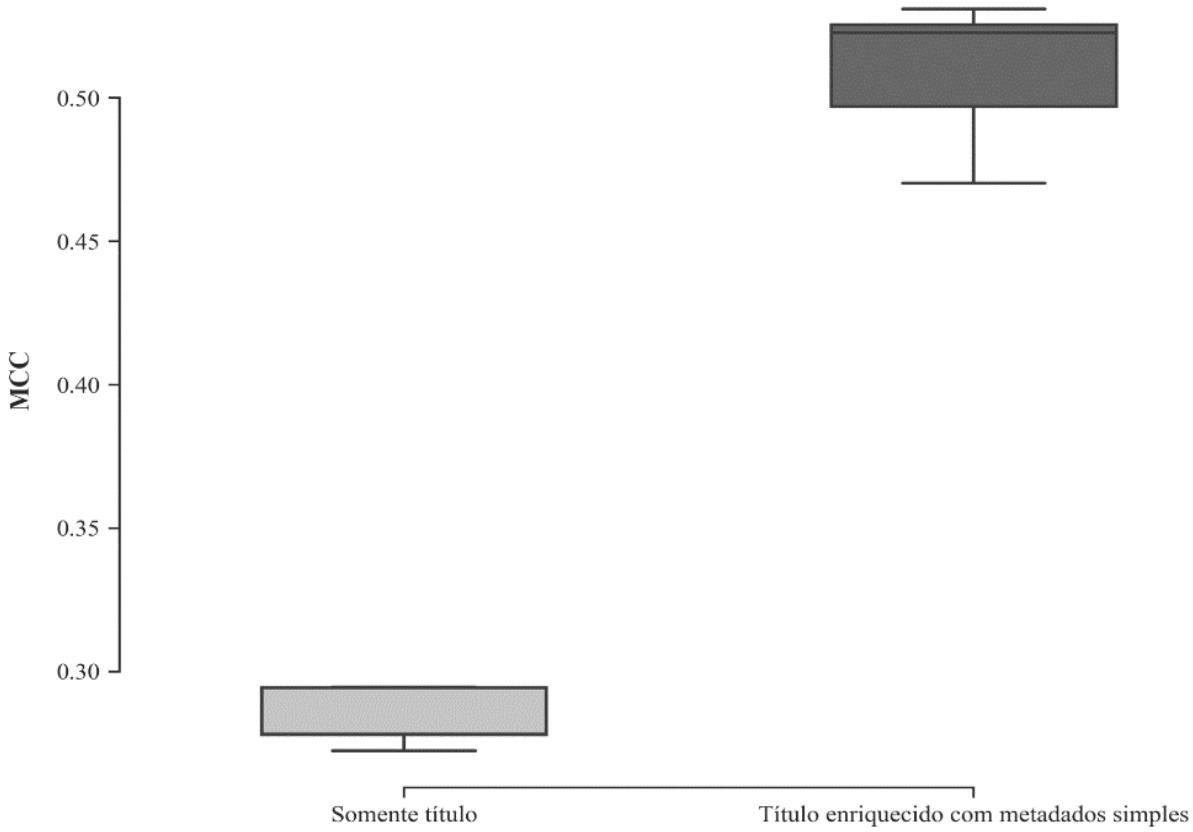
Para verificar a hipótese de que o título do artigo enriquecido com metadados simples, por si só, contém informações indicativas para inferir o impacto da publicação, foram construídas diferentes versões do conjunto de dados. A

principal contém exclusivamente os títulos dos artigos e as suas respectivas classes (alto impacto/baixo impacto). A outra versão inclui, além do título, metadados adicionais simples, disponíveis antes mesmo do acesso à publicação ser disponibilizado ao público, como o número de autores, o ano de publicação e a presença de sinais de pontuação no título. Essa variação foi deliberadamente concebida para avaliar se a adição de informações simples melhora significativamente o desempenho preditivo em comparação com o uso exclusivo do título.

4.1 Análise de desempenho preditivo usando o critério agrupamento por quartis

A avaliação comparativa entre os modelos baseou-se na métrica Matthews Correlation Coefficient (MCC), devido à sua robustez em cenários de desbalanceamento de classes. A Figura 1 apresenta a distribuição dos valores de MCC obtidos durante a validação cruzada (k = 5) para as abordagens baseadas exclusivamente no título e no título enriquecido com metadados. Observa-se uma distinção visual clara entre as distribuições. O modelo treinado apenas com títulos apresentou desempenho moderado, com valores de MCC próximos da média de 0,28. Em contrapartida, a inclusão de metadados simples (quantidade de autores, pontuação e ano) aumentou consistentemente a capacidade preditiva do classificador, resultando em valores de MCC superiores a 0,47 em todas as iterações de teste, com média aproximada de 0,51. Para validar a significância dessa diferença, aplicou-se o teste de postos sinalizados de Wilcoxon unicaudal. A análise estatística confirmou a superioridade do modelo enriquecido ($W = 15,0$; $p = 0,031$). Como o valor de p obtido é inferior ao nível de significância adotado ($\alpha = 0,05$), rejeita-se a hipótese nula. Os resultados evidenciam que, para a tarefa de classificação, a densidade informacional adicional fornecida pelos metadados bibliográficos disponíveis antes da disponibilização do artigo é estatisticamente determinante para a melhoria do desempenho do modelo BERT.

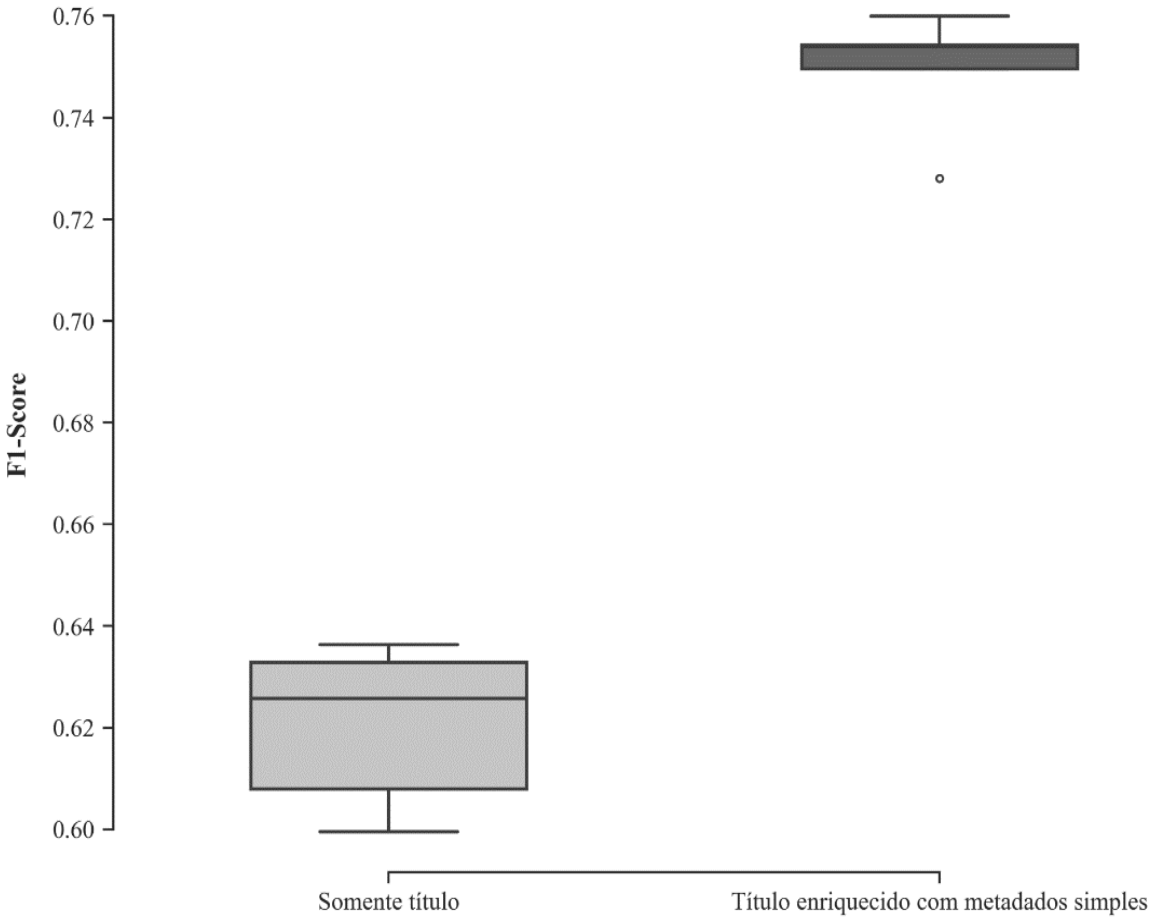
Figura 1 - Comparação do desempenho do BERT (5-fold CV)



Fonte: Dados da pesquisa.

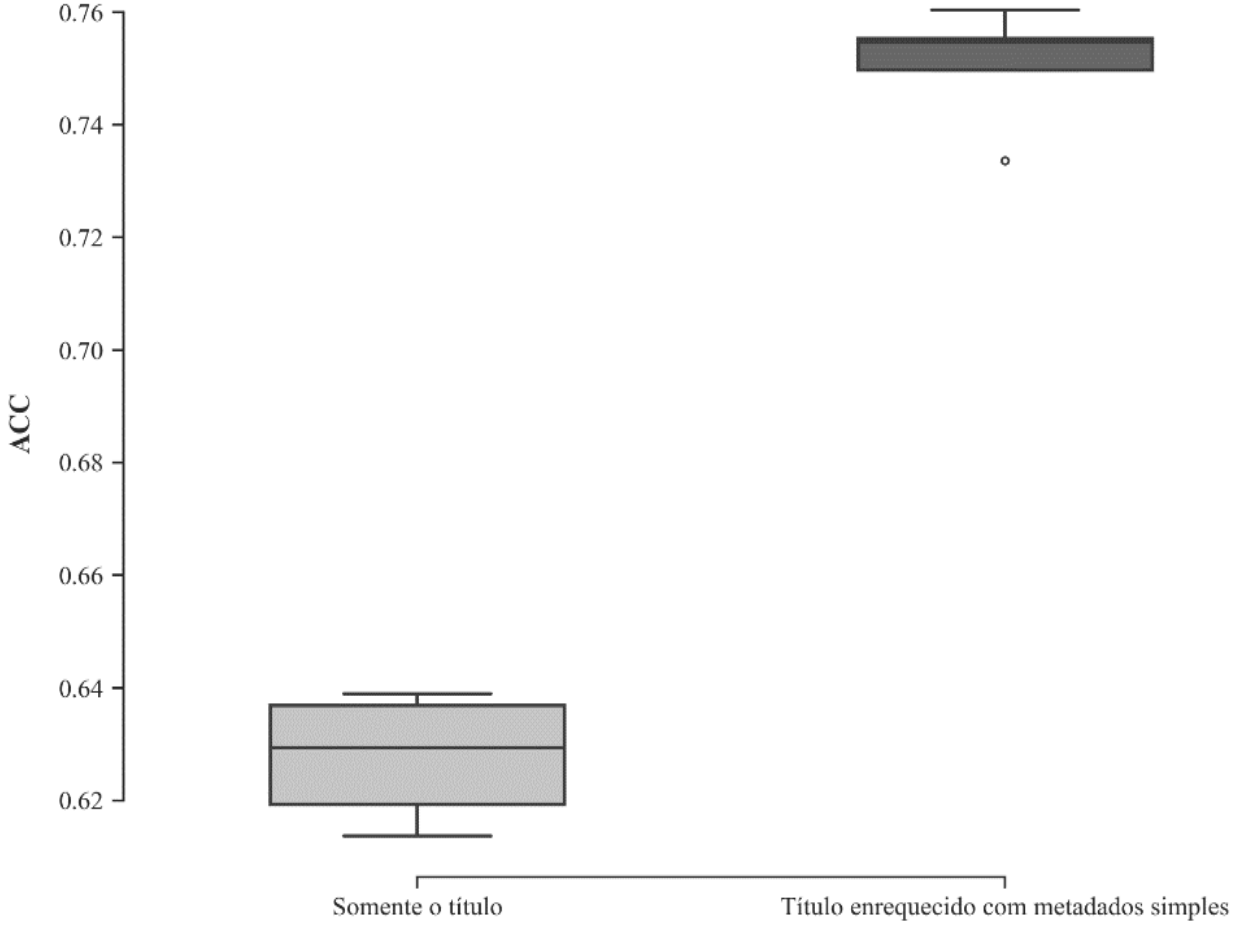
Para corroborar a robustez dos achados observados no MCC, o comportamento dos classificadores foi avaliado também sob a ótica do F1-Score ponderado (Figura 2) e da Acurácia (Figura 3). O padrão de superioridade do modelo enriquecido com metadados repetiu-se integralmente nestas métricas. Enquanto o modelo baseado apenas em títulos estagnou em patamares entre 0,60 e 0,63 para o F1-Score ponderado, a versão com metadados alcançou valores consistentemente superiores, oscilando entre 0,73 e 0,76. Comportamento idêntico foi registrado para a Acurácia. A validação estatística via teste de Wilcoxon unicaudal reproduziu o resultado anterior ($W = 15,0$; $p = 0,031$) para ambas as métricas adicionais, rejeitando a hipótese nula ao nível de significância de $\alpha = 0,05$. Essa convergência de resultados através de diferentes indicadores reforça a conclusão de que a informação contextual dos metadados oferece um ganho de generalização real e estatisticamente significativo ao modelo.

Figura 2 - Comparação de desempenho (5-fold CV) entre as abordagens: Distribuição do F1-Score Ponderado



Fonte: Dados da pesquisa.

Figura 3 - Comparação de desempenho (5-fold CV) entre as abordagens: Acurácia

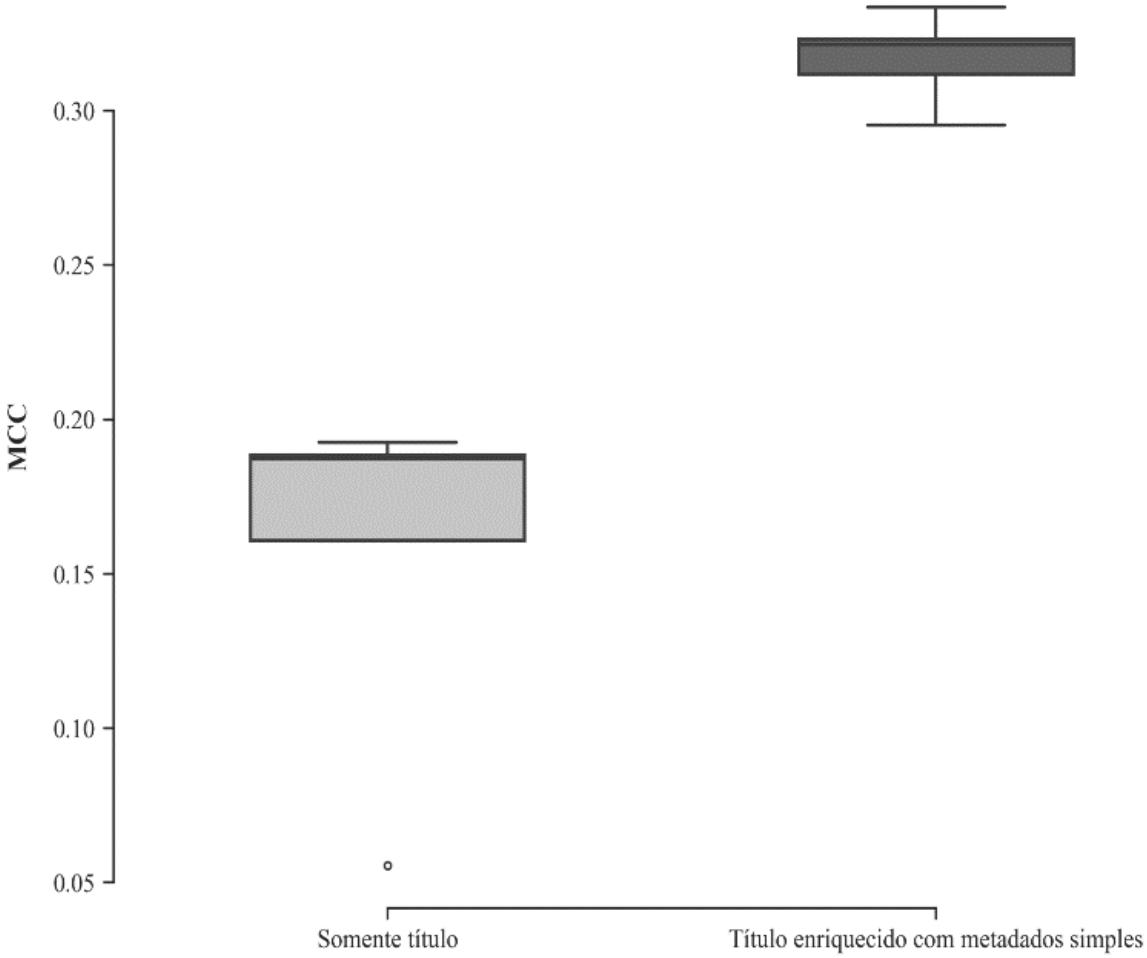


Fonte: Dados da pesquisa.

Ao contrário de trabalhos anteriores, como os de Hirako, Sasano e Takeda (2024), que utilizaram o texto completo dos artigos, e Baba, Baba e Ikeda (2019), que se basearam principalmente em resumos e metadados estruturados, o presente estudo centra-se exclusivamente no potencial preditivo dos metadados simples do artigo, disponíveis antes da publicação. Esta abordagem minimalista constitui uma contribuição inédita para a literatura, ao demonstrar empiricamente que, mesmo com entradas extremamente reduzidas, é possível obter resultados competitivos na tarefa de predição de impacto. Os resultados mostram que o título, por ser uma síntese concisa e cuidadosamente elaborada do conteúdo do artigo, contém sinais importantes para prever o seu impacto futuro e que a adição de metadados simples melhora significativamente o desempenho do modelo, algo que não tinha sido investigado diretamente em trabalhos anteriores.

Os resultados obtidos reforçam a hipótese central deste trabalho: é possível prever, com um desempenho competitivo, se um artigo será muito citado ou não, nos anos seguintes à sua publicação, utilizando apenas o seu título enriquecido com metadados simples. Isto permite aos autores da publicação avaliarem o título mesmo antes de a publicação ser submetida à apreciação dos pares, o que pode aumentar a qualidade do título. É importante mencionar que o nosso resultado não depende do critério adotado para distinguir artigos "Alto impacto" de "Baixo impacto". A Figura 4 apresenta a distribuição dos valores de MCC obtidos durante a validação cruzada (k = 5) para as abordagens baseadas exclusivamente no título e no título enriquecido com metadados para o critério média. A análise estatística confirmou a superioridade do modelo enriquecido (W = 15,0; p = 0,031).

Figura 4 - Comparação do Desempenho do BERT (5-fold CV)



Fonte: Dados da pesquisa.

Os resultados são promissores, mas devem ser usados com cautela. Eles não substituem a avaliação humana, servindo apenas como apoio à tomada de decisões.

A aplicação prática da nossa proposta consiste em um sistema de suporte à decisão editorial: um classificador preditivo que estima objetivamente o potencial de citação de um manuscrito. Ao integrar este classificador a Grandes Modelos de Linguagem (LLMs) – que geram opções de títulos baseadas no resumo (Teh; Uwasomba, 2024) – nossa ferramenta permite que cientistas testem e selecionem títulos que, além de maior probabilidade de impacto, estejam perfeitamente alinhados às suas descobertas. No entanto, o sistema opera como

um assistente; a palavra final cabe ao investigador, que deve validar se a sugestão mantém o rigor técnico e a fidelidade aos resultados da pesquisa.

Por último, como principais limitações deste trabalho, cabe destacar a restrição ao uso de informação textual básica, sem levar em conta o conteúdo completo dos artigos, resumos. Isto foi feito propositadamente porque, em determinadas circunstâncias, é possível que não disponhamos de toda a informação do artigo. Além disso, a análise baseou-se num recorte específico de dados (físicos), mas acreditamos que estes resultados podem ser generalizados a outras áreas do conhecimento ou bases bibliográficas. Como orientações futuras, propõe-se:

a) a incorporação de dados mais ricos, como resumos, palavras-chave e redes de coautoria, com o objetivo de melhorar o desempenho dos modelos;
b) a análise do impacto de diferentes estilos de redação de títulos sobre a atratividade e a difusão dos artigos;
c) a aplicação da metodologia proposta em diferentes áreas científicas, para avaliar sua solidez e capacidade de generalização.

Adicionalmente, espera-se que os resultados aqui apresentados contribuam para uma melhor compreensão dos fatores que influenciam o impacto das publicações científicas e inspirem novas investigações centradas na ciência de dados aplicada à bibliometria.

5 Considerações finais

O objetivo principal deste trabalho foi desenvolver e avaliar modelos preditivos capazes de classificar artigos científicos de acordo com o seu potencial de receber citação a curto prazo, utilizando como principal fonte de informação os títulos dos artigos e metadados simples.

A motivação para esta investigação surgiu da crescente procura por métodos que permitam antecipar o impacto das publicações científicas, uma vez que as citações, tradicionalmente utilizadas como métrica de relevância e que servem de base para outras métricas, só se acumulam com o tempo. Ao explorar o potencial informativo dos elementos disponíveis no momento da submissão,

procuramos contribuir com estratégias mais rápidas e eficientes para a identificação de artigos promissores.

Os resultados obtidos demonstraram que é possível, com níveis razoáveis de precisão, prever se um artigo estará entre os mais citados utilizando apenas informações simples concatenadas ao seu título. Esta descoberta corrobora a hipótese central deste trabalho, segundo a qual a informação presente nos metadados, mesmo sem acesso ao conteúdo completo do artigo, já contém elementos indicativos para que os modelos preditivos identifiquem padrões associados a um maior potencial de citação.

Contudo, o estudo mostrou que, embora o uso exclusivo do título já forneça sinais relevantes, a inclusão de metadados adicionais, como o ano de publicação, a autoria e a presença de pontuação no título, pode contribuir para melhorar a capacidade preditiva dos modelos. Isto sugere que, apesar da simplicidade dos dados textuais principais, existe um valor acrescentado em considerar características complementares para refinar as previsões.

Para concluir, este trabalho contribui para o debate sobre novas formas de avaliar os cientistas. A ideia é utilizar métodos complementares que levem em consideração o potencial futuro, e não apenas os resultados já obtidos. No entanto, a decisão final deve ser sempre humana, levando também em consideração a capacidade do cientista de ampliar as suas redes de colaboração, a qualidade dos seus trabalhos anteriores, a sua capacidade de gestão, e outros fatores.

Financiamento

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Agradecimentos

Os autores agradecem à *American Physical Society* (APS) por disponibilizar os *APS Dataset* para pesquisa, que foram fundamentais para a realização deste trabalho. Também agradecemos aos colegas e colaboradores que contribuíram com sugestões e discussões relevantes durante o desenvolvimento desta pesquisa.

Referências

ALGABA, Andres *et al*. Large language models reflect human citation patterns with a heightened citation bias. *In*: FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2025, Albuquerque. **Proceedings** [...]. Stroudsburg: Association for Computational Linguistics, 2025. p. 6829-6864.

BABA, Takahiro; BABA, Kensuke; IKEDA, Daisuke. Citation count prediction using abstracts. **Journal of Web Engineering**, Gistrup, v. 18, n. 1-3, p. 207-228, 2019. Disponível em: https://doi.org/10.13052/jwe1540-9589.18136. Acesso: 20 mar. 2026.

BROWN, Tom B. *et al*. Language models are few-shot learners. *In*: NEURAL INFORMATION PROCESSING SYSTEMS, 33., 2020, San Diego. **Proceedings** [...]. New York: Curran Associates, 2020. p. 1877-1901.

CHICCO, David; JURMAN, Giuseppe. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. **BMC Genomics**, New York, v. 21, n. 6, p. 1-13, 2020. Disponível em: https://doi.org/10.1186/s12864-019-6413-7. Acesso: 20 mar. 2026.

DEMŠAR, Janez. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine Learning Research**, Massachusetts, v. 7, p. 1-30, 2006.

DEVLIN, Jacob *et al*. BERT: pre-training of deep bidirectional transformers for language understanding. *In*: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2019, Minneapolis. **Proceedings** [...]. Stroudsburg: Association for Computational Linguistics, 2019. p. 4171-4186.

GALKE, Lukas *et al*. Using titles vs. full-text as source for automated semantic document annotation. *In*: KNOWLEDGE CAPTURE CONFERENCE, 9, 2017, Austin. **Proceedings** [...]. New York: Association for Computing Machinery, 2017.

GARFIELD, Eugene. Citation indexes for science: a new dimension in documentation through association of ideas. **Science**, Washington, v. 122, n. 3159, p. 108-111, 1955. Disponível em: https://doi.org/10.1126/science.122.3159.108. Acesso: 20 mar. 2026.

HIRAKO, Jun; SASANO, Ryohei; TAKEDA, Koichi. CiMaTe: citation count prediction effectively leveraging the main text. **arXiv**, Ithaca, 6 Oct., 2024. Disponível em: https://doi.org/10.48550/arXiv.2410.04404. Acesso: 20 mar. 2026.

JEONG, Chanwoo *et al*. A context-aware citation recommendation model with BERT and graph convolutional networks. **Scientometrics**, New York, v. 124, n.

3, p. 1907-1922, 2020. Disponível em: https://doi.org/10.1007/s11192-020-03561-y. Acesso: 20 mar. 2026.

JI, Taoran *et al*. Citation forecasting with multi-context attention-aided dependency modeling. **ACM Transactions on Knowledge Discovery from Data**, New York, v. 18, n. 6, p. 1-23, 2024. Disponível em: https://doi.org/10.1145/3649140. Acesso: 20 mar. 2026.

KOHAVI, Ron. A study of cross-validation and bootstrap for accuracy estimation and model selection. *In*: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 14., 1995, Montreal. **Proceedings** [...]. San Francisco: Morgan Kaufmann, 1995. p. 1137-1143.

KOYEJO, Oluwasanmi *et al*. Consistent binary classification with generalized performance metrics. *In*: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 27., 2014, Montreal. **Proceedings** [...]. New York: Curran Associates, 2014.

MA, Anqi; LIU, Yu; XU, Xiujuan; DONG, Tao. A deep-learning based citation count prediction model with paper metadata semantic features. **Scientometrics**, New York, v. 126, n. 8, p. 6803-6823, 2021. Disponível em: https://doi.org/10.1007/s11192-021-04033-7. Acesso: 20 mar. 2026.

NØRGAARD, Birgitte; LIE, Karen E.; LUND, Hans. Predictors of citation rates and the problem of citation bias: a scoping review. **Journal of Clinical Epidemiology**, Amsterdam, v. 190, p. 1-13, 2026. Disponível em: https://doi.org/10.1016/j.jclinepi.2025.112057. Acesso: 20 mar. 2026.

NEWMAN, Mark E. J. Prediction of highly cited papers. **Europhysics Letters**, Mulhouse, v. 105, n. 2, p. 28002, 2014. Disponível em: https://doi.org/10.1209/0295-5075/105/28002. Acesso: 20 mar. 2026.

SILVA, Maurício Coelho; WENDT, Lucas George; MOURA, Ana Maria Mielniczuk; ARAUJO, Ronaldo Ferreira. O sistema de recompensa científico na ótica da Ciência Aberta: dimensões de avaliação, características e desafios. **Transinformação**, Campinas, v. 36, p. 1-18, 2024. Disponível em: https://doi.org/10.1590/2318-0889202436e2410680. Acesso em: 20 de março de 2026.

STEGEHUIS, Clara; LITVAK, Nelly; WALTMAN, Ludo. Predicting the long-term citation impact of recent publications. **Journal of informetrics**, Amsterdam, v. 9, n. 3, p. 642-657, 2015. Disponível em: https://doi.org/10.1016/j.joi.2015.06.005. Acesso: 20 mar. 2026.

TEH, Phoey Lee; UWASOMBA, Chukwudi Festus. Impact of large language models on scholarly publication titles and abstracts: a comparative analysis. **Journal of Social Computing**, Beijing, v. 5, n. 2, p. 105-121, 2024. Disponível em: https://doi.org/10.23919/JSC.2024.0011. Acesso: 20 mar. 2026.

VASWANI, Ashish *et al*. Attention is all you need. *In*: NEURAL INFORMATION PROCESSING SYSTEMS, 30., 2017, Long Beach. **Proceedings** [...]. New York: Curran Associates, 2017. p. 5998-6008.

VERGOULIS, Thanasis. *et al*. Simplifying impact prediction for scientific articles. *In*: EDBT/ICDT 2021 Joint Conference Workshops, 2021, Nicosia. **Proceedings** [...]. Aachen: CEUR-WS.org, 2021.

VITAL JR., Adilson *et al*. Predicting citation impact of research papers using GPT and other text embeddings. **Physica A: Statistical Mechanics and its Applications**, Amsterdam, v. 674, p. 130789, 2025. Disponível em: https://doi.org/10.1016/j.physa.2025.130789. Acesso: 20 mar. 2026.

WANG, Dashun; SONG, Chaoming; BARABÁSI, Albert-László. Quantifying long-term scientific impact. **Science**, Washington, v. 342, n. 6154, p. 127-132, 2013. Disponível em: https://doi.org/10.1126/science.1237825. Acesso: 20 mar. 2026.

ZHANG, Fang; WU, Shengli. Predicting citation impact of academic papers across research areas using multiple models and early citations. **Scientometrics**, New York, v. 129, p. 4137-4166, 2024. Disponível em: https://doi.org/10.1007/s11192-024-05086-0. Acesso: 20 mar. 2026.

First impressions in science: predicting paper impact from article titles and basic metadata

Abstract: The assessment of scientific impact relies largely on citation-based indicators, which only consolidate over long periods, limiting their usefulness for early editorial decisions, research evaluation, and information management. This study investigates to what extent article titles, used alone or enriched with simple bibliographic metadata available prior to publication, can anticipate future citation impact. The problem is formulated as a binary classification task, distinguishing high-impact and low-impact articles. Experiments are conducted on 78,707 articles from Physical Review A, published by the American Physical Society, with impact defined using citation quartile grouping. A model based on contextual linguistic representations is employed and evaluated using stratified 5-fold cross-validation, comparing the exclusive use of titles with their combination with simple metadata, namely publication year, number of authors, and punctuation. Results are assessed using the Matthews Correlation Coefficient, a normalized metric whose maximum value is 1 and values close to 0 indicate random performance. The findings show limited performance for titles alone (MCC ≈ 0.283) and a substantial improvement when metadata are included (MCC ≈ 0.509), which is statistically significant (Wilcoxon, $p < 0.05$). These results highlight the potential of predictive approaches as support tools for research evaluation, editorial policies, and scholarly communication.

Keywords: citation prediction; scientific impact; BERT; article titles; scientometrics

Declaração de autoria

Concepção e elaboração do estudo: João Pedro Cavalcanti Azevedo, Antônio de Abreu Batista Junior, Yonara Costa Magalhães e Jesús P. Mena Chalco
Coleta de dados: João Pedro Cavalcanti Azevedo
Análise e interpretação de dados: João Pedro Cavalcanti Azevedo, Antônio de Abreu Batista Junior, Yonara Costa Magalhães e Jesús P. Mena Chalco
Redação: João Pedro Cavalcanti Azevedo, Antônio de Abreu Batista Junior, Yonara Costa Magalhães e Jesús P. Mena Chalco
Revisão crítica do manuscrito: João Pedro Cavalcanti Azevedo, Antônio de Abreu Batista Junior, Yonara Costa Magalhães e Jesús P. Mena Chalco

Declaração de disponibilidade de dados

O conjunto de dados que dá suporte aos resultados deste estudo pode ser solicitada aos autores (ver autoria para correspondência).

Autoria para correspondência

Antônio de Abreu Batista Junior
antonio.batista@ufma.br

Editor-chefe

Thiago Henrique Bragato Barros

Como citar

MAGALHÃES, Yonara Costa, AZEVEDO, João Pedro Cavalcanti, BATISTA JUNIOR, Antônio de Abreu, MENA-CHALCO, Jesús Pascual. Primeiras impressões na ciência: prevendo o impacto de artigos a partir de títulos e metadados básicos. **Em Questão**, Porto Alegre, v. 32, e-149912, 2026. DOI: https://doi.org/10.1590/1808-5245.32.149912

Recebido: 30/08/2025
Aceito: 27/01/2026