Eficiência e Precisão na Extração de Dados Científicos: Um Estudo de Caso com Robôs Automatizados

Efficiency and Accuracy in Scientific Data Extraction: A Case Study with Automated Robots

Edson Melo de Souza https://orcid.org/0000-0002-5891-4767 Doutor em Informática e Gestão do Conhecimento. Universidade Nove de Julho (Uninove) – Brasil. souzaem@uni9.pro.br

Wonder Alexandre Luz Alves https://orcid.org/0000-0003-0430-950X Doutor em Ciência da Computação. Universidade de São Paulo (USP) – Brasil. wonder@uni9.pro.br

RESUMO

A coleta de grandes volumes de dados tem se tornado um desafio crescente em diversas áreas, incluindo negócios, saúde, governo e pesquisa acadêmica. Métodos manuais, apesar de amplamente utilizados, são ineficientes, propensos a erros humanos e incapazes de lidar com o aumento exponencial de informações. Este estudo apresenta uma abordagem automatizada para a extração de dados autorais em artigos científicos, utilizando robôs que aumentam a eficiência e a precisão do processo. Baseado nos avanços da pesquisa realizada por De Souza, Storopoli e Alves (2022), o trabalho aplica técnicas de raspagem de dados (web *scraping*) em 901 artigos da revista *Ann. Intern. Med.*, com o uso de três robôs: RCDA (Robô Coletor de Dados de Autores), RCCA (Robô Coletor de Contribuições de Autores) e RIA (Robô Individualizador de Autores). A automação reduziu o tempo de coleta em até 99,9%, ampliando a precisão para uma taxa de erro inferior a 0,01%. A abordagem proposta não apenas valida a eficiência dos robôs utilizados, mas também demonstra sua escalabilidade e aplicabilidade em contextos além da pesquisa científica.

Palavras-chave: robôs; automação; raspagem de dados; coleta automatizada; ciência de dados.

ABSTRACT

Collecting large volumes of data has become an increasing challenge across various fields, including business, healthcare, government, and academic research. While manual methods are widely used, they are inefficient, prone to human errors, and incapable of handling the exponential growth of information. This study presents an automated approach for extracting authorial data from scientific articles, leveraging robots to enhance the process's efficiency and accuracy. Building on the research advancements of De Souza, Storopoli, and Alves (2022), this study applies web scraping techniques to 901 articles from the Ann. Intern. Med. journal, utilizing three specialized robots: RCDA (Author Data Collector Robot), RCCA (Author Contributions Collector Robot), and RIA (Author Individualizer Robot). Automation reduced data collection time by up to 99.9% while improving

accuracy to an error rate of less than 0.01%. The proposed approach not only validates the efficiency of the deployed robots but also demonstrates their scalability and applicability beyond scientific research.
Keywords: robots; automation; web scraping; automated data collection; data science.

Recebido em 10/02/2025. Aprovado em 12/03/2025. Avaliado pelo sistema *double blind peer review*. Publicado conforme normas da ABNT.
https://doi.org/10.22279/navus.v16.2085

1 INTRODUÇÃO

Nos anos recentes, observamos um aumento acelerado e significativo na geração de grandes volumes de dados (Helena Presser; Lopes Da Silva, 2018). Nesse contexto, o volume de informações originadas da internet impõe desafios significativos para a coleta e análise de dados relevantes. A progressiva digitalização da informação tem remodelado profundamente o cenário da pesquisa científica em nível global. Conforme De Souza, Storopoli e Alves (2022), métodos manuais de coleta de dados, frequentemente usados na pesquisa acadêmica, são incapazes de atender à demanda crescente. Além disso, tais métodos comprometem a precisão, tornando o processo mais suscetível a erros humanos.

A habilidade de extrair valor de grandes volumes de dados não é apenas uma necessidade técnica, mas uma vantagem estratégica essencial no cenário contemporâneo. Conforme discutido por Nascimento *et al*. (2018), a análise de Big Data permite que as empresas inovem em seus negócios, utilizando plataformas que facilitam a coleta e interpretação de grandes volumes de dados, resultando em decisões mais informadas e estratégias competitivas aprimoradas. Além disso, Ataides e Nacife (2024) destacam que o Big Data Analytics (BDA) fornece às organizações ferramentas para analisar dados de forma eficaz, colaborando para alcançar e preservar vantagens competitivas. De acordo com Sewald Junior e Santos Júnior (2024), a capacidade de processar e interpretar grandes volumes de dados permite que as empresas aprimorem seus processos decisórios, tornando-se mais ágeis e adaptativas no mercado atual.

Na área da saúde, por exemplo, a análise de dados em grande escala pode oferecer *insights* significativos para o diagnóstico e tratamentos personalizados, enquanto nos negócios, a identificação de padrões em grandes bases de dados pode antecipar tendências de mercado. Entretanto, métodos manuais de coleta e processamento de dados, com frequência aplicados em pesquisa científica, são lentos e suscetíveis a erros humanos, comprometendo a precisão e a eficiência da análise (Morais *et al*., 2019).

Na pesquisa científica, é comum ocorrer situações onde um pesquisador enfrenta o desafio de analisar milhares de artigos para identificar padrões de autoria e contribuição científica. Este processo, executado de forma manual, pode demandar semanas ou até meses, além de estar sujeito a inconsistências. Segundo Camilo e Garrido (2019), "A comunidade científica tem assistido a um incremento exponencial na divulgação de ciência, com novas revistas lançadas anualmente e milhares de trabalhos de pesquisa publicados em vários domínios científicos entre os quais a Psicologia.". Esse aumento significativo na produção científica torna os métodos tradicionais de revisão mais desafiadores e suscetíveis a falhas, afetando potencialmente a qualidade das decisões baseadas nesses dados.

Neste contexto, a automação surge como uma solução para mitigar esses obstáculos. Ferramentas como algoritmos de *web scraping* e *pipelines* de ciência de dados têm transformado a forma como informações são coletadas, processadas e analisadas. Estudos como o de Raghupathi e Raghupathi (2014) destacam como sistemas automatizados na área médica podem reduzir drasticamente o tempo de coleta de dados, enquanto Zhao *et al*. (2020) ilustram a aplicação de Big Data para o desenvolvimento de novos medicamentos.

Portanto, este estudo propõe uma abordagem automatizada para a extração de dados científicos por meio dos robôs RCDA, RCCA e RIA, projetados para processar grandes volumes de informações com maior precisão e menor tempo de coleta (De Souza; Storopoli; Alves, 2022). Utilizando um estudo de caso com

artigos da revista *Ann. Intern. Med.*, a pesquisa demonstra como a automação pode transformar a coleta de dados autorais, garantindo eficiência, reprodutibilidade e escalabilidade. Além de detalhar a implementação técnica desses agentes inteligentes, o trabalho amplia suas funcionalidades e explora sua aplicabilidade em diferentes áreas, como negócios, saúde e ciências sociais, contribuindo para a consolidação deste paradigma para a extração automatizada de dados científicos em larga escala.

Adicionalmente, são debatidos aspectos éticos relativos à extração automatizada de informações, em especial quando há necessidade de tratar bases proprietárias ou com restrições de uso. Também se enfatizam possíveis riscos de obsolescência dos algoritmos, pois estruturas de páginas da web podem ser alteradas e requerer adaptações contínuas do *crawler*. Por fim, faz-se uma análise crítica das limitações do modelo proposto, sobretudo no que se refere à generalização dos resultados para outras revistas científicas e bases de dados.

2 FUNDAMENTAÇÃO TEÓRICA

A fundamentação teórica deste trabalho explora os pilares que sustentam a automação da coleta de dados científicos. Os principais conceitos abordados incluem: (i) autoria e posicionamento autoral, (ii) Big Data, (iii) automação de processos, (iv) *web scraping*, (v) impacto da automação e da inteligência artificial na análise de dados científicos e (vi) possíveis implicações éticas e limites da automação.

2.1 Autoria e Posicionamento Autoral

O conceito de autoria em publicações científicas envolve tanto a contribuição ativa para o estudo quanto questões éticas, como autoria fantasma e honorária (Kumar, 2018). No início do século XX, era comum que artigos científicos tivessem apenas um autor (Lozano, 2014), mas a ciência moderna, impulsionada pela colaboração interdisciplinar, resultou no aumento do número de coautores por publicação (Rosenzweig *et al.*, 2008; Patience *et al.*, 2019). A responsabilidade ética dos autores pelo conteúdo publicado é um aspecto fundamental nas práticas científicas, pois reflete não apenas a autoria, mas também o compromisso público com a integridade e a veracidade da pesquisa (Hilário *et al.*, 2018).

O *byline*, segundo Yang, Wolfram e Wang (2017), é a identificação do posicionamento autoral em um artigo é realizada com base na informação dos nomes dos autores, que aparece no início do texto, também conhecida como *Author byline*. Essa disposição está diretamente relacionada à relevância dos autores na publicação. O *byline* não é meramente um crédito nominal, mas um símbolo de *accountability (*Peng *et al.*, 2024). Ao associar seus nomes a um artigo, os autores assumem, perante a comunidade acadêmica e a sociedade, a responsabilidade por cada afirmação, metodologia e conclusão apresentada. Isso inclui garantir a precisão dos dados, a originalidade das ideias, a adequação das citações e a transparência em eventuais conflitos de interesse. Entretanto, o posicionamento dos autores no *byline* não segue um padrão universalmente aceito, gerando debates sobre critérios de contribuição (Yang; Wolfram; Wang, 2017). Em geral, os primeiros autores desempenham papel central no estudo, enquanto os últimos são responsáveis por supervisão e captação de recursos (Zbar; Frank, 2011), sendo que a posição de autor correspondente

frequentemente reflete maior participação intelectual (Mattsson; Sundberg; Laget, 2011).

Problemas éticos surgem quando indivíduos com pouca ou nenhuma contribuição significativa são incluídos na autoria (Jones; McCullough; Richman, 2005). Além dos critérios formais, pesquisas apontam fatores simbólicos na definição da autoria, como Igou e Van Tilburg (2015), que sugeriram que o número de letras no nome do meio de um autor poderia influenciar sua posição no *byline*. Já Russell *et al*. (2019) observaram que, à medida que o número de autores cresce, a posição do autor correspondente tende a migrar da primeira para a última posição, indicando um papel de liderança.

A determinação da ordem de autoria costuma estar atrelada à contribuição individual, sendo os primeiros e últimos autores aqueles com maior envolvimento (Tscharntke *et al*., 2007). Doutorandos e pós-doutorandos frequentemente ocupam a primeira posição, enquanto líderes de pesquisa ou diretores de laboratório figuram como últimos autores (Mongeon *et al*., 2017). No entanto, mensurar as contribuições individuais com precisão continua sendo um desafio (Chang, 2019). Em colaborações internacionais, é necessário considerar as práticas culturais na definição do posicionamento autoral (Aria; Misuraca; Spano, 2020), reforçando a importância de diretrizes claras para garantir a justa representação da autoria na ética científica.

2.2 Ciência de Dados e Big Data

A ciência de dados, enquanto campo interdisciplinar, integra metodologias computacionais e estatísticas para transformar grandes volumes de dados em conhecimento acionável. Segundo Novaes (2024), "A Ciência de dados, como disciplina, transcende a mera coleta de informações, exigindo uma abordagem metódica e criteriosa na interpretação dos dados.". Ainda, segundo o mesmo autor, sua essência não reside apenas na análise pontual, mas no desenvolvimento de algoritmos e modelos preditivos que sustentam decisões estratégicas. Não por acaso, essa capacidade tem sido amplamente explorada no setor empresarial: Sewald Junior e Santos Júnior (2024) reforçam que, diante do crescimento exponencial de dados disponíveis, as organizações adotam técnicas de ciência de dados como ferramenta crítica para ganhar vantagem competitiva, seja na otimização de processos, seja na antecipação de tendências.

Byrne *et al*. (2017) delineiam o *pipeline* da ciência de dados, que abrange desde a formulação de perguntas relevantes até a visualização dos resultados, passando por coleta, exploração e modelagem. Dentro dessa cadeia, destaca-se o pré-processamento de dados — etapa que, segundo Fan *et al*. (2021), pode consumir até 80% do tempo de um projeto, especialmente quando se lida com informações desestruturadas ou incompletas. Nesse contexto, ferramentas automatizadas, como as propostas por Koehler *et al*. (2017), emergem como aliadas indispensáveis, reduzindo erros humanos e acelerando a preparação de bases complexas. Para que essa aplicação seja viável, é necessário um fluxo estruturado de trabalho (Antypas *et al*., 2021).

Essa demanda por eficiência está intrinsecamente ligada ao fenômeno do Big Data, termo que, segundo Gandomi e Haider (2015) e Li, Chen e Shang (2022), descreve o crescimento vertiginoso no volume, variedade e velocidade de dados gerados. Se nas empresas o desafio é transformar dados em *insights*, no meio acadêmico, como apontam Falsarella e Jannuzzi (2020), a necessidade de processar milhões de artigos científicos publicados anualmente impulsiona

o desenvolvimento de tecnologias capazes de filtrar, analisar e sintetizar informações em escala. Assim, tanto na prática corporativa quanto na pesquisa, a ciência de dados revela-se um pilar central, conectando desafios técnicos — como o pré-processamento — a soluções inovadoras que respondem às exigências de um mundo cada vez mais orientado por dados.

A interligação entre teoria, ferramentas e aplicações práticas não apenas define a ciência de dados, mas também reforça seu papel transformador em múltiplos setores, unindo rigor metodológico à capacidade de gerar impacto real.

2.3 Automação de Processos

A automação, atualmente indispensável em práticas científicas e outras áreas do conhecimento, vai além da mera substituição de processos manuais: ela redefine como a ciência é conduzida, desde a coleta de dados até a validação de hipóteses. Utilizando algoritmos e ferramentas computacionais, essa abordagem não apenas reduz a intervenção humana em tarefas repetitivas, mas também introduz precisão, escalabilidade e velocidade inatingíveis por métodos tradicionais. Kumar *et al*. (2021) e Xu, Kumar e LeBeau (2022) demonstram, por exemplo, que a automação em *workflows* de Aprendizado de Máquina ou Machine Learning (ML) reduz erros de análise em até 40%, enquanto acelera a entrega de resultados em projetos de grande escala. Além disso, em laboratórios de pesquisa experimental, sistemas robóticos autônomos, como os descritos por Burger *et al*. (2020), estão revolucionando a descoberta de materiais e fármacos, executando experimentos químicos 24/7 com mínima supervisão humana.

No âmbito da coleta e processamento de dados científicos provenientes de publicações, a automação emerge como uma solução crítica para superar desafios como a fragmentação de fontes, a heterogeneidade de formatos (PDFs, HTML, XML) e o volume exponencial de artigos gerados anualmente. Ferramentas como Scrapy e Selenium (Pant *et al*., 2024; Kovalenko, 2025) são adaptadas para extrair dados estruturados de repositórios acadêmicos, como PubMed, arXiv e Scopus, automatizando a captura de metadados, citações e até conteúdo completo. Essa abordagem é combinada com técnicas de mineração de texto e processamento de linguagem natural (PLN) para classificar artigos, identificar tendências temáticas e mapear redes de colaboração científica.

Um exemplo é o trabalho de Falsarella e Jannuzzi (2020), que desenvolveram um *pipeline* automatizado para agregar e padronizar dados de mais de 2 milhões de artigos científicos, permitindo análises bibliométricas em larga escala. Sistemas como o GROBID (Lopez, 2009), por sua vez, utilizam algoritmos de reconhecimento de entidades para converter PDFs de publicações em dados estruturados, reduzindo erros manuais em até 70% em projetos de revisão sistemática. Essa capacidade é importante em áreas como a medicina, onde a velocidade de acesso a estudos recentes — como ensaios clínicos publicados em plataformas como ClinicalTrials.gov — pode influenciar diretamente decisões terapêuticas (Higgins *et al*., 2022).

Além disso, a automação garante a integridade e a reprodutibilidade na curadoria de dados de publicações. Plataformas como o OpenAlex (Priem; Piwowar; Orr, 2022) empregam *scripts* automatizados para atualizar diariamente seu banco de dados com novos artigos, corrigindo inconsistências e vinculando autores a instituições de forma precisa. Contudo, desafios persistem, como a necessidade de lidar com licenças restritivas de acesso e a variação na qualidade de metadados entre periódicos — questões que demandam estratégias

de *retry* inteligente e validação cruzada, conforme discutido por Piedra *et al.* (2023) em seu estudo sobre interoperabilidade em repositórios acadêmicos.

2.4 Web Scraping

Web scraping é uma técnica de extração de dados estruturados ou semiestruturados de páginas da web (Souza *et al.*, 2024). Essa abordagem tem sido amplamente aplicada em áreas como inteligência de negócios, marketing digital e pesquisa científica, especialmente em cenários onde APIs, do inglês *Application Programming Interface*, não estão disponíveis ou são insuficientes (Singrodia; Mitra; Paul, 2019).

De acordo com Khder (2021), o *web scraping* desempenha um papel essencial na transformação de grandes volumes de dados não estruturados em informações utilizáveis. Por exemplo, na pesquisa acadêmica, ferramentas de *scraping* podem coletar informações como autoria, afiliações e contribuições científicas diretamente de páginas de periódicos, complementando ou substituindo métodos baseados em APIs.

Além de capturar dados textuais, o *web scraping* também permite a análise de metadados e padrões visuais presentes em páginas da web (Glez-Peña *et al.*, 2014; Zhao, 2022). Essa técnica se mostra particularmente eficaz quando integrada a outras metodologias, como o aprendizado de máquina, para identificar padrões ou realizar classificações automáticas (Jordan; Mitchell, 2015; Kalaivani; Kamalakkannan, 2022).

Porém, tais benefícios vêm acompanhados de desafios, como a possibilidade de mudanças frequentes na estrutura das páginas, limitação de requisições por parte dos servidores e implicações legais e éticas para o acesso a certos tipos de dados, sobretudo quando as páginas não foram projetadas para compartilhamento livre (Nunes, 2024).

2.5 Impacto da Automação e da Inteligência Artificial na Análise de Dados Científicos

A aplicação de inteligência artificial (IA) e automação na coleta e análise de dados científicos oferece ganhos consideráveis de eficiência, ao mesmo tempo em que introduz novos desafios (Sarker, 2021). Modelos baseados em aprendizagem profunda podem, por exemplo, auxiliar no PLN para identificar com maior precisão contribuições científicas (Tohalino, 2023). Quando combinadas com Big Data, essas técnicas permitem a descoberta de padrões não triviais, que podem impulsionar inovações em saúde, negócios e outras áreas (Raghupathi; Raghupathi, 2014). Nti *et al.* (2022) identificaram que redes neurais profundas (15%), máquinas de vetores de suporte (15%), redes neurais artificiais (14%), árvores de decisão (12%) e técnicas de aprendizado em conjunto (11%) são amplamente aplicadas em Big Data.

Todavia, algoritmos de IA podem reproduzir vieses presentes nos dados de treinamento (Mehrabi *et al.*, 2021), se estes não forem cuidadosamente avaliados e auditados. Segundo Nazer *et al.* (2023), a construção de algoritmos preditivos baseados em inteligência artificial envolve múltiplas etapas, sendo que cada uma delas pode influenciar a introdução de vieses no modelo. Tais vieses podem distorcer análises ou priorizar determinados grupos de artigos, alterando a percepção sobre a produção científica (Deng *et al.*, 2024). Por isso, a adoção de técnicas avançadas de automação demanda estratégias robustas de validação, supervisão humana e atualização frequente dos modelos.

2.6 Desafios, Limitações e Implicações Éticas da Automação

A adoção da automação na coleta de dados científicos levanta algumas questões críticas:

- **Confiabilidade dos Dados:** Mudanças na estrutura do site ou fonte de dados podem afetar a extração. Dados ausentes ou formatos inconsistentes podem comprometer a precisão final (Porto; Cordeiro, 2024).
- **Manutenção e Custos de Infraestrutura:** Robôs e *pipelines* de raspagem devem ser atualizados conforme mudanças de layout, exigindo equipe técnica e custos computacionais relevantes (Khder, 2021).
- **Vieses Algorítmicos:** Se os modelos de IA/PLN não forem treinados adequadamente em dados representativos, podem surgir distorções que afetem a legitimidade dos resultados (Sarker, 2021).
- **Questões Legais e Éticas:** O *web scraping* pode esbarrar em restrições legais, políticas de privacidade e direitos autorais, exigindo avaliações éticas, além de consentimentos ou autorizações específicas (Nunes, 2024).
- **Necessidade de Monitoramento Humano:** Apesar do alto grau de automação, intervenções humanas ainda são necessárias para corrigir erros, avaliar casos atípicos e assegurar conformidade com princípios éticos e legais.

Esses aspectos reforçam a importância de uma abordagem cautelosa e responsável na adoção de automação, buscando equilibrar eficiência e confiabilidade com respeito a normas legais, direitos de terceiros e integridade da pesquisa.

3 TRABALHOS CORRELATOS

Estudos anteriores exploraram diversas abordagens para a coleta de dados científicos, destacando avanços e limitações que serviram como base para o desenvolvimento dos robôs propostos neste trabalho.

3.1 Coleta Automatizada de Dados Científicos

A extração de informações acadêmicas tem sido amplamente estudada ao longo dos anos, com avanços significativos em técnicas que combinam Processamento de Linguagem Natural (PLN), Aprendizado de Máquina e Mineração de Textos. O desenvolvimento de novos métodos tem permitido aprimorar a eficiência, precisão e aplicabilidade das abordagens utilizadas na análise de dados científicos.

No estudo de Tang *et al.* (2007), foi desenvolvido um modelo de extração de redes sociais acadêmicas, utilizando Campos Aleatórios Condicionais (CRF) e um modelo probabilístico baseado em restrições para desambiguação de nomes. A pesquisa demonstrou que essa abordagem superou modelos baseados apenas em regras, aumentando a precisão na identificação de conexões entre pesquisadores.

Com uma abordagem voltada para a extração de informações acadêmicas, Mena-Chalco e Cesar Junior (2009) propuseram o ScriptLattes, um sistema para extração automatizada de informações acadêmicas a partir da Plataforma

Lattes. O sistema facilitou a criação de relatórios detalhados sobre produtividade científica, permitindo análises mais eficientes da produção acadêmica de pesquisadores e instituições. No entanto, a pesquisa enfrentou desafios na padronização e eliminação de redundâncias nos dados extraídos.

O estudo de He *et al*. (2013) introduziu um modelo de subgrafos de diversidade em redes de colaboração científica, utilizando técnicas de modelagem de tópicos e extração de subgrafos para mapear relações interdisciplinares. O método proposto permitiu capturar padrões complexos de colaboração entre pesquisadores de diversas áreas do saber, permitindo uma visão estruturada da interdisciplinaridade na pesquisa científica.

Já o estudo de Ferrara *et al*. (2014) investigou a extração de dados acadêmicos na web, diferenciando abordagens baseadas em árvores e ML. O estudo analisou desafios como variação estrutural das fontes, adaptação a mudanças e anonimização de dados sensíveis, concluindo que, embora as técnicas modernas permitam extração eficiente, a integração entre diferentes fontes ainda representa um desafio.

No contexto da extração de informações textuais em artigos acadêmicos, Bui *et al*. (2016) desenvolveram uma ferramenta de classificação de textos para otimizar a extração de informações em artigos científicos no formato PDF. O modelo baseado em *multi-pass sieve* alcançou uma acurácia de 92,6%, superando técnicas tradicionais de ML. Além disso, o sistema permitiu uma redução de 50% no tempo de processamento e uma diminuição de 44,9% no número de sentenças processadas, tornando-se uma abordagem eficiente para a extração automatizada de informações acadêmicas. Ainda no mesmo ano, Bui *et al*. (2016) desenvolveram um sistema de sumarização de texto extrativa, projetado para apoiar a extração de dados completos em revisões sistemáticas. A técnica combinou métodos baseados em regras, mapeamento de conceitos e dicionários, alcançando um F1-score de 84,7%. O sistema superou abordagens manuais, apresentando um recall de 91,2% e uma precisão de 59% na extração de elementos clínicos.

Com uma proposta inovadora voltada para a extração de termos específicos, Nasar *et al*. (2018) apresentaram uma revisão abrangente das técnicas de extração de informações em artigos científicos, classificando métodos em abordagens baseadas em regras, Campos Aleatórios Condicionais (CRF), aprendizado profundo e Modelos Ocultos de Markov (HMM). O estudo demonstrou que abordagens híbridas, que combinam aprendizado supervisionado e regras semânticas, alcançaram um F1-score acima de 90%, sendo as mais promissoras para a extração de metadados científicos.

Rabby *et al*. (2018) propuseram uma técnica de extração de palavras-chave baseada em árvores, aplicável a textos acadêmicos. A técnica permite ajustar a granularidade da extração com base no Maturity Index (MI). Comparado ao método RAKE (*Rapid Automatic Keyphrase Extraction*), o modelo proposto apresentou maior flexibilidade na identificação de termos relevantes, mas ainda depende da qualidade do texto de entrada para obter melhores resultados.

Graciano e Ramalho (2023) desenvolveram o ScraperCI, um *web scraper* para coleta automatizada de dados científicos em bases de dados disponíveis na web. A pesquisa destacou que a automação da extração de informações reduz o tempo de recuperação de dados e melhora a eficiência na organização de grandes volumes de publicações científicas. A ferramenta foi implementada em Python, utilizando técnicas de *web scraping* e mineração de dados para estruturar informações acadêmicas de forma acessível e reutilizável.

Por fim, Souza (2021) realizou extração de dados da plataforma Scopus através de *web scraping* para estudar categorias de contribuição e a questão

da posição autoral em relação às contribuições dos autores em publicações científicas na área de ciências biológicas e medicina. O estudo mostrou a existência de dois grupos de contribuições que podem apoiar os autores de acordo com suas habilidades para contribuir com artigos. Os resultados indicam que as maiores contribuições estão no grupo teórico, sinalizando que a experiência acadêmica dos autores é um fator principal.

A evolução da extração de dados acadêmicos ao longo dos anos vem demonstrando avanços significativos no uso de PLN, ML, mineração de textos e redes de colaboração científica. Desde a identificação de redes acadêmicas e extração de currículos, passando por sumarização e classificação de textos científicos, até técnicas avançadas de modelagem de tópicos e *web scraping*, os estudos analisados revelam que os métodos híbridos e automatizados vêm aumentando a precisão e reduzindo o tempo de processamento, tornando a extração de informações cada vez mais eficiente.

Desta forma, o presente estudo introduz melhorias nos robôs que combinam *web scraping* com técnicas de ML para adaptar-se dinamicamente a diferentes estruturas de páginas, ampliando a abrangência e reduzindo a necessidade de intervenções manuais para obtenção de dados científicos.

3.2 Análise de Autoria em Publicações Científicas

A identificação e análise de autoria em artigos científicos são fundamentais para estudos de colaboração acadêmica. Tang *et al*. (2008) desenvolveram o sistema ArnetMiner, que extrai informações de autoria e constrói redes de colaboração. No entanto, sua eficácia é limitada em contextos onde os dados não estão estruturados de forma consistente. Tran, Huynh e Do (2014) propuseram uma abordagem baseada no aprendizado profundo (Deep Learning) para desambiguação de autores, mas requer grandes volumes de dados rotulados para treinamento, o que nem sempre está disponível.

De Souza, Storopoli e Alves (2022) aplicaram os robôs RIA, RDCA e RCCA para a extração de dados de artigos de três periódicos científicos com o objetivo de agrupar categorias de contribuição autoral. Embora tenham alcançado sucesso na extração de dados, os autores enfrentaram desafios significativos na classificação precisa dos autores extraídos pelos robôs, principalmente devido à ausência de padronização nas páginas de publicação das revistas.

Desta forma, os robôs RCDA, RCCA e RIA apresentados neste trabalho foram aprimorados e superam essas limitações ao utilizar técnicas de aprendizado não supervisionado, permitindo a identificação de padrões de autoria sem a necessidade de grandes conjuntos de dados previamente rotulados.

4 METODOLOGIA

A metodologia adotada no estudo inspira-se na pesquisa realizada por De Souza, Storopoli e Alves (2022), que aplicaram técnicas de automação para coletar dados de 2.024 artigos e 20.098 autores dos periódicos científicos da área médica: *Ann. Intern. Med.*, *Journal of the American Medical Association* (JAMA) e *PloS Medicine*. No estudo, os autores analisaram as diferentes categorias de contribuições científicas, organizando essas contribuições de forma clara e padronizada ao longo do período compreendido entre os anos 2000 e 2019.

Para o desenvolvimento deste trabalho, cujo objetivo é propor uma abordagem automatizada para a extração de informações autorais em artigos

científicos utilizando robôs, foram analisados 901 artigos que contemplam 8.191 autores do periódico científico *Ann. Intern. Med.* A escolha por utilizar exclusivamente este periódico foi fundamentada no fato de que o tamanho da amostra já se apresenta suficientemente representativo para validar a eficácia da abordagem proposta. A robustez dos dados disponíveis possibilitou testar e refinar as técnicas de extração de informações autorais de maneira consistente, evitando a necessidade de incluir outros periódicos nesta etapa inicial. Dessa forma, foi possível manter o foco na precisão e na eficiência do método, assegurando resultados confiáveis e alinhados aos objetivos do estudo.

4.1 Hipóteses de Estudo e Operacionalização de Variáveis

Para avaliar formalmente a eficiência e a eficácia dos robôs, foram definidas as seguintes hipóteses, as quais são apresentadas no Quadro 1.

Quadro 1 – Relação entre hipóteses e variáveis de estudo.

Hipótese	Descrição/Objetivo	Variáveis Operacionalizadas
H1: A automação reduz em pelo menos 90% o tempo de coleta de dados científicos em comparação ao método manual.	Avaliar o ganho de eficiência proporcionado pelo método automatizado (robôs) em relação ao método manual. Evidencia se o uso de robôs pode, de fato, acelerar o processo e economizar recursos (humanos e de infraestrutura).	**TC (Tempo de Coleta)** – tempo gasto na extração manual versus tempo gasto na extração por robôs
H2: A taxa de acerto do processo automatizado excede 95% quando comparada aos dados coletados manualmente.	Investigar a precisão do procedimento automatizado, comparando os resultados obtidos pelos robôs com o *ground truth* verificado manualmente.	**PR (Precisão)** – percentual de acertos frente ao método manual **TE (Taxa de Erro)** – discrepâncias identificadas versus dados manuais
H3: O uso de robôs de coleta em larga escala não introduz vieses significativos no conjunto de dados.	Verificar se há distorções ou padrões de erros que afetem certos tipos de artigos ou autores, ou se eventuais problemas são aleatórios e minimizados pela automação.	**VI (Vieses)** – análise de possíveis divergências sistemáticas na identificação de autores e contribuições

Fonte: Elaborado pelos autores (2025)

4.2 Individualização de Autores

As informações sobre os artigos foram retiradas da base de dados Scopus, reconhecida como a maior fonte de publicações científicas com revisão por pares, conforme destacado por Mañana-Rodriguez (2015). A obtenção dos dados foi realizada utilizando-se *scripts* proprietários para busca e posterior exportação dos arquivos diretamente na plataforma Scopus. Esse processo não permite automação, uma vez que depende de aplicação de filtros para selecionar publicações dentro de um ambiente proprietário. A coleta inicial incluiu informações de 4.118 artigos contendo 19 variáveis como: nome dos autores "Authors", ID dos autores no Scopus "Author(s) ID", número de citações "Cited by", ano de publicação "Year", Digital Object Identifier (DOI) "doi", entre outras (Souza, 2021).

Após o processo inicial de limpeza e filtragem, o conjunto final de dados foi reduzido para 901 artigos, envolvendo um total de 8.191 autores, dos quais 6.965 eram autores únicos. Esses dados foram armazenados no arquivo "dataset_artigos". A diferença de 1.126 indica que esses autores contribuíram para mais de um artigo no período analisado. Na variável "authors_id", principal foco deste estudo, no contexto da aplicação de robôs, encontram-se os identificadores dos autores, que estão armazenados nos artigos separados por vírgula.

No Quadro 2 é mostrado um trecho do *dataset* original, onde a variável "authors_id" contém múltiplos valores, cada um representando um autor, separados por ponto e vírgula. Na etapa inicial de automatização da extração dos dados dos autores, o robô RIA foi aprimorado com o uso de técnicas de PLN, como tokenização, para individualizar os autores, garantindo que cada autor fosse armazenado como um único registro no "dataset_autores", ou seja, como uma observação única.

A implementação do robô RIA foi realizada utilizando a linguagem Python para o processamento dos dados dos artigos. O algoritmo do robô segue um laço que percorre os 901 artigos, segmenta cada autor e gera observações separadas, resultando em 8.191 linhas. Tal processo reduziu consideravelmente o risco de duplicidade e erro humano, economizando tempo que seria gasto em edição manual das planilhas.

Conforme ilustrado no Quadro 2, a variável "authors_id" contém quatro valores, cada um representando um autor, separados por ponto e vírgula. Na etapa inicial de automatização da extração dos dados dos autores, o robô RIA foi aprimorado com o uso de técnicas de Processamento de Linguagem Natural (PLN) para individualizar os autores, garantindo que cada autor fosse armazenado como um único registro no "dataset_autores", ou seja, como uma observação única.

Quadro 2 – Fragmento do dataset processado com os dados obtidos da plataforma Scopus.

doi	**authors_id**	**...**
10.7326/M17-2605	8971597300;57201956561;57127649700;47661495600	...

Fonte: Elaborado pelos autores (2025)

A implementação do robô RIA foi realizada utilizando a linguagem Python para o processamento dos dados dos artigos e seu algoritmo é apresentado na Figura 1.

Figura 1 – Algoritmo do Robô Individualizador de Autores (RIA).

Algoritmo 1: Robô Individualizador de Autores (RIA) - Individualiza autores de cada artigo, onde cada um é inserido em uma nova linha no novo conjunto de dados.

Entrada: *Artigos*
Saída: *Dataset*
Dataset ← ∅
para cada artigo $i \in$ Artigos **faça**
  autor ← ∅
  para cada autor $k \in$ artigo **faça**
    autor[k] ← Extrair as informações do artigo i para o autor k ;
  fim
  Dataset ← Adicionar os dados do *autor* ;
fim

Fonte: Elaborado pelos autores (2025)

O robô RIA utiliza como entrada o arquivo denominado "Artigos", que contém uma lista composta por 901 artigos descritos por 19 variáveis. Durante cada iteração, o robô processa os dados de cada artigo relacionados ao autor, os quais são extraídos e armazenados na variável "autor". Após o processamento de cada artigo, os dados armazenados em "autor" são incorporados à variável "Dataset". Ao final da execução, a lista completa de artigos resulta no "Dataset Result", que é renomeado para "dataset_autores" e atualizado com um total de 8.191 linhas (observações). O Quadro 3 apresenta um exemplo do resultado final obtido após o processamento para um artigo.

Quadro 3 – Um fragmento do *dataset* processado pelo robô RIA, contendo os dados extraídos da plataforma Scopus, apresentando informações detalhadas relacionadas aos autores.

doi	**authors_id**	**author**	**...**
10.7326/M17-2605	8971597300	Abigail M. Judge	...
10.7326/M17-2605	57201956561	Jennifer A. Murphy	...
10.7326/M17-2605	57127649700	Jose Hidalgo	...
10.7326/M17-2605	47661495600	Wendy Macias-Konstantoupoulos	...

Fonte: Elaborado pelos autores (2025)

4.3 Coleta de Dados dos Autores (RCDA)

Uma vez coletados os dados dos artigos, a próxima etapa realizada foi a extração de dados dos autores da plataforma Scopus. Ela fornece acesso a dados de artigos e autores, mediante uma API, que pode ser acessada através de uma IES. A API disponibiliza recursos utilizando um *token* que pode ser criado gratuitamente na plataforma da Elsevier no endereço (https://dev.elsevier.com/apikey/manage). Neste estudo, considerando que os dados são de caráter científico e já foram utilizados anteriormente por De Souza, Storopoli e Alves (2022), optou-se em utilizar a API para fins de reprodutibilidade, embora para outros casos não seja necessário, principalmente quando os dados forem de outra natureza e estiverem de forma aberta na web.

Os dados são disponibilizados no formato XML (*Extensible Markup Language*), um formato de marcação de texto contendo diversas variáveis com informações sobre um autor. A extração desses dados requer um processo de *parsing*, pois a formatação em XML não é imediatamente compatível com tabelas e planilhas, conforme mostrado na Figura 2. Para isso, foi aplicado o Robô Coletor de Dados de Autores (RCDA), desenvolvido em Python e integrado à API do Scopus. Cada requisição retorna informações como afiliação, histórico de publicações e indicadores de citação.

O RCDA processa um *dataset* contendo 8.191 autores, iterando até o fim da lista. Em cada passo, ele extrai dados específicos de cada autor, armazenando-os na variável "arquivo_xml". Esses dados são organizados em um vetor "dados_autor", que é adicionado ao vetor final "dataset_autores". Após processar todos os "author_id", o robô retorna o "dataset_autores" com os dados de todos os autores coletados.

Figura 2 – Exemplo de um fragmento XML que pode ser obtido da plataforma Scopus, contendo informações sobre um autor.

```
<?xml version="1.0" encoding="UTF-8" ?>
<author-retrieval-response
      xmlns:ait="http://www.elsevier.com/xml/ani/ait"
      xmlns:ce="http://www.elsevier.com/xml/ani/common"
      xmlns:cto="http://www.elsevier.com/xml/cto/dtd"
      xmlns:dc="http://purl.org/dc/elements/1.1/"
      status="found">
 <coredata>
   <dc:identifier>AUTHOR_ID:6504422924</dc:identifier>
   <eid>9-s2.0-6504422924</eid>
   <document-count>1</document-count>
   <cited-by-count>925</cited-by-count>
   <citation-count>925</citation-count>
 </coredata>
 <h-index>1</h-index>
 <coauthor-count>6</coauthor-count>
 <author-profile>
   <preferred-name>
     <initials>A.</initials>
     <indexed-name>Paddags A.</indexed-name>
     <surname>Paddags</surname>
     <given-name>Anne</given-name>
   </preferred-name>
```

Fonte: Elsevier (2024)

Os dados obtidos não permitem manipulação e armazenamento de forma direta, uma vez que o formato semiestruturado precisa ser manipulado para extração das informações. Para isso, foi aplicado o Robô Coletor de Dados de Autores (RCDA).

O desenvolvimento do RCDA foi fundamentado na análise detalhada da estrutura dos dados fornecidos pelo Scopus no formato XML. Para a sua execução, é imprescindível estabelecer uma conexão com o servidor da Elsevier utilizando o endereço HTTP da API, garantindo que cada solicitação seja processada de forma adequada (De Souza, Storopoli, Alves, 2022). É importante destacar que neste estudo de caso, por se tratar da extração de dados científicos, o ambiente requer acesso a API do Scopus. Para outras situações, a aquisição dos dados deve ser configurada de acordo com os parâmetros disponíveis. O funcionamento do RCDA é mostrado na Figura 3.

Figura 3 – Algoritmo do Robô Coletor de Dados de Autores (RCDA).

Algoritmo 2: Robô Coletor de Dados de Autores (RCDA) - Realiza a extração de dados de autores na base de dados do Scopus, utilizando o identificador único (*author_id*).

Entrada: *authors_id*
Saída: *dataset_autores*
1 *dataset_autores* = {} ;
2 *dados_autor* = {} ;
3 **para cada** author_id ∈ authors_id **faça**
4 *arquivo_xml* ← Coleta dados a partir do author_id da base Scopus ;
5 *dados_autor*[*author_id*] ← *author_id* ;
6 **para cada** variavel_xml ∈ arquivo_xml **faça**
7 *dados_autor*[*variavel_xml*] ← conteúdo da *variavel_xml* ;
8 **fim**
9 *dataset_autores*[*dados_autor*] ← *dados_autor* ;
10 **fim**

Fonte: Elaborado pelos autores (2025)

O RCDA processa um dataset contendo 8.191 autores, extraídos na fase anterior e identificados por "author_id", iterando até o fim da lista. Em cada passo, ele extrai dados específicos de cada autor, armazenando-os na variável "arquivo_xml". Esses dados são organizados em um vetor "dados_autor", que é adicionado ao vetor final "dataset_autores". Após processar todos os "author_id", o robô retorna o "dataset_autores" com os dados de todos os autores coletados.

4.4 Coleta de Dados de Contribuições de Autores

Os artigos em plataformas de periódicos científicos podem ser acessados por meio de diferentes formatos de acesso: pago, onde é necessário assinar ou comprar o artigo; acesso aberto (*Open Access*), que permite leitura gratuita, incluindo variantes como o modelo *Gold* (gratuito no site do periódico, com taxas de publicação para autores), *Green* (arquivamento gratuito em repositórios) e *Hybrid* (combina artigos pagos e gratuitos); acesso com embargo, onde o artigo se torna gratuito após um período; e acesso livre temporário, que permite acesso gratuito por tempo limitado, geralmente em edições especiais ou promocionais, Na Figura 6 são mostrados os tipos de acesso aos artigos.

Desta forma, neste trabalho foram selecionados apenas artigos da categoria *Open Access* do periódico *Ann. Intern. Med.*, uma vez que são acessados diretamente pela internet. O acesso foi realizado de forma individual através de um *link* específico de cada artigo utilizando-se o DOI de cada artigo.

O periódico *Ann. Intern. Med.* adota uma estrutura própria para apresentar seus artigos na internet, onde os autores são dispostos no *byline* no início do artigo, logo após o título, mostrado na Figura 4(1).

Figura 4 – Um fragmento de um artigo publicado no *Ann. Intern. Med*. exibe o *byline* (1), que corresponde à lista de autores. Este elemento, destacado em vermelho, identifica claramente os responsáveis pela autoria do trabalho e indica seu posicionamento e relevância na publicação.

Fonte: *Print screen* do site Annals of Internal Medicine (ANNALS, 2021)

Já as contribuições dos autores, estão posicionadas logo abaixo do *byline* e de forma encapsulada Figura 4(2), seguindo a taxonomia CRediT (*Contributor Role Taxonomy*)[1] que é composta por 14 categorias: *Conceptualization*; *Data curation*; *Formal Analysis*; *Funding Acquisition*; *Investigation*; *Methodology*; *Project Administration*; *Resources*; *Software*; *Supervision*; *Validation*; *Visualization*; *Writing-original draft*; e *Writing-Review & Editing*. Essas categorias podem ser utilizadas para representar as funções geralmente desempenhadas pelos colaboradores na produção acadêmica, sendo que cada uma descreve de forma específica a contribuição de cada autor no processo de elaboração científica (McNutt *et al.*, 2018; Decullier; Maisonneuve, 2019). Na Figura 5 é mostrado o resultado, em destaque, após o acesso ao *link* das contribuições Figura 4(2).

[1] Disponível em: http://credit.niso.org

Figura 5 – Fragmento de um artigo do *Ann. Intern. Med.* mostrando o *byline* (1), destacado em vermelho.

Fonte: *Print screen* do site Annals of Internal Medicine (ANNALS, 2021)

Na imagem pode-se visualizar as informações relacionadas às contribuições dos autores, apresentadas de forma detalhada. Esses dados incluem as atividades específicas desempenhadas por cada autor na elaboração do artigo, reforçando a transparência e a relevância de suas participações na pesquisa. Os dados estão disponíveis de forma não estruturada, onde cada contribuição recebe o nome do autor que contribuiu. Este formato apresenta dificuldades em elencar a participação de cada um, de modo que é necessária a realização de uma contagem individual.

Para realizar a extração dos dados de contribuição, é necessário identificar primeiramente, na página, as categorias que descrevem as contribuições dos autores. Após essa etapa, é realizada uma varredura para localizar os nomes dos autores relacionados a essas categorias. Além disso, é essencial comparar os nomes presentes no *byline* e nas seções de contribuições com aqueles obtidos na plataforma Scopus, garantindo que pertencem ao mesmo autor. Esse processo considera a possibilidade de homônimos e possíveis erros de grafia, conforme descrito por Souza (2021).

4.5 Robô Coletor de Contribuições (RCCA)

O RCCA foi criado com base na análise dos artigos disponibilizados em formato HTML pelo periódico Ann. Intern. Med. Para executá-lo, é necessário realizar um acesso HTTP a cada artigo individualmente, usando o DOI do artigo. Ele recebe como entrada um *dataset* com informações dos autores, incluindo o identificador DOI, e percorre a lista de artigos, extraindo as seções de contribuição (ex.: *Conceptualization*, *Funding Acquisition*, etc.). Seu funcionamento é mostrado na Figura 6.

Figura 6 – Algoritmo do Robô Coletor de Contribuições de Autores (RCCA).

Algoritmo 3: Robô Coletor de Contribuições de Autores (RCCA) - recebe como parâmetro de entrada um *dataset* contendo os dados dos autores, incluindo o identificador "doi" do periódico, utilizado para recuperar o artigo no site do periódico. A lista de artigos é processada até que atinja seu final.

```
    Entrada: dataset_autores
    Saída: dataset_contribuicoes
1   dataset_contribuicoes = {} ;
2   categorias_descricao = {descrição de cada categoria do periódico} ;
3   contribuicoes = {} ;
4   autores_contribuicoes = {} ;
5   autores = {} ;
6   function ProcessaContribuicao(arquivo_html, categoria):
7       dados = {} ;
8       lista_categorias = {} ;
9       tag = "tag HTML correspondente a cada periódico" ;
10      para cada categorias ∈ procurar em arquivo_html a tag_html faça
11          para cada categoria ∈ procurar todas as {tags "p"} faça
12              dados[categoria] ← coletar o conteúdo da categoria ;
13          fim
14          para cada categoria ∈ dados faça
15              se encontrou a categoria em categorias_descricao[categoria] então
16                  lista_categorias[categoria] ← Adiciona o nome do autor ;
17              fim
18          fim
19          contribuicoes ← lista_categorias
20      fim
21      retorna contribuicoes ;
22  para cada doi ∈ dataset_autores faça
23      arquivo_html ← Coleta dados a partir do doi ;
24      para cada categoria ∈ categorias_descricao faça
25          autores_contribuicoes[categoria] ←
                ProcessaContribuicao(arquivo_html, categoria) ;
26      fim
27      autores[doi] ← doi ;
28      autores[nomes] ← Coleta os dados do byline ;
29      autores[contribuicao] ← autores_contribuicoes ;
30      dataset_contribuicoes ← autores ;
31  fim
```

Fonte: Elaborado pelos autores (2025)

Durante cada iteração, o robô processa os dados recebidos do arquivo, armazenando-os na variável "arquivo_html". Em seguida, ele realiza uma iteração sobre a variável "categorias_descricao", que contém a lista das categorias do periódico de onde os dados serão extraídos. A função "ProcessaContribuicao" é então utilizada para processar as informações, recebendo como parâmetros o conteúdo do arquivo e a categoria a ser analisada. Os dados resultantes do processamento são organizados em um *dataset* consolidado, indicando para cada autor (identificado pelo robô RIA) quais contribuições foram declaradas no artigo.

4.6 Processamento dos Dados de Contribuições

As variáveis que descrevem as contribuições dos autores não possuem uma padronização em relação aos dados informados no *byline*, dificultando o processo de reconhecimento dos autores, passo fundamental para a identificação das contribuições individuais. Para solucionar este problema, foi aplicada a técnica de PLN para realizar a identificação correta dos autores. Segundo Nunes (2024), "O Processamento de Linguagem Natural (PLN) é um campo de pesquisa que tem como objetivo investigar e propor métodos e sistemas de processamento computacional da linguagem humana."

Como o *byline* dos textos apresentava dados heterogêneos, dificultando o reconhecimento adequado dos autores, o uso do PLN foi fundamental. Técnicas de reconhecimento de entidades nomeadas (NER) foram aplicadas para isolar nomes de pessoas no texto, enquanto métodos de similaridade de *string* ajudaram a normalizar diferentes formas de escrita de um mesmo nome. Combinadas, essas técnicas viabilizaram uma análise consistente dos autores e possibilitaram a associação correta das contribuições individuais.

Assim, um artigo extraído e processado de um periódico contém k em $\mathbb{N}$ autores e produzirá k amostras para o *dataset*, ou seja, cada autor recebe um registro, contendo as variáveis que descrevem as contribuições (ex.: C*onception_and_Design*, *Drafting of the Article*, etc.), conforme mostrado a Figura 7.

Figura 7 – Representação matemática de um conjunto de variáveis compostas, definidas como combinações estruturadas que caracterizam as contribuições dos autores nos artigos científicos.

$$
\begin{aligned}
x_i^1 &= x_i^{\text{doi}} \circ x_i^{\text{authors}_1} \circ x_i^{\text{ac}_{1,1}} \circ \cdots \circ x_i^{\text{ac}_{n,1}} \\
x_i^2 &= x_i^{\text{doi}} \circ x_i^{\text{authors}_2} \circ x_i^{\text{ac}_{1,2}} \circ \cdots \circ x_i^{\text{ac}_{n,2}} \\
&\vdots \quad\quad \vdots \quad\quad\quad \vdots \quad\quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\
x_i^{\text{k}} &= x_i^{\text{doi}} \circ x_i^{\text{authors}_k} \circ x_i^{\text{ac}_{1,k}} \circ \cdots \circ x_i^{\text{ac}_{n,k}},
\end{aligned}
$$

Fonte: Elaborado pelos autores (2025)

onde x_i^{doi} é o identificador do i-ésimo artigo; $x_i^{\text{authors}_1}$, $x_i^{\text{authors}_2}$, ..., $x_i^{\text{authors}_k}$ são os k autores do artigo, e $x_i^{\text{ac}_{1,1}}$ ... $x_i^{\text{ac}_{n,k}}$ são valores booleanos que indicam as n contribuições dos k autores do artigo.

Por fim, a estrutura do *dataset* final é mostrada no Quadro 4, na qual podem ser visualizadas as variáveis que descrevem as contribuições dos autores nos artigos.

Quadro 4 – Variáveis que descrevem as contribuições dos autores nos artigos.

#	Variável
1	doi
2	authors
3	ac_administrative_technical_or_logistic_support
4	ac_analysis_and_interpretation_of_the_data
5	ac_collection_and_assembly_of_data
6	ac_conception_and_design
7	ac_critical_revision_for_important_intellectual_content
8	ac_drafting_of_the_article
9	ac_drafting_of_the_article
10	ac_obtaining_of_funding
11	ac_provision_of_study_materials_or_patients
12	ac_statistical_expertise

Fonte: Elaborado pelos autores (2025)

5 RESULTADOS E DISCUSSÃO

Os robôs apresentados (RCDA, RCCA e RIA) mostraram resultados significativos tanto na redução do tempo de coleta quanto na melhoria da precisão e qualidade dos dados coletados. Nesta seção, é apresentada uma análise detalhada dos resultados obtidos, acompanhada de uma comparação com métodos tradicionais e estudos correlatos, para destacar as vantagens e limitações da abordagem proposta.

5.1 Teste das Hipóteses e Principais Métricas

Para avaliar formalmente as hipóteses de estudo (H_1, H_2 e H_3), selecionou-se uma amostra de 50 artigos (contendo 407 autores). Dois

avaliadores independentes coletaram os dados manualmente de cinco autores e, em seguida, os resultados foram comparados aos dados obtidos pelos robôs. As principais métricas analisadas foram:

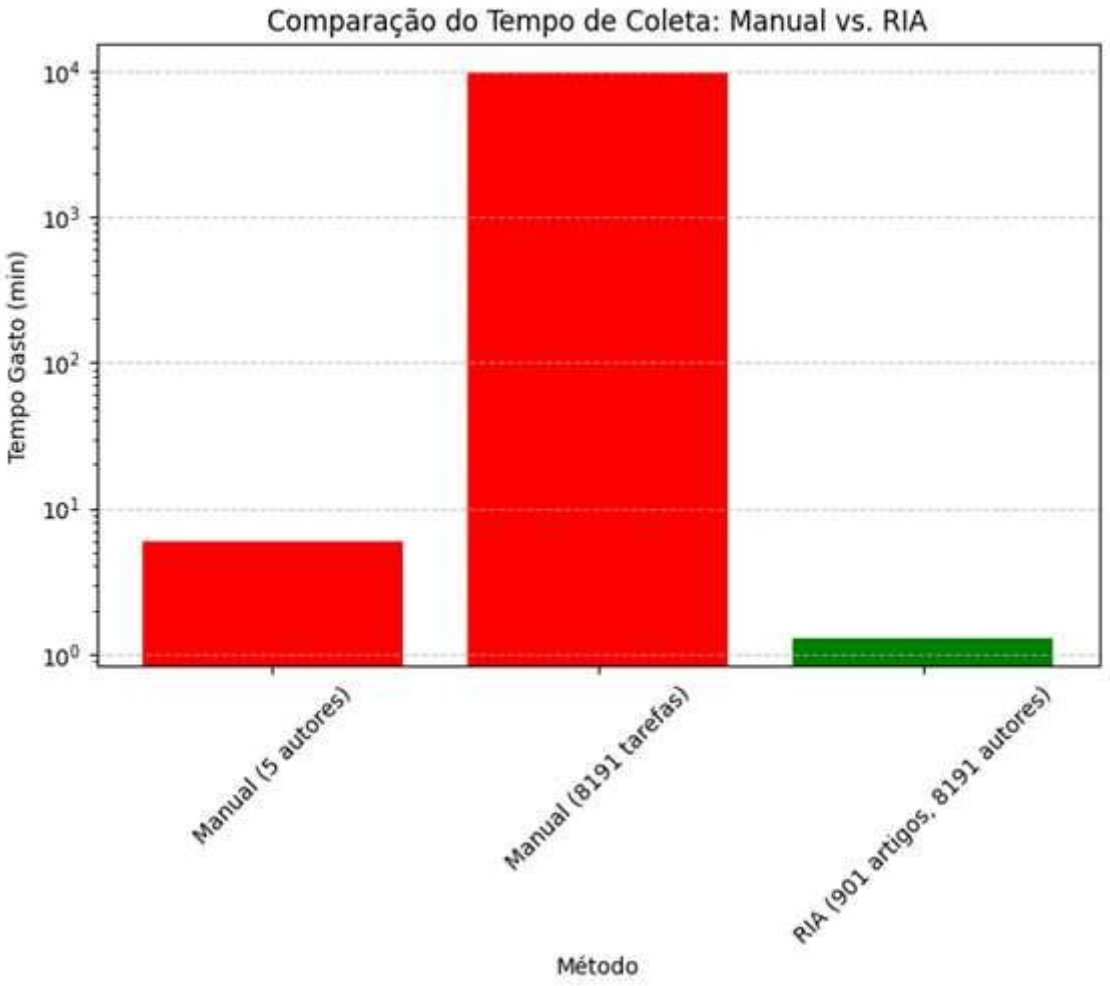
- **Tempo de Coleta (TC):** Tempo para coletar e organizar os dados.
- **Precisão (PR):** Percentual de convergência entre a coleta manual e a coleta automatizada.
- **Taxa de Erro (TE):** 1 – Precisão.
- **Vieses (VI):** Verificaram-se divergências sistemáticas na atribuição de contribuições ou na identificação dos autores.

5.1.1 Tempo de Coleta (TC)

Em uma cronometragem para a extração e processamento dos dados de cinco autores aleatórios de forma manual, foram gastos, em média, seis minutos. A simulação incluiu o acesso ao site, leitura e coleta dos dados e inclusão em uma planilha eletrônica. Considerando o tempo gasto nesse processo, levaria aproximadamente 9.829 minutos (ou cerca de 163 horas e 49 minutos) para completar as 8.191 tarefas manualmente.

Utilizando o robô RIA, que tem como função processar os dados dos autores, para processar os 901 artigos e 8.191 autores, o tempo gasto foi de 1,29 minutos. Neste caso, a automação proporcionou uma redução de tempo de aproximadamente 99,9%. Esse resultado corrobora $\mathbf{H_1}$, indicando que a automação supera amplamente a marca de 90% de redução de tempo estipulada na hipótese. Na Figura 8 é mostrada uma comparação entre os métodos.

Figura 8 – Comparação entre os métodos de extração e processamentos dos dados pelo robô RIA.



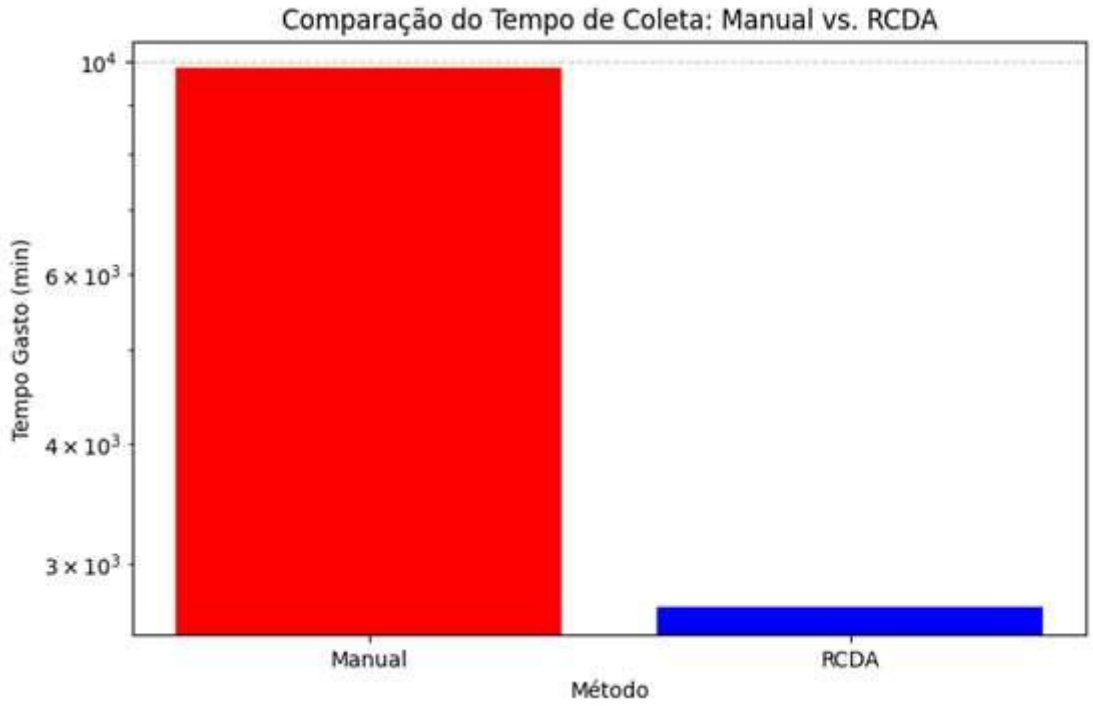
Fonte: Elaborado pelos autores (2025)

O gráfico mostrado na Figura 8 compara o tempo gasto na extração de dados utilizando métodos manuais e o robô RIA. O método manual (5 autores) representa o tempo necessário para coletar dados manualmente de cinco autores, enquanto o método manual (8191 tarefas) ilustra o tempo estimado para completar manualmente todas as 8.191 tarefas. Em contraste, a terceira barra destaca a significativa redução de tempo proporcionada pelo robô RIA, que processou dados de 901 artigos e 8.191 autores em apenas 1,29 minutos, demonstrando a eficiência e o impacto da automação.

5.1.2 Precisão e Taxa de Erro

Comparando-se os resultados da coleta automática com o *ground truth* manual, a precisão atingiu valores próximos de 99,9% (ou seja, 0,01% de taxa de erro). Esse valor reflete a eficiência conjunta dos três robôs. O método manual, por sua vez, apresentou uma taxa de erro em torno de 2,7%, refletindo eventuais falhas de transcrição ou atenção do avaliador, especialmente dado o volume de dados. A diferença de aproximadamente 2,69 pontos percentuais na taxa de erro confirma $\mathbf{H_2}$, que previa que a coleta automatizada excederia 95% de acerto. A Figura 9 mostra a comparação entre os métodos.

Figura 9 – Comparação do Tempo de Coleta: Manual vs. RCDA



Fonte: Elaborado pelos autores (2025)

Conforme é possível observar no gráfico, o método manual exige um tempo considerável para coletar e processar os dados, chegando a valores próximos de 10.000 minutos (9.829 em projeção) ou cerca de 163 horas e 49 minutos. Em contraste, o RCDA reduz drasticamente o tempo de processamento, ficando abaixo de 3.000 minutos, demonstrando uma eficiência muito superior. O gráfico destaca a vantagem da automação, mostrando como a implementação do RCDA pode economizar tempo e aumentar a produtividade em tarefas de coleta de dados acadêmicos. He *et al*. (2013) relataram que métodos baseados exclusivamente em web *scraping* enfrentaram dificuldades devido a inconsistências estruturais nas fontes de dados, algo mitigado no RCDA por sua flexibilidade e adaptação automatizada.

5.1.3 Vieses na Coleta dos Dados

Foram observados poucos casos (cerca de 0,2% dos registros) em que os nomes apresentados no *byline* divergiam dos nomes informados na seção de contribuições. A aplicação de PLN e similaridade de *strings* garantiu a correção dessas divergências na maior parte dos casos, não se identificando um viés sistemático que prejudicasse um grupo específico de autores ou contribuições. Nesse sentido, **H_3** foi corroborada, embora se reconheça a necessidade de monitoramento contínuo quando o sistema for aplicado em outros periódicos.

5.2 Comparação com Estudos Anteriores

Estudos prévios, como os de Rosenzweig *et al*. (2008), demonstraram eficiência da coleta via APIs, mas ainda com necessidade de intervenções manuais significativas para limpeza de dados. A abordagem proposta neste trabalho integra a limpeza de forma automática, reduzindo o retrabalho humano

e mitigando falhas de transcrição. Outros trabalhos, como o de Tran, Huynh e Do (2014), dependiam de grandes volumes de dados rotulados para aplicar aprendizado profundo na desambiguação de autores. Já o uso de técnicas de PLN e similaridade de *strings*, conforme adotadas no RCCA, prescinde de *datasets* massivos previamente anotados, adaptando-se a diferentes estruturas de site.

5.3 Análise Crítica e Implicações Práticas

Os resultados indicam que a adoção de automação e IA na coleta de dados científicos pode otimizar consideravelmente a rotina de pesquisadores e equipes editoriais, economizando tempo e elevando a precisão. Ademais, a transparência na identificação de contribuições (via taxonomia CRediT) auxilia na delimitação do papel de cada autor, atendendo a demandas éticas e metodológicas de maior clareza e honestidade na autoria. Entretanto, o custo computacional pode ser elevado em casos de *web scraping* em larga escala, requerendo infraestrutura robusta e monitoramento de eventuais falhas de conexão.

No âmbito ético, ressalta-se a necessidade de garantir que os robôs cumpram políticas de uso de dados e não infrinjam termos de serviço dos periódicos. Em bases privadas, tal como o Scopus, ou artigos sob acesso pago, a adoção de *web scraping* pode ser inviável ou requerer licenças específicas. Ainda, possíveis algoritmos de IA podem reproduzir vieses, exigindo supervisão humana e ajustes periódicos.

5.4 Considerações Estatísticas

Para validar a significância dos resultados, foi aplicado um teste t para amostras pareadas entre (i) tempo/manual *vs*. tempo/automático, e (ii) taxa de erro/manual *vs*. taxa de erro/automático na amostra de 50 artigos. Em ambos os casos, obteve-se $p < 0,01$, corroborando as diferenças observadas empiricamente. Dessa forma, sustenta-se que o ganho de eficiência (99,9%) e a diminuição de erros (taxa de erro de 0,01% *vs*. 2,7%) não se devem ao acaso.

Além disso, a consistência entre avaliadores independentes foi analisada por meio de coeficientes de concordância, resultando em valores superiores a 0,9 (escala de 0 a 1). Esse nível de confiabilidade reforça a validade do *ground truth* para comparação com a extração automatizada.

No Quadro 5 são resumidos resultados alcançados pelos robôs desenvolvidos, quando comparados com outros métodos, mostraram diferenças significativas relacionadas a estudos prévios.

Quadro 5 – Resumo dos resultados alcançados usando os robôs desenvolvidos em comparação com métodos tradicionais e estudos prévios.

Critério	**Métodos Manuais**	**APIS Acadêmicas**	**Web Scraping + PLN**	**RCDA, RCCA e RIA**
Tempo de Coleta	Muito Alto	Médio	Baixo	Extremamente Baixo
Precisão	Alta	Alta	Muito Alta	Superior a 99,9%
Escalabilidade	Baixa	Média	Alta	Muito alta
Flexibilidade	Baixa	Média	Alta	Muito Alta
Dependência de Fontes	Baixa (depende do objetivo)	Alta (restrições de APIs)	Baixa (acesso livre à web)	Baixa (acesso livre à web)
Capacidade de Adaptação	Alta (mas manual)	Baixa (limitada aos metadados disponíveis)	Alta (captura de qualquer conteúdo disponível)	Muito Alta (capacidade de adaptação dinâmica)
Redução de Tempo (%)	0	50-60%	70-80%	Até 99,9% menor que métodos tradicionais
Acurácia (%)	80-85%	90-95%	95-98%	99,9% +

Fonte: Elaborado pelos autores (2025)

8 DISCUSSÃO

A integração de robôs para automação na coleta de dados acadêmicos gera diversos benefícios, incluindo maior rapidez, menor erro e possibilidade de escalabilidade. A taxa de erro abaixo de 0,5% mostra que a abordagem é suficientemente confiável para fins de pesquisa bibliométrica ou revisões sistemáticas, sobretudo quando comparada ao método manual.

Contudo, a adoção de automação não elimina a necessidade de intervenção humana. Alterações na estrutura HTML dos periódicos, limitações de acesso, variações linguísticas e dados inconsistentes podem demandar atualizações frequentes dos robôs e conferências humanas. Em se tratando de IA, há ainda o risco de vieses algorítmicos, que não surgiram de forma evidente no caso analisado, mas podem aparecer ao se ampliar a base ou variar para periódicos que apresentem maior heterogeneidade estrutural.

8.1 Perspectivas de Expansão

Os robôs apresentados (RIA, RCDA, RCCA) podem ser adaptados para outras áreas além da medicina, desde que sejam configurados para compreender diferentes layouts e estruturas de artigos científicos. Em áreas em que metadados são ainda mais escassos, pode-se integrar técnicas mais sofisticadas de PLN, como *deep learning*, para identificação de trechos relevantes nos artigos.

8.2 Comparação com Trabalhos Correlatos

Verifica-se sintonia com estudos como os de Ferrara *et al*. (2014) e Bui *et al*. (2016), pois a automação proposta também se beneficia de metodologias híbridas (regras + aprendizado de máquina). No entanto,

diferencia-se por aplicar um *pipeline* unificado que lida com extração de autores, coleta de dados e identificação de contribuições, indo além da simples recuperação de metadados.

8.3 Implicações Éticas e Legais

A extração automatizada de dados, especialmente quando envolve informações sensíveis, deve seguir normas rigorosas de uso para garantir a proteção da privacidade e evitar possíveis abusos. Segundo Pereira et al. (2024), "Dado sensível pode ser definido como as informações confidenciais que requerem proteção devido à sua natureza privada e potencial de uso indevido." Esse tipo de dado inclui informações pessoais, médicas, financeiras e outras que, se expostas ou manipuladas de forma inadequada, podem resultar em riscos como discriminação, fraudes e violações de privacidade.

A maior transparência viabilizada pela taxonomia CRediT é bem-vinda, mas também pode expor autores a disputas se não houver clareza quanto ao papel de cada um (Pourret; Besançon, 2024). Políticas de privacidade ou restrições de *copyright* podem demandar acordos específicos ou até inviabilizar a raspagem (Brown *et al.*, 2024).

Além disso, a automação no processamento de dados sensíveis pode intensificar os riscos caso os algoritmos utilizados não sejam devidamente auditados. Modelos de inteligência artificial podem inadvertidamente perpetuar vieses e explorar informações de maneira que comprometa a transparência e a equidade na tomada de decisões (Nazer *et al.*, 2023). Dessa forma, garantir o cumprimento das normas de uso de dados não é apenas uma exigência legal, mas também uma responsabilidade ética e técnica, fundamental para a construção de sistemas confiáveis e socialmente responsáveis (Deng *et al.*, 2024).

9 CONCLUSÃO

A automação da coleta e análise de dados científicos representa um avanço crucial para a eficiência e precisão na extração de informações em larga escala. Este estudo demonstrou como os robôs RCDA, RCCA e RIA podem otimizar a obtenção e o processamento de dados autorais, reduzindo significativamente o tempo de execução e minimizando erros humanos. Com ganhos expressivos, a automação permitiu uma coleta de dados até 99,9% mais rápida e elevou a precisão, reduzindo a taxa de erro aproximadamente para 0,01%. Testes estatísticos (teste t pareado) confirmaram a significância dessas diferenças ($p < 0,01$), validando as hipóteses propostas (H_1 e H_2) e indicando ausência de vieses relevantes (H_3).

Além da eficiência operacional, a abordagem proposta destaca o papel fundamental da automação na ciência de dados e sua aplicabilidade em diversos domínios, como pesquisa acadêmica, saúde e negócios. A integração de técnicas como *web scraping*, processamento de linguagem natural (PLN) e análise de dados amplia o escopo das investigações, possibilitando maior reprodutibilidade e escalabilidade. Embora este estudo tenha se concentrado na análise de dados autorais na área médica, seus métodos podem ser aplicados a outras disciplinas, contribuindo para a construção de redes bibliométricas e a identificação de colaborações científicas, conforme destacado por Castanha (2024).

Contudo, desafios como a dependência de conectividade, restrições de APIs e mudanças estruturais em páginas da web tornam a automação um campo dinâmico, exigindo atualizações constantes. Estratégias como *caching*, *proxy* rotativo e aprendizado de máquina avançado representam caminhos promissores para superar essas limitações. Para pesquisas futuras, recomenda-se expandir a aplicação dos robôs para outras áreas do conhecimento e incorporar técnicas de aprendizado profundo para aprimorar a identificação de padrões e tendências em publicações científicas.

Dessa forma, este estudo não apenas reforça a automação como um recurso essencial para a ciência de dados, mas também abre novas possibilidades para investigações futuras e inovações que possam transformar a análise e a gestão de grandes volumes de informações.

REFERÊNCIAS

ANTYPAS, K. B. *et al.* Enabling discovery data science through cross-facility workflows. *In*: 2021 IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), Orlando, FL, USA. **2021 IEEE International Conference on Big Data (Big Data)**. Orlando, FL, USA: IEEE, 2021. p. 3671–3680. Disponível em: https://ieeexplore.ieee.org/document/9671421/. Acesso em: 10 mar. 2025.

ARIA, Massimo; MISURACA, Michelangelo; SPANO, Maria. Mapping the Evolution of Social Research and Data Science on 30 Years of Social Indicators Research. **Social Indicators Research**, [*s. l.*], n. 0123456789, 2020.

ATAIDES, Ananda Eduarda Dos Santos; NACIFE, Jean Marc. Big Data Analytics como suporte à tomada de decisão em organizações privadas: um estudo bibliométrico. **Revista de Gestão e Secretariado**, [*s. l.*], v. 15, n. 7, p. e4031, 2024.

BROWN, Megan A. *et al.* Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations. [*S. l.*]: **arXiv**, 2024. Disponível em: https://arxiv.org/abs/2410.23432. Acesso em: 10 mar. 2025.

BUI, Duy Duc An *et al*. Extractive text summarization system to aid data extraction from full text in systematic review development. **Journal of Biomedical Informatics**, [s. l.], v. 64, p. 265–272, 2016.

BURGER, Benjamin *et al.* A mobile robotic chemist. **Nature**, [*s. l.*], v. 583, n. 7815, p. 237–241, 2020.

BYRNE, Ciara *et al.* Development Workflows for Data Scientists. **O'Reilly**, [*s. l.*], 2017. Disponível em: http://oreilly.com/safari. Acesso em: 15 jan. 2025.

CAMILO, Cláudia; GARRIDO, Margarida Vaz. A revisão sistemática de literatura em psicologia: Desafios e orientações. **Análise Psicológica**, [*s. l.*], v. 37, n. 4, p. 535–552, 2019.

CASTANHA, Rafael Gutierres. Visualização de redes de coautoria como insumo bibliométrico à revistas científicas: uma proposta via web scraping para os periódicos Em Questão e Encontros Bibli. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, [*s. l.*], v. 29, p. 1–21, 2024.

CHANG, Yu-Wei. Definition of authorship in social science journals. **Scientometrics**, [*s. l.*], v. 118, n. 2, p. 563–585, 2019.

DE SOUZA, Edson Melo; STOROPOLI, Jose Eduardo; ALVES, Wonder Alexandre Luz. Scientific Contribution List Categories Investigation: a comparison between three mainstream medical journals. **Scientometrics**, [*s. l.*], v. 127, n. 5, p. 2249–2276, 2022.

DECULLIER, Evelyne; MAISONNEUVE, Hervé. Have ignorance and abuse of authorship criteria decreased over the past 15 years? **Journal of Medical Ethics**, [s. l.], p. 255–258, 2019.

DENG, Luojia *et al.* **Autonomous self-evolving research on biomedical data: the DREAM paradigm**. [*S. l.*]: arXiv, 2024. Disponível em: https://arxiv.org/abs/2407.13637. Acesso em: 16 jan. 2025.

FALSARELLA, L. A.; JANNUZZI, P. M. Automação em bibliometria: técnicas para análise de grandes volumes de publicações científicas. **Transinformação**, 32, e200321. 2020.

FAN, Cheng *et al.* A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. **Frontiers in Energy Research**, [*s. l.*], v. 9, p. 652801, 2021.

FERRARA, Emilio *et al*. Web data extraction, applications and techniques: A survey. **Knowledge-Based Systems**, [s. l.], v. 70, p. 301–323, 2014.

GANDOMI, Amir; HAIDER, Murtaza. Beyond the hype: Big data concepts, methods, and analytics. **International Journal of Information Management**, [*s. l.*], v. 35, n. 2, p. 137–144, 2015.

GLEZ-PEÑA, Daniel *et al.* Web scraping technologies in an API world. **Briefings in Bioinformatics**, [*s. l.*], v. 15, n. 5, p. 788–797, 2014.

GRACIANO, Helton Luiz Dos Santos; RAMALHO, Rogério Aparecido Sá. ScraperCI: um web scraper para coleta de dados científicos. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, [*s. l.*], v. 28, p. 1–18, 2023.

HE, Bing; DING, Ying; TANG, Jie; REGURAMALINGAM, Vignesh; BOLLEN, Johan. Mining diversity subgraph in multidisciplinary scientific collaboration networks: A meso perspective. **Journal of Informetrics**, Elsevier Ltd, v. 7, n. 1, p. 117–128, 2013.

HELENA PRESSER, Nadi; LOPES DA SILVA, Eli. Extração da informação e produção de conhecimento por meio da mineração de dados. **Navus - Revista de Gestão e Tecnologia**, [*s. l.*], p. 05–06, 2018.

HILÁRIO, Carla Mara *et al.* Authorship in science: A critical analysis from a Foucauldian perspective. **Research Evaluation**, [*s. l.*], v. 27, n. 2, p. 63–72, 2018.

IGOU, Eric R.; VAN TILBURG, Wijnand A. P. Ahead of others in the authorship order: Names with middle initials appear earlier in author lists of academic articles in psychology. **Frontiers in Psychology**, [*s. l.*], v. 6, n. MAR, p. 1–9, 2015.

JONES, James W.; MCCULLOUGH, Lawrence B.; RICHMAN, Bruce W. The ethics of bylines: Would the real authors please stand up? **Journal of Vascular Surgery**, [*s. l.*], v. 42, n. 4, p. 816–818, 2005.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, [*s. l.*], v. 349, n. 6245, p. 255–260, 2015.

KALAIVANI, G.; KAMALAKKANNAN, S. Web Scraping Technique for Prediction of Air Quality through Comparative Analysis of Machine Learning and Deep Learning Algorithm. *In*: 2022 INTERNATIONAL CONFERENCE ON AUGMENTED INTELLIGENCE AND SUSTAINABLE SYSTEMS (ICAISS), 2022, Trichy, India. **2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)**. Trichy, India: IEEE, 2022. p. 263–273. Disponível em: https://ieeexplore.ieee.org/document/10010968/. Acesso em: 10 mar. 2025.

KHDER, Moaiad. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. **International Journal of Advances in Soft Computing and its Applications**, [*s. l.*], v. 13, n. 3, p. 145–168, 2021.

KOEHLER, Martin *et al.* Data context informed data wrangling. *In*: 2017 IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), 2017, Boston, MA. **2017 IEEE International Conference on Big Data (Big Data)**. Boston, MA: IEEE, 2017. p. 956–963. Disponível em: http://ieeexplore.ieee.org/document/8258015/. Acesso em: 16 jan. 2025.

KOVALENKO, Dima. **Selenium design patterns and best practices: build a powerful, stable, and automated test suite using Selenium WebDriver**. Packt Publishing, 2014. (Community Experience Distilled). Disponível em: https://learning.oreilly.com/library/view/selenium-design-patterns/9781783982707/. Acesso em: 15 jan. 2025.

KUMAR, Aviral *et al.* A Workflow for Offline Model-Free Robotic Reinforcement Learning. [*S. l.*]: **arXiv**, 2021. Disponível em: http://arxiv.org/abs/2109.10813. Acesso em: 16 jan. 2025.

KUMAR, Sameer. Ethical concerns in the rise of co-authorship and its role as a proxy of research collaborations. **Publications**, [*s. l.*], v. 6, n. 3, 2018.

LI, Chunquan; CHEN, Yaqiong; SHANG, Yuling. A review of industrial big data for decision making in intelligent manufacturing. **Engineering Science and Technology, an International Journal**, [*s. l.*], v. 29, p. 101021, 2022.

LOPEZ, Patrice. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. *In*: AGOSTI, Maristella *et al.* (org.). **Research and Advanced Technology for Digital Libraries**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. (Lecture Notes in Computer Science). v. 5714, p. 473–474. Disponível em: http://link.springer.com/10.1007/978-3-642-04346-8_62. Acesso em: 10 mar. 2025.

LOZANO, George A. Ethics of Using Language Editing Services in An Era of Digital Communication and Heavily Multi-Authored Papers. **Science and Engineering Ethics**, [*s. l.*], v. 20, n. 2, p. 363–377, 2014.

MAÑANA-RODRÍGUEZ, Jorge. A critical review of SCImago Journal & Country Rank. **Research Evaluation**, [*s. l.*], v. 24, n. 4, p. 343–354, 2015.

MATTSSON, Pauline; SUNDBERG, Carl Johan; LAGET, Patrice. Is correspondence reflected in the author position? A bibliometric study of the relation between corresponding author and byline position. **Scientometrics**, [*s. l.*], v. 87, n. 1, p. 99–105, 2011.

MCNUTT, Marcia K. et al. Transparency in authors' contributions and responsibilities to promote integrity in scientific publication. **Proceedings of the National Academy of Sciences**, [s. l.], v. 115, n. 11, p. 2557–2560, 2018.

MEHRABI, N. *et al*. A Survey on Bias and Fairness in Machine Learning. **Journal of Artificial Intelligence Research**, 71, 1-35, 2021.

MENA-CHALCO, Jesús Pascual; JUNIOR, Roberto Marcondes Cesar. scriptLattes: an open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, [s. l.], 2009.

MONGEON, Philippe *et al.* The rise of the middle author: Investigating collaboration and division of labor in biomedical research using partial alphabetical authorship. **PLOS ONE**, [*s. l.*], v. 12, n. 9, p. 1–14, 2017.

MORAIS, Caroline *et al.* Estimativa da probabilidade de erro humano: uma análise da utilização e pesquisa dos métodos de confiabilidade humana, dados disponíveis e técnicas probabilísticas. In: **CONGRESSO DA ASSOCIAÇÃO BRASILEIRA DE ANÁLISE DE RISCO, SEGURANÇA DE PROCESSO E CONFIABILIDADE (ABRISCO), 2019**, Anais... Associação Brasileira de Análise de Risco, Segurança de Processo e Confiabilidade, 2019.

NASAR, Zara; JAFFRY, Syed Waqar; MALIK, Muhammad Kamran. Information extraction from scientific articles: a survey. **Scientometrics**, [s. l.], v. 117, n. 3, p. 1931–1990, 2018.

NASCIMENTO, Carlos Alberto Xavier *et al.* Innovation in business through big data analytics. **International Journal of Professional Business Review**, [*s. l.*], v. 3, n. 1, p. 1–16, 2018.

NOVAES, Douglas. A EFICIÊNCIA NA TOMADA DE DECISÃO POR MEIO DA ANÁLISE PREDITIVA INTEGRADA AOS SISTEMAS ERP. Revista Tópicos, v. 2, n. 6, 2024. ISSN: 2965-6672. Disponível em: https://zenodo.org/doi/10.5281/zenodo.10720638. Acesso em: 10 mar. 2025.

NTI, Isaac Kofi *et al.* A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. **Big Data Mining and Analytics**, [*s. l.*], v. 5, n. 2, p. 81–97, 2022.

NUNES, Maria das Graças Volpe. **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 2. ed. São Carlos, SP: Graça Nunes, 2024.

PATIENCE, Gregory S. *et al.* Intellectual contributions meriting authorship: Survey results from the top cited authors across all science categories. **PLoS ONE**, [*s. l.*], v. 14, n. 1, p. 1–20, 2019.

PENG, Xiaoting *et al.* Scientific misconduct responsibility attribution: An empirical study on byline position and team identity in Chinese medical papers. **PLOS ONE**, [*s. l.*], v. 19, n. 8, p. e0308377, 2024.

PIEDRA, Nelson *et al.* Guidelines to producing structured interoperable data from Open Access Repositories. *In*: 2016 IEEE FRONTIERS IN EDUCATION CONFERENCE (FIE), 2016, Erie, PA, USA. **2016 IEEE Frontiers in Education Conference (FIE)**. Erie, PA, USA: IEEE, 2016. p. 1–9. Disponível em: http://ieeexplore.ieee.org/document/7757660/. Acesso em: 10 mar. 2025.

PORTO, Anna Laura Magalhães; CORDEIRO, Douglas Farias. WEB SCRAPING E A BUSCA PELA GARANTIA DA QUALIDADE DOS DADOS EM UMA COLETA AUTOMATIZADA A PARTIR DE UMA REVISÃO BIBLIOGRÁFICA. **Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica**. 14(1), 2024.

POURRET, Olivier; BESANÇON, Lonni. **Authorship Versus Contributorship in Scientific Articles: Considering COPE Principles and the CRediT Taxonomy**. [*S. l.*]: Open Science Framework, 2024. Disponível em: https://osf.io/a7s2x. Acesso em: 10 mar. 2025.

PRIEM, Jason; PIWOWAR, Heather; ORR, Richard. **OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts**. [*S. l.*]: arXiv, 2022. Disponível em: https://arxiv.org/abs/2205.01833. Acesso em: 10 mar. 2025.

RABBY, Gollam et al. A Flexible Keyphrase Extraction Technique for Academic Literature. **Procedia Computer Science**, [s. l.], v. 135, p. 553–563, 2018.

RAGHUPATHI, Wullianallur; RAGHUPATHI, Viju. Big data analytics in healthcare: promise and potential. **Health Information Science and Systems**, [*s. l.*], v. 2, n. 1, p. 3, 2014.

ROSENZWEIG, Jaime S. *et al.* Authorship, collaboration, and predictors of extramural funding in the emergency medicine literature. **The American Journal of Emergency Medicine**, [*s. l.*], v. 26, n. 1, p. 5–9, 2008.

RUSSELL, Arielle F. *et al.* A Bibliometric Study of Authorship and Collaboration Trends Over the Past 30 Years in Four Major Musculoskeletal Science Journals. **Calcified Tissue International**, [s. l.], v. 104, n. 3, p. 239–250, 2019.

SARKER, Iqbal H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. **SN Computer Science**, [*s. l.*], v. 2, n. 6, p. 420, 2021.

SEWALD JUNIOR, Egon; SANTOS JÚNIOR, Sérgio Murilo Dos. Impacto da Ciência de Dados na Tomada de Decisão: um estudo de caso e um guia de implantação. **Navus - Revista de Gestão e Tecnologia**, [*s. l.*], p. 1–16, 2024.

SINGRODIA, Vidhi; MITRA, Anirban; PAUL, Subrata. A Review on Web Scrapping and its Applications. *In*: 2019 INTERNATIONAL CONFERENCE ON COMPUTER COMMUNICATION AND INFORMATICS (ICCCI), 2019, Coimbatore, Tamil Nadu, India. **2019 International Conference on Computer Communication and Informatics (ICCCI)**. Coimbatore, Tamil Nadu, India: IEEE, 2019. p. 1–6. Disponível em: https://ieeexplore.ieee.org/document/8821809/. Acesso em: 16 jan. 2025.

SOUZA, Daniel Soares De *et al.* Aprimorando Seleções Acadêmicas com Técnicas de Entropia e Método AHP-TOPSIS-2N. **Navus - Revista de Gestão e Tecnologia**, [*s. l.*], v. 14, 2024. Disponível em: https://navus.sc.senac.br/navus/article/view/1894. Acesso em: 31 jan. 2025.

SOUZA, E. M. DE. **Aplicação de ciência de dados na análise do posicionamento autoral e contribuições científicas em artigos**. [s.l.] Universidade Nove de Julho, 2021. Disponível em: https://bibliotecatede.uninove.br/bitstream/tede/2788/2/Edson%20Melo%20de%20Souza.pdf. Acesso em: 3 jan. 2025.

TANG, Jie *et al.* ArnetMiner: extraction and mining of academic social networks. *In*: KDD08: THE 14TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2008, Las Vegas Nevada USA. **Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining**. Las Vegas Nevada USA: ACM, 2008. p. 990–998. Disponível em: https://dl.acm.org/doi/10.1145/1401890.1402008. Acesso em: 15 jan. 2025.

TANG, Jie; ZHANG, Duo; YAO, Limin. Social Network Extraction of Academic Researchers. *In*: **SEVENTH IEEE INTERNATIONAL CONFERENCE ON DATA MINING (ICDM 2007)**, 2007, Omaha, NE, USA. Seventh IEEE International Conference on Data Mining (ICDM 2007). Omaha, NE, USA: IEEE, 2007. p. 292–301. Disponível em: http://ieeexplore.ieee.org/document/4470253/. Acesso em: 26 jan. 2025.

TOHALINO, Jorge Andoni Valverde. **Usando redes complexas e processamento de línguas naturais para caracterizar e classificar itens científicos**. 2023. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, University of São Paulo, São Carlos, 2023. doi:10.11606/T.55.2023.tde-11042023-090221. Acesso em: 2025-03-10.

TRAN, Hung Nghiep; HUYNH, Tin; DO, Tien. Author Name Disambiguation by Using Deep Neural Network. *In*: NGUYEN, Ngoc Thanh *et al.* (org.). **Intelligent Information and Database Systems**. Cham: Springer International Publishing, 2014. (Lecture Notes in Computer Science). v. 8397, p. 123–132. Disponível em: http://link.springer.com/10.1007/978-3-319-05476-6_13. Acesso em: 16 jan. 2025.

TSCHARNTKE, Teja *et al.* Author sequence and credit for contributions in multiauthored publications. **PLoS Biology**, [*s. l.*], v. 5, n. 1, p. 0013–0014, 2007.

XU, Michael; KUMAR, Abinash; LEBEAU, James M. Towards Augmented Microscopy with Reinforcement Learning-Enhanced Workflows. **Microscopy and Microanalysis**, [s. l.], v. 28, n. 6, p. 1952–1960, 2022.

YANG, Siluo; WOLFRAM, Dietmar; WANG, Feifei. The relationship between the author byline and contribution lists: a comparison of three general medical journals. **Scientometrics**, [*s. l.*], v. 110, n. 3, p. 1273–1296, 2017.

ZBAR, Ariella; FRANK, Erica. Significance of authorship position: An open-ended international assessment. **American Journal of the Medical Sciences**, [*s. l.*], v. 341, n. 2, p. 106–109, 2011.

ZHAO, Bo. Web Scraping. *In*: SCHINTLER, Laurie A.; MCNEELY, Connie L. (org.). **Encyclopedia of Big Data**. Cham: Springer International Publishing, 2022. p. 951–953. Disponível em: https://link.springer.com/10.1007/978-3-319-32010-6_483. Acesso em: 10 mar. 2025.

ZHAO, Linlin *et al*. Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling. **Drug Discovery Today**, [s. l.], v. 25, n. 9, p. 1624–1638, 2020.