What about data science? An analysis of the market based on Job posts

Barbara A. P. Barata[1], **Adrielson F. Justino**[1], **Antonio F. L. Jacob Junior**[1], **Fábio M. F. Lobato**[1,2]

[1]Programa de Pós-Graduação em Engenharia de Computação e Sistemas (PECS)
Universidade Estadual do Maranhão - (UEMA)

[2]Instituto de Engenharia e Geociências
Universidade Federal do Oeste do Pará (UFOPA)

antoniojunior@professor.uema.br, fabio.lobato@ufopa.edu.br

Abstract. *The growing importance of data science in the business landscape is indisputable. Consequently, there is also an increase in job posts related to these rising professions. The automatic analysis of this data can benefit professionals who want to enter this area or are looking for opportunities in the job market, as well as for universities and companies. In this context, the goal of the present study is to collect and analyze job posts related to data science from different sources using text-mining techniques. To this end, 6,000 job posts were analyzed on the leading platforms in the area, such as Indeed, NewScientist, Efinancialcareers, and Pharmiweb. The results provide valuable insights into key technologies, behavioral skills, ad sites, and general requirements. The findings of this study have the potential to guide possible updates for the development of technical and interpersonal skills according to labor market trends, helping people looking for relocation on the market and allowing building curricula that are more in line with market demands.*

Resumo. *A importância crescente da ciência de dados no cenário empresarial é indiscutível. Consequentemente, também se observa um crescimento na quantidade de anúncios de vagas relacionadas a essas profissões em ascensão. A análise automática desses dados pode trazer benefícios tanto para profissionais que desejam ingressar nessa área ou buscar oportunidades no mercado de trabalho, quanto para universidades e empresas. Neste contexto, o objetivo deste artigo é coletar e analisar trabalhos relacionados à ciência de dados provenientes de diferentes fontes, utilizando técnicas de mineração de texto. Para tal, foram analisados mais de 6.000 anúncios de vagas nas principais plataformas da área, como Indeed, NewScientist, Efinancialcareers e Pharmiweb. Os resultados obtidos fornecem insights valiosos sobre as principais tecnologias, habilidades comportamentais, sites de anúncios e requisitos gerais. As descobertas deste estudo têm o potencial de direcionar possíveis atualizações para o desenvolvimento de competências técnicas e interpessoais de acordo com as tendências do mercado de trabalho, auxiliando pessoas em busca de recolocação profissional e permitindo a construção de currículos mais alinhados com as demandas do mercado.*

1. Introdução

O século XXI inaugurou a era do *big data* e da economia de dados, na qual o DNA de dados tornou-se um constituinte intrínseco das organizações [Cao 2017]. Sendo assim, o crescimento dos dados gerados por negócios, governo, vida cotidiana e pesquisa científica lançaram muitos esforços para compreendê-los e utilizá-los [Brady 2019]. À vista disso, a capacidade de entender a estrutura e o conteúdo do discurso humano expandiu consideravelmente a dimensionalidade dos conjuntos de dados disponíveis [Agarwal and Dhar 2014]. Na chamada sociedade da informação, os dados se tornam a principal matéria-prima do contexto econômico. Um indivíduo que possui um *smartphone* deixa de ser apenas um consumidor para potencialmente se tornar também um produtor massivo de dados [Reis et al. 2020].

Neste contexto, a natureza e a organização do trabalho está experimentando mudanças em escalas sem precedentes. A pós-pandemia afetou a economia, demografia, além de avanços tecnológicos [Scully-Russ and Torraco 2020]. Por conseguinte, os mercados de trabalho globais demandam novos perfis de negócios e força de trabalho [Jagannathan et al. 2019]. Frente à estas transformações, percebe-se que a investigação de tal fenômeno é de extrema relevância, uma vez que facilita a identificação e o estudo sistemático de noções relacionadas ao fator humano e seu papel no ambiente profissional, como as competências técnicas e as habilidades sociais [Papoutsoglou et al. 2019]. Segundo [Di Battista et al. 2023], há uma demanda crescente por requalificação devido às lacunas de habilidades e talentos que estão afetando a evolução da indústria. Neste contexto, compreender e acompanhar as habilidades necessárias nesse mercado representa um desafio contínuo para aqueles interessados em ingressar no mercado ou progredir em suas carreiras.

No que tange ao escopo deste estudo, dados obtidos por meio do *Google Trends*[1] refletem a comparação entre os termos *"big data"* e *"data science"* nos últimos dois anos em todo o mundo, evidenciando um vasto interesse pelo termo *"data science"*, que está em crescente utilização em pesquisas. A ciência de dados é um paradigma interdisciplinar e abrangente, onde diferentes teorias e modelos são combinados para transformar dados em conhecimento (e valor) [Grossi et al. 2021]. Neste sentido, a área de Ciência de Dados abre várias possibilidades, ao mesmo passo que a formação para essa área torna-se um desafio. Outro aspecto crítico é a dispersão das informações de vagas de emprego. Atualmente existem muitos sites dedicados ao recrutamento e divulgação de empregos como consequência da digitalização do mercado de trabalho [Khaouja et al. 2021]. As ofertas descrevem as atividades e atribuições a desenvolver nas vagas disponíveis, junto aos conhecimentos, competências, formação e experiência requeridos. Um anúncio de emprego típico também contém informações sobre a disposição da contratação, faixa salarial e demais aspectos operacionais [Paletta and Moreiro González 2021].

Considerando a multitude de plataformas de recrutamento e o volume de dados gerado, a extração de conhecimento de tais dados representa um desafio de pesquisa importante. À luz desta problemática, o presente estudo tem como objetivo analisar ofertas de vagas relacionadas à ciência de dados utilizando métodos de mineração de textos. Para tal, foram coletados aproximadamente 6.000 anúncios de diversas plataformas, seguindo as melhores práticas de ciência de dados como análise exploratória, modelagem e

[1] http://surl.li/hplpu

validação dos *insights* obtidos. Os resultados alcançados permitiram identificar as principais tecnologias, habilidades comportamentais, sites de anúncios e requisitos gerais. Tais informações têm o potencial de guiar pessoas que buscam recolocação no mercado e também na atualização de estruturas curriculares de cursos de graduação e pós-graduação, além de permitir a construção de políticas públicas fomentadoras para este setor.

O restante do artigo encontra-se organizado como segue. Na Seção 2 são apresentados os trabalhos relacionados. Na Seção 3 são descritos os materiais e métodos. Na Seção 4 são apresentados os resultados decorrentes das análises. Por fim, na Seção 5 são apresentadas as conclusões e sugestões de trabalhos futuros.

2. Trabalhos relacionados

Para permitir um melhor direcionamento sobre a manipulação de dados textuais relacionados ao mercado de trabalho, foram levantados na literatura alguns trabalhos relacionados à Mineração de Texto para descoberta de conhecimento em anúncios de emprego. [Gurcan and Cagiltay 2019] propuseram uma metodologia semi-automática para analisar o conteúdo de anúncios de emprego *online* voltados para engenharia de *software* e *big data*. Os anúncios foram consultados pelo site Indeed e o método de modelagem de tópicos utilizado foi o *Latent Dirichlet Allocation* (LDA). Este método foi aplicado visando evidenciar os domínios de conhecimento e conjuntos de habilidades requeridos. Em [Debao et al. 2021] os autores buscaram entender as características dos requisitos para vagas de *big data* a partir dos aspectos da cidade para onde a vaga é destinada, faixa salarial, histórico educacional e experiência requerida. Os dados foram obtidos de um site chinês de recrutamento, o Zhaopin. O algoritmo utilizado para o agrupamento de texto foi o *k-means*. Neste estudo é levantado algumas deficiências como o formato desigual das informações, destacando a relevância de combinação de mais tecnologias de mineração de texto para um resultado mais robusto.

[Ao et al. 2022] direcionaram seus estudos para comparar vários métodos usados na literatura para extrair requisitos de habilidade do texto de anúncios de emprego. Os autores utilizaram três abordagens, duas baseadas na classificação de habilidades de artigos econômicos específicos e uma baseada no Dicionário Europeu de Habilidades e Competências; e também na aplicação da modelagem de tópicos LDA. O modelo LDA mostrou um melhor desempenho, entretanto, os autores destacaram que o método também apresentou tópicos de difícil interpretabilidade e anotação. Como indicação de trabalho futuro tem-se o estudo de estratégias/métodos para melhorar a interpretação dos tópicos. [Ternikov 2022] evidenciou a demanda para habilidades na esfera de Tecnologia da Informação na região da Comunidade de Estados Independentes, realizando o mapeamento entre os conjuntos de habilidades necessárias e as ocupações de trabalho. Para identificação de habilidades foram usados processamento de linguagem natural, agrupamento hierárquico e regras de associação. Os autores sugerem que a rede educacional precisa estar conectada às atualizações do conjunto de habilidades das tendências do mercado de trabalho, sendo que mudanças acontecem frequentemente e indicar formas e/ou ferramentas para categorização e extração de competências é primordial.

[Gurcan 2019] extraiu as competências essenciais solicitadas em vagas de *big data* utilizando modelagem de tópicos por meio do LDA. Os resultados obtidos permitiram a construção de uma taxonomia de competências para *big data*. O traba-

lho forneceu *insights* significativos para aperfeiçoar o entendimento basilar das características e tendências de trabalhos da área em questão. As descobertas do estudo têm implicações significativas para os principais atores de *big data* de diferentes aspectos, sendo a indústria, academia e as comunidades relacionadas. [Alibasic et al. 2022] propuseram uma abordagem baseada em dados para identificar tendências de empregos por meio de um estudo de caso na indústria de petróleo e gás. A abordagem proposta utilizou uma variedade de ferramentas de análise de dados que são: *Latent Semantic Indexing* (LSI), LDA, *Factor Analysis* e *Non-Negative Matrix Factorization* (NMF). Esse estudo foi capaz de identificar a discrepância entre as habilidades cobertas pelo sistema educacional e as habilidades exigidas no mercado de trabalho.

Por meio da avaliação dos trabalhos relacionados verificou-se que a aplicação da Mineração de Texto em anúncios de emprego extrai informações precisas dos dados textuais relacionados ao mercado de trabalho. Os estudos exploraram competências, requisitos e tendências em diferentes setores. Ao utilizar técnicas como *N-grama*, modelagem de tópicos, agrupamento e processamento de linguagem natural, é possível revelar informações valiosas sobre demandas e habilidades necessárias em várias ocupações. Os resultados têm implicações relevantes para a indústria, academia e comunidades, contribuindo para o aprimoramento do alinhamento entre a educação e o mercado de trabalho. É destacada a importância de explorar estratégias para aprimorar a interpretação e extração de informações dos textos analisados, promovendo uma compreensão mais abrangente das tendências e características do mercado de trabalho em diversas áreas. Essas descobertas também fornecem orientações valiosas para profissionais e instituições envolvidas no desenvolvimento de habilidades.

Considerando a tendência crescente no mercado de ciência de dados e o interesse contemporâneo nessa área, o presente estudo contribui com: i) a análise de anúncios de emprego de diferentes plataformas; ii) a investigação de técnicas de mineração de texto, como o uso de *N-grama*, modelagem de tópicos; e iii) a construção de um léxico específico para a área. As avaliações são realizadas de forma qualitativa, fornecendo subsídios tanto para profissionais interessados em vagas na área de ciência de dados quanto para futuras pesquisas nesse campo em constante evolução. O diferencial deste estudo é a abordagem mais abrangente na análise de anúncios de emprego, especialmente no mercado de ciência de dados. Destacam-se diferentes técnicas de mineração de texto e sua aplicação qualitativa, enquanto outros estudos focam em abordagens mais específicas e suas metodologias individuais.

3. Materiais e Métodos

O processo adotado foi o CRISP-DM (*Cross-Industry Standard Process for Data Mining*), o qual foi escolhido por ser amplamente reconhecido e valorizado na área [Wirth and Hipp 2000, Costa et al. 2022, Lobato et al. 2023]. O CRISP-DM é composto pelas etapas apresentadas na Figura 1, a saber: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implementação. As próximas subseções apresentam o detalhamento dessas etapas no escopo do trabalho.

3.1. Entendimento do negócio

Em um primeiro momento, foi realizada uma investigação preliminar para obter informações sobre ciência de dados, o papel dos profissionais dessa área, a importância

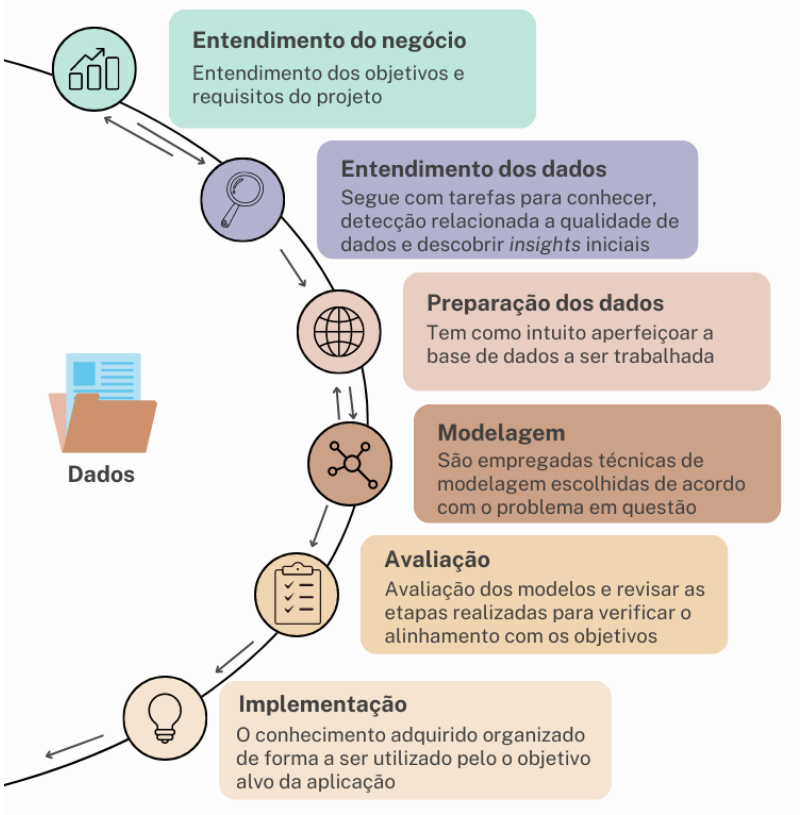

Figura 1. Fases do método CRISP-DM.

da análise de anúncios de emprego e sua conexão com os objetivos desta pesquisa. Posteriormente, foi conduzida uma revisão da literatura para identificar estudos relacionados à extração de informações de vagas de emprego, os quais foram apresentados na Seção 2. Além disso, foram feitas pesquisas na literatura cinza, que engloba uma ampla gama de formatos, como relatórios, documentos de trabalho, documentos governamentais, entre outros. Tais ações visaram a obtenção de informações valiosas visando enriquecer a pesquisa e auxiliar na tomada de decisões.

3.2. Entendimento e preparação dos dados

A ferramenta *Diffbot*[2] foi utilizada para extrair dados de várias plataformas de divulgação de vagas. Por meio dessa ferramenta é possível coletar informações de diversas fontes na internet, incluindo notícias, artigos, produtos, mídias sociais e vagas de emprego. A obtenção de dados no *Diffbot* se deu por meio da pesquisa na categoria de trabalho. Em seguida, aplicou-se o filtro de busca no título, utilizando a palavra-chave *"Data Scientist"*, o que resultou em 6.000 anúncios de vagas provenientes de diversos sites relacionados a oportunidades de emprego nessa área.

A plataforma, por sua vez, realiza automaticamente a identificação de padrões nos dados, como títulos, descrições e outros, e organiza essas informações em colunas correspondentes na planilha, permitindo baixar os dados coletados. Essa abordagem evita a necessidade de manipulação de informações não estruturadas, resultando em economia

[2]https://www.diffbot.com/

Tabela 1. Campos do conjunto de dados coletados e suas informações.

Campos	descrição	Quantidade
Id	Número do identificador	6.000
Title	Título do cargo do anúncio da vaga correspondente	6.000
Text	Corpo do texto da descrição da vaga	5.831
humanlanguage	Idioma do texto	6.000
pageUrl	URL da fonte dos anúncios	6.000
requirements	Informações relacionadas aos requisitos	2.258
tasks	Informações sobre as tarefas do cargo	2.981

de tempo e esforço. Para compreender a composição da base de dados, foram avaliados alguns aspectos, como a quantidade de dados em cada coluna, a fim de ter uma noção da dimensão dos mesmos. A Tabela 1 apresenta os campos presentes na base de dados, incluindo *id*, *title*,*text*, *humanlanguage*, *Uniform Resource Locator*(URL), *requirements* e *tasks*. É importante destacar que nem todos os anúncios continham os dois últimos campos (*requirements* e *tasks*).

Ao se deparar com anúncios de emprego, é salutar compreender a estrutura dos mesmos a fim de identificar as informações potencialmente úteis. Diversos recursos *online*, como o *site Indeed*o blog da *Gupy*, o *LinkedIn* e o *Glassdoor*, fornecem orientações e diretrizes para a criação de anúncios de emprego eficazes. Esses materiais abordam aspectos como a estrutura do anúncio, os elementos essenciais a serem incluídos, a linguagem adequada, os requisitos e habilidades desejados, bem como informações sobre a empresa e a cultura organizacional. Ao entender a estrutura desses anúncios é possível extrair informações valiosas para análises, identificando melhor as tendências e demandas do mercado de trabalho. Sendo assim, a partir do entendimento dos dados, foram selecionados os campos principais para análise, como o título e o texto dos anúncios. Outros campos, como URL, idioma e requisitos, foram manipulados de forma secundária. No caso das URLs, foi realizada uma operação de limpeza para remover o prefixo e o sufixo do caminho. Além disso, foram utilizadas expressões regulares para extrair informações sobre anos de experiência mencionados nos anúncios.

O pré-processamento do conteúdo textual foi realizado baseando-se nos trabalhos [de Almeida et al. 2017, Cirqueira et al. 2018, Silva et al. 2021] e consistiu nos seguintes passos: a) remoção de duplicados, que contou com o *id* para a verificação; b) filtragem do idioma "inglês", dado a sua maior frequência; c) padronização do texto; d) remoção de acentuação, *stopwords*, pontuações e espaços excessivos. Algumas palavras específicas, como *"homedepot", "mythdhr", "jobs", "job", "cookies"*, foram adicionadas a lista de *stopwords*. Essas etapas foram essenciais para garantir a qualidade e a consistência dos dados antes da análise. Selecionamos 1.691 descrições de vagas para ser realizada a leitura com o intuito de contribuir com as anotações utilizadas na construção do léxico.

3.3. Modelagem

Para identificar padrões nos anúncios de emprego, utilizamos diferentes técnicas de análise. Adotamos a abordagem de geração de *N-grama*, com enfoque em bigramas e trigramas, considerando o entendimento do domínio de aplicação descrito na Seção 3.1. As análises de *N-gramas* foram realizadas com os títulos das vagas, onde foi utilizada a representação baseada em *term frequency–inverse document frequency* (TF-IDF). Este es-

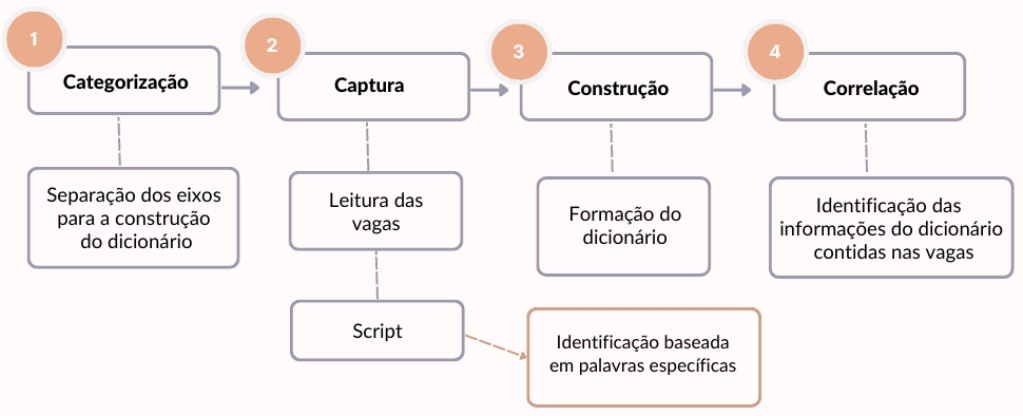

Figura 2. Etapas da construção do léxico.

tudo também aplicou a modelagem de tópicos para extrair conhecimento das descrições de vagas, usando a técnica BerTopic. A técnica é capaz de fornecer tópicos mais contextualmente específicos [Baird et al. 2022]; ao contrário do LDA, mais prevalente nos trabalhos relacionados, que possui difícil anotação. O BerTopic gera a incorporação de documentos com modelos de linguagem baseados em transformadores pré-treinados, agrupa essas incorporações e, finalmente, gera representações de tópicos com o procedimento TF-IDF baseado em classe [Grootendorst 2022]. No desenvolvimento da modelagem de tópicos, é inicialmente realizado o treinamento, seguido pela geração de tópicos para análises preliminares. Os tópicos são compostos por palavras-chave que refletem os principais temas nos documentos. O BerTopic determina automaticamente a quantidade de tópicos, com opção de redução após o treinamento. A etapa seguinte implica em uma análise qualitativa, na qual os tópicos são avaliados e ajustados, se requerido. O objetivo é determinar um número de tópicos realista e facilmente interpretável por um especialista humano, seguindo o conceito de *human-in-the-loop* [Wu et al. 2022].

Outra análise empregada envolveu a criação de um léxico de vagas de emprego, seguindo as etapas de categorização, captura, construção e correlação, conforme descrito na Figura 2. Na fase de categorização, quatro eixos foram definidos para a construção do dicionário a partir da análise dos elementos dos anúncios. Esses eixos incluem *conhecimentos*, *tecnologias*, *linguagem de programação* e *habilidades comportamentais*. Após essa definição, o processo prossegue com a parte da extração das informações das descrições a fim de construir o dicionário. Essa etapa emprega uma abordagem semi-automática, que combina a leitura das descrições com a utilização de um *script* de pesquisa de termos para realizar as anotações necessárias. Durante a leitura das descrições, palavras-chave relevantes relacionadas aos eixos definidos são adicionadas a uma lista. Além disso, o *script* faz buscas baseada em termos específicos identificados na leitura da amostra de dados, contribuindo para a composição do léxico.

3.4. Avaliação e implementação

Para a avaliação optou-se pela utilização da abordagem embasada em *grounded theory*, sendo um método de pesquisa qualitativo que permite aos pesquisadores discernir processos explícitos e implícitos em seus dados [Charmaz and Thornberg 2021]. Neste trabalho a abordagem foi conduzida de maneira que os(as) autores(as) fizessem a leitura de

uma amostragem relevante dos dados e cruzassem com os resultados obtidos. Destaca-se também que para a realização do cálculo amostral, foi levado em consideração o tamanho da população, o nível de confiança de 95% e uma margem de erro de 2. Com base nesses parâmetros, o tamanho da amostra foi determinado como 1.691. Outro aspecto pertinente foi o contato com *tech recruiters*, que possibilitou validar alguns *insights* obtidos com os resultados das análises.

3.5. Tecnologias

A linguagem de programação utilizada no desenvolvimento dos *scripts* de análise foi o Python, considerando o seu vasto repositório de bibliotecas para mineração de dados, em especial, para a mineração de textos. Utilizaram-se *notebooks* no ambiente de execução *Google Colaboratory*[3] na sua versão Colab Pro+. Resumidamente, as seguintes bibliotecas foram utilizadas para a preparação e modelagem dos dados: o *pandas* [4] foi utilizado para a manipulação dos dados; o *Natural Language Toolkit*[5](NLTK) foi utilizado para o pré-processamento de textos, em especial, para a remoção de palavras vazias (*stopwords*); nesta etapa também foram utilizadas expressões regulares implementadas na biblioteca (RE); o *scikit-learn*[6] foi utilizado para a representação de dados textuais utilizando TF-IDF e para a geração dos bigramas e trigramas. Para a visualização de dados utilizaram-se o matplotlib/pyplot[7], o *Seaborn*[8] e o *WordCloud*[9]. Visando a reprodutibilidade, a base utilizada e o dicionário encontram-se disponíveis no repositório `https://github.com/fabiolobato/ENIAC23-Jobs`.

4. Resultados

Após o pré-processamento dos dados, foi realizada a análise de um total de 5.709 anúncios de emprego. A Figura 3 apresenta uma representação visual das descrições das vagas em forma de nuvem de palavras. Por meio dessa análise, é possível identificar várias informações relevantes, como habilidades, conhecimentos e requisitos mencionados. Essas informações destacam a importância da combinação de termos semelhantes na compreensão aprofundada dos anúncios. A análise minuciosa da Figura 3 permite identificar outros aspectos não mapeados nas perguntas de pesquisa, como a preocupação da comunidade com a inclusão e equidade. É possível verificar a ocorrência de expressões como *"sexual orientation"*, *"gender identity"*, *"race color"*, *"equal oportunity"* e *"religion"*.

Com o objetivo de examinar as principais plataformas de anúncios de vagas, realizou-se a contagem da frequência dos *sites* com base na coluna *url*. A representação das dez plataformas mais frequentes pode ser observada na Figura 4. Por meio de sua análise é possível inferir que há uma variedade nas fontes encontradas, o que permite uma compreensão mais abrangente das descrições das vagas. Diferentemente do esperado, plataformas bem conhecidas no Brasil como o LinkedIn, 99jobs e Infojobs, não estão entre as opções predominantes. Destarte, este achado de pesquisa tem o potencial de

[3]https://colab.research.google.com/
[4]https://pandas.pydata.org/
[5]https://www.nltk.org/
[6]https://scikit-learn.org/stable/
[7]https://matplotlib.org/stable/tutorials/introductory/pyplot.html
[8]https://seaborn.pydata.org/
[9]https://pypi.org/project/wordcloud/

Figura 3. Representação em nuvem de palavras das descrições dos anúncios.

guiar profissionais que estão buscando colocação na área de ciência de dados no cenário internacional.

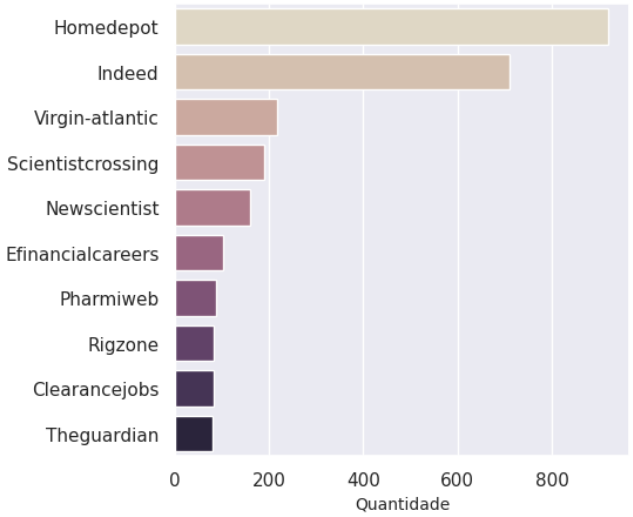

Figura 4. Plataformas com a maior quantidade de anúncios de vagas de ciência de dados.

Os resultados da análise dos *N-gramas* nos títulos dos anúncios são apresentados na Tabela 2, a qual exibe 12 palavras mais comuns nos documentos. Pode-se observar que o termo mais frequente é *senior data scientist* (386), indicando a necessidade de um alto nível de experiência para essas posições. No entanto, também é notável a presença do termo *junior data scientist* com apenas 37 ocorrências, sugerindo que o mercado oferece poucas oportunidades para profissionais iniciantes. Sendo assim, ter um entendimento dos títulos mencionados nas vagas é fundamental para compreender quais cargos estão

em demanda e se há o surgimento de novas formas de denominação para essas posições.

Tabela 2. Representação dos *N-gramas* construídos baseado nos títulos das vagas.

N-gramas	Frequência
Senior data scientist	386
Data scientist jobs	97
Lead data scientist	87
Associate data scientist	85
Principal data scientist	82
Machine learning	52
Jobs united states	44
Junior data scientist	37
Data scientist intern	29
Data analytics	16
Program data scientist	14

Durante a análise dos anúncios de emprego, foi possível examinar a distribuição dos requisitos de anos de experiência solicitados pelas empresas. Os dados foram representados na Figura 5, que apresenta a quantidade de vagas para cada ano de experiência. Conforme observado no gráfico, há uma variedade nas demandas das empresas em relação à experiência dos candidatos. A faixa mais comum é de 5 anos de experiência, representando um total de 565 vagas. Em seguida, temos as faixas de 3 anos, com 466 vagas, e 8 anos, com 453 vagas. Essas informações são relevantes para profissionais que buscam oportunidades no mercado de trabalho, pois oferecem percepções sobre as demandas e preferências das empresas em relação à experiência dos candidatos.

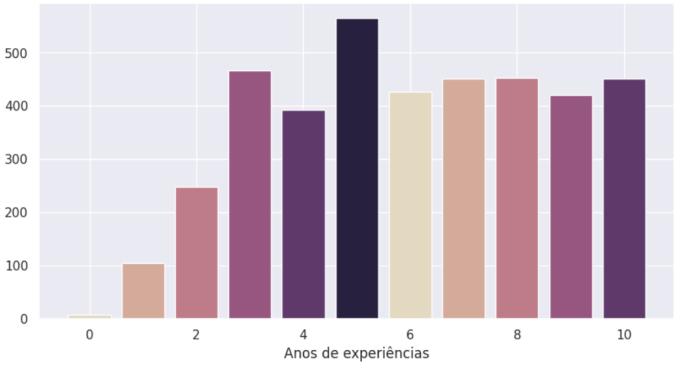

Figura 5. Ano mínimo de experiência indicado nas vagas.

Em relação à modelagem de tópicos, inicialmente o BerTopic gerou 139 tópicos. Visando melhorar a legibilidade e seguindo a estratégia *human-in-the-loop*, foram testados diversos valores para o número de tópicos (50, 40, 30 e 20). Após a leitura dos tópicos e considerando o conhecimento das descrições das vagas, foi selecionado o valor de 20 tópicos. Os tópicos -1, 4, 14, 15, 16 e 18 foram omitidos, pois continham palavras irrelevantes para as análises como posições preenchidas, vagas fechadas, problemas na coleta (edição de vaga em aberto) *etc*. A Tabela 3 apresenta os tópicos identificados juntamente com suas palavras-chave. Esses tópicos abrangem uma variedade de assuntos, como visão computacional, análise de dados e segurança da informação e informações sobre candidatura.

Tabela 3. Representação dos tópicos selecionados com os termos.

Tópicos	Palavras-chave
Dados e negócios	*Data, business, experience, work, analysis, learning, science, team, analytics, machine*
Pesquisa em bioinformática	*Bioinformatics, data, experience, computational, biology, genomics, scientist, data scientist, discovery, research*
Modelagem e busca	*Search, model, data, personalization, information retrieval, building, retrieval, model, productionalization*
Análise de clientes	*Clients, data, business, analytics, eval, function return, return eval, services, return, function*
Pesquisa de dados	*Data, research, education, university, experience, analysis, student, students, work, science*
Dados em grande escala	*Data, large, business, work, age, work experience, position, large scale, basis disability, federal*
Empregabilidade e carreira	*Search, joband great, joband, search questions, fantastic expert, flexible fantastic, hereother benefits, hereother, expert resources, itbecome*
Empregabilidade e oportunidades	*Latest posted, currently, scientist, posted, data scientist, position, scientist currently, position closed, open, closed*
Análise de dados	*Data, button, business, experience, analysis, work, analytics, click, skills, statistical*
Visão computacional	*Image, computer vision, vision, data, learning, development, experience, AI, computer*
Oportunidade e emprego	*Data, ZoomInfo, offered, h1b sponsorship, h1b, sponsorship, quality, talk, looking*
Empresa e análises	*Lowe, lowes, science, data, mooresville, quantitative analytic, related, analytic, modeling*
Segurança da informação	*Programmer, closedexpired, architect full, scientist cybersecurity, learning blockchain, developer senior, cybersecurity engineer, engineer machine, stack developer*
Mercado de trabalho	*Salary, glassdoor, scientist, data scientist, salaries, average, remote data, data, indeed, figures*

Ao criar o léxico, foi possível realizar uma análise meticulosa descrições de vagas com base na frequência das palavras. Ao considerar os critérios utilizados na elaboração do léxico, como as tecnologias, habilidades, conhecimentos específicos, requisitos e linguagens de programação, foi possível examinar as vagas e identificar a coocorrência de palavras, resultando na construção de uma representação com base nos termos mais comumente mencionados.

Os resultados da análise das tecnologias mais frequentes são apresentados na Figura 6. Por meio de sua análise é perceptível visualizar uma variedade de tecnologias, abrangendo banco de dados relacional como *MySQL*, banco de dados não relacional como *BongoDB*, sistema operacional *Linux*, visualização de dados, controle de versão e bibliotecas para manipulação de dados. Essas constatações indicam a relevância e demanda por domínio relacionadas a essas tecnologias no mercado de trabalho. Na Figura 7 são apresentadas as linguagens de programação mais recomendadas, revelando o predomínio do *Python*, seguido pela linguagem *R*.

Ao abordar o campo multidisciplinar da Ciência de Dados, é evidente que existem diversos conhecimentos associados à área. A Figura 8 apresenta os conhecimentos mais mencionados nas vagas de emprego analisadas, incluindo *"Machine Learning"*, *"Statistics"*, *"Data visualization"*, entre outros. Esses *insights* destacam a necessidade dos profissionais possuírem um sólido domínio de conceitos e aplicações relevantes para suas atividades diárias de trabalho. As análises das descrições dos anúncios confirmam a importância da atualização constante para os atores, coadunando com as conclusões de outros trabalhos do estado da arte [Ternikov 2022, Gurcan 2019]. A Figura 9 apresenta as habilidades mais mencionadas, juntamente com sua frequência de ocorrência.

A Figura 10 destaca o requisito de formação avançada para os profissionais, incluindo pós-doutorado e mestrado, além da importância da experiência prévia. Isso indica que os empregadores valorizam candidatos com um alto nível de qualificação acadêmica e experiência relevante na área. Em comparação com o estudo de [Debao et al. 2021], que analisou vagas relacionadas a *big data*, percebe-se que a exigência de formação é menor

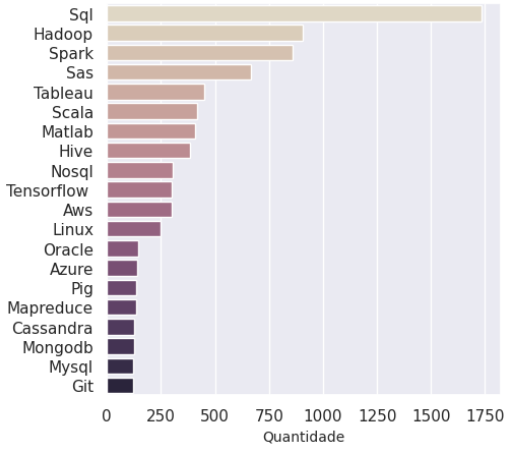

Figura 6. As tecnologias mais frequentes encontradas nas descrições das vagas.

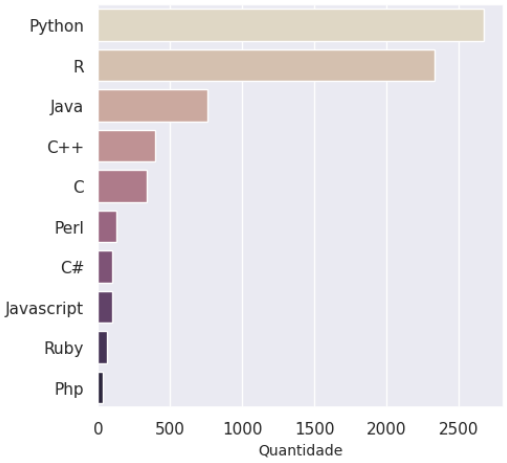

Figura 7. Linguagens de programação mais mencionadas nos anúncios.

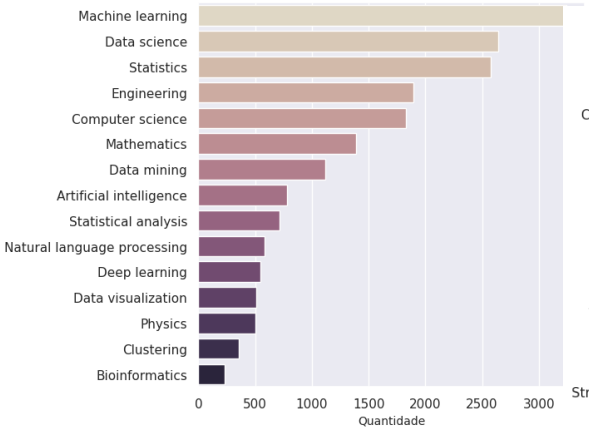

Figura 8. Conhecimentos mais recorrentes nas descrições das vagas.

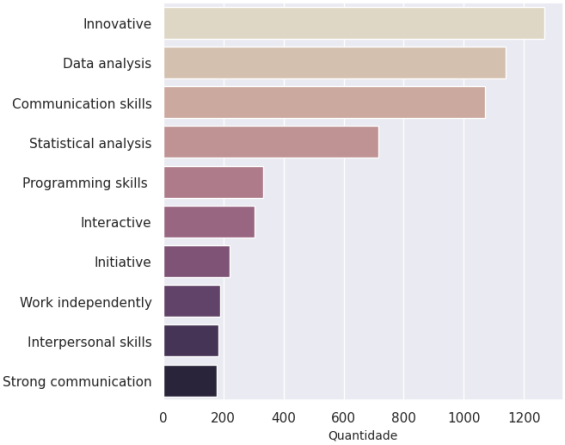

Figura 9. Habilidades mais recorrentes.

nessa área em comparação à ciência de dados.

Neste sentido, hipotetizamos que as vagas de cientistas de dados júniors eram preenchidas por meio da capacitação de colaboradores que já atuam nas empresas. Esta hipótese foi corroborada por *tech recruiters* contatados pelos(as) autores(as). Sendo assim, conclui-se que a entrada nesta área não ocorre por via direta, ou seja, por meio da contratação de cientistas de dados júnior por chamada pública. Este é um indicativo interessante para profissionais que procuram recolocação na área tecnologia e vêm migrando para ciência de dados.

Juntos, esses resultados fornecem informações importantes para que os interessados nas vagas estejam cientes dessas tendências e considerem adquirir ou aprimorar suas habilidades nesses eixos indicados para melhorar suas chances de sucesso na busca por emprego. Além disso, empregadores e instituições de ensino podem utilizar essas informações para ajustar seus currículos e programas de treinamento, a fim de atender

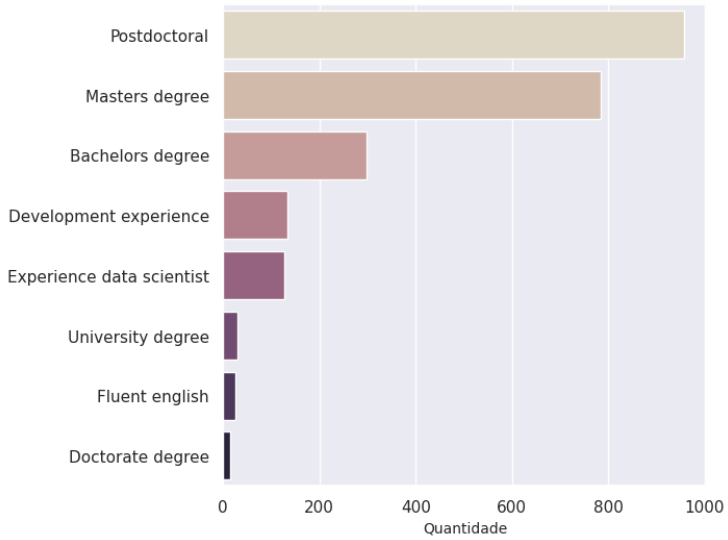

Figura 10. Indicação dos requisitos predominantes nas descrições.

às demandas do mercado de trabalho e garantir uma melhor correspondência entre as habilidades dos candidatos e as necessidades das vagas disponíveis.

5. Considerações finais

É inegável a importância da ciência dos dados no mundo corporativo. Considerando as dinâmicas mercadológicas, carreiras relacionadas à ciência de dados têm tido crescente interesse e trazendo consigo um aumento significativo nas ofertas de vagas. Nesse contexto, o presente artigo apresentou uma análise das descrições de vagas relacionadas à ciência de dados provenientes de diversas fontes. O diferencial deste estudo é a abordagem mais abrangente na análise de anúncios de empregos frente ao estado da arte. Para tal, foi realizado uma coleta de 6.000 anúncios de diferentes plataformas, visando obter uma ampla e representativa amostra do mercado de trabalho nessa área. As análises visaram identificar as habilidades, conhecimentos, cargos prevalentes, tecnologias e requisitos exigidos nas vagas de emprego.

Durante a condução deste trabalho foram utilizadas boas práticas de ciência de dados, o que possibilitou a realização de análises exploratórias, modelagem utilizando as abordagens de *N-grama*, modelagem de tópicos e a construção do léxico voltado para vagas; e validação dos resultados obtidos utilizando a abordagem *grounded-theory*, seguindo uma abordagem baseada no *human-in-the-loop*. Essas etapas foram fundamentais para compreender as características e os requisitos das vagas de ciência de dados, bem como para identificar padrões e tendências relevantes para a área. Seguiram-se também princípios de ciência aberta, onde as bases e os artefatos construídos foram publicamente disponibilizadas em um repositório aberto.

Os resultados alcançados permitiram identificar tendências da área, como cargos mais prevalentes, habilidades, conhecimento, tecnologias e requisitos. Tais informações têm o potencial de guiar pessoas que buscam recolocação no mercado e também na atualização de estruturas curriculares de cursos de graduação e pós-graduação, além de permitir a construção de políticas públicas fomentadoras para este mercado tão em alta.

A aplicação de técnicas de mineração de textos e a análise de dados provenientes de diferentes fontes se revelaram ferramentas poderosas para compreender as dinâmicas do mercado de trabalho e orientar decisões importantes relacionadas à seleção e qualificação de profissionais nessa área em constante evolução. Como trabalhos futuros, pretende-se agregar mais técnicas de mineração de texto para identificar conjuntos diferentes de requisitos entre os cargos analisados.

Agradecimentos

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)-DT-308334/2020; pela Fundação Amazônia de Amparo a Estudos e Pesquisas (FAPESPA) PRONEM-FAPESPA/CNPq nº 045/2021; e pelo Acordo de Cooperação Técnica N° 02/2021 (Processo N° 38328/2020 -TJ/MA).

Referências

Agarwal, R. and Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for is research.

Alibasic, A., Upadhyay, H., Simsekler, M. C. E., Kurfess, T., Woon, W. L., and Omar, M. A. (2022). Evaluation of the trends in jobs and skill-sets using data analytics: a case study. *Journal of Big Data*, 9(1):32.

Ao, Z., Horvath, G., Sheng, C., Song, Y., and Sun, Y. (2022). Skill requirements in job advertisements: A comparison of skill-categorization methods based on explanatory power in wage regressions. *arXiv preprint arXiv:2207.12834*.

Baird, A., Xia, Y., and Cheng, Y. (2022). Consumer perceptions of telehealth for mental health or substance abuse: a twitter-based topic modeling analysis. *JAMIA open*.

Brady, H. E. (2019). The challenge of big data and data science. *Annual Review of Political Science*, 22:297–323.

Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, 50(3):1–42.

Charmaz, K. and Thornberg, R. (2021). The pursuit of quality in grounded theory. *Qualitative research in psychology*, 18(3):305–327.

Cirqueira, D., Pinheiro, M. F., Jacob, A., Lobato, F., and Santana, Á. (2018). A literature review in preprocessing for sentiment analysis for brazilian portuguese social media. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*.

Costa, G. d. S., Couto, D. C., Junior, A. F. J., and Lobato, F. M. (2022). Feminismo e redes sociais online: uma análise de tweets sobre o dia internacional da mulher. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 169–180. SBC.

de Almeida, G. R., Cirqueira, D. R., and Lobato, F. M. (2017). Improving social crm through eletronic word-of-mouth: a case study of reclameaqui. In *Anais Estendidos do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 107–110. SBC.

Debao, D., Yinxia, M., and Min, Z. (2021). Analysis of big data job requirements based on k-means text clustering in china. *PloS one*, 16(8):e0255419.

Di Battista, A., Grayling, S., and Hasselaar, E. (2023). Future of jobs report 2023. Technical report, World Economic Forum, Geneva, Switzerland.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Grossi, V., Giannotti, F., Pedreschi, D., Manghi, P., Pagano, P., and Assante, M. (2021). Data science: a game changer for science and innovation. *International Journal of Data Science and Analytics*, 11:263–278.

Gurcan, F. (2019). Extraction of core competencies for big data: Implications for competency-based engineering education. *International Journal of Engineering Education*, 35(4):1110–1115.

Gurcan, F. and Cagiltay, N. E. (2019). Big data software engineering: Analysis of knowledge domains and skill sets using lda-based topic modeling. *IEEE access*.

Jagannathan, S., Ra, S., and Maclean, R. (2019). Dominant recent trends impacting on jobs and labor markets-an overview. *International Journal of Training Research*.

Khaouja, I., Kassou, I., and Ghogho, M. (2021). A survey on skill identification from online job ads. *IEEE Access*, 9:118134–118153.

Lobato, F., Poça, M., and Canto, V. (2023). Análise, otimização e acompanhamento de um serviço de psicologia universitário: uma abordagem baseada em ciência de dados. *Revista da CGU*, 15(27).

Paletta, F. C. and Moreiro González, J. A. (2021). A transformação digital e os impactos no mercado de trabalho: estudo dos anúncios de emprego na web para profissionais da informação no setor privado. *Information research*, 26(3).

Papoutsoglou, M., Ampatzoglou, A., Mittas, N., and Angelis, L. (2019). Extracting knowledge from on-line sources for software engineering labor market: A mapping study. *IEEE Access*, 7:157595–157613.

Reis, L. C. R., da Fonseca, M. I., et al. (2020). Big data: Um novo campo de atuação para bibliotecários. *Prisma. Com*, 2020(41):231–250.

Scully-Russ, E. and Torraco, R. (2020). The changing nature and organization of work: An integrative review of the literature. *Human Resource Development Review*, 19(1):66–93.

Silva, L. E., Schneider, E. T. R., Gumiel, Y. B., da Luz, M. A. P., Paraiso, E. C., Moro, C., et al. (2021). Experiments on portuguese clinical question answering. In *Anais da X Brazilian Conference on Intelligent Systems*. SBC.

Ternikov, A. (2022). Soft and hard skills identification: insights from it job advertisements in the cis region. *PeerJ Computer Science*, 8.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.