UNIVERSIDADE ESTADUAL DO MARANHÃO

CENTRO DE CIÊNCIAS TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO E SISTEMAS

MESTRADO PROFISSIONAL EM ENGENHARIA DE COMPUTAÇÃO E SISTEMAS

EXTRAÇÃO DE CONHECIMENTO DE ANÚNCIOS DE EMPREGO: UM ESTUDO DE CASO EM VAGAS DE CIÊNCIA DE DADOS.

BÁRBARA ADRIANA PIRES BARATA

Trabalho apresentado ao curso de Mestrado Profissional em Engenharia da Computação e Sistemas na Universidade Estadual do Maranhão como pré-requisito para obtenção do título de Mestre sob orientação do(a) Prof. Dr. Fábio Manoel França Lobato e Coorientação do(a) Prof. Dr. Antonio Fernando Lavareda Jacob Jr.

2024

UNIVERSIDADE ESTADUAL DO MARANHÃO

CENTRO DE CIÊNCIAS TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO E SISTEMAS

MESTRADO PROFISSIONAL EM ENGENHARIA DE COMPUTAÇÃO E SISTEMAS

BÁRBARA ADRIANA PIRES BARATA

EXTRAÇÃO DE CONHECIMENTO DE ANÚNCIOS DE EMPREGO: UM ESTUDO DE CASO EM VAGAS DE CIÊNCIA DE DADOS.

2024

Barata, Bárbara Adriana Pires
Extração de conhecimento de anúncios de emprego: um estudo de caso em vagas de ciências de dados. / Bárbara Adriana Pires Barata. – São Luis, MA, 2024.
107 f

Dissertação (Dissertação em Engenharia de Computação e Sistemas) – Universidade Estadual do Maranhão, 2024.

Orientador: Prof. Dr. Fábio Manoel França Lobato
Coorientador: Prof. Dr. Antonio Fernando Lavareda Jacob Jr.

1.Ciência de dados. 2.Análise de vagas. 3.Mineração de texto. 4.Análise de dados. 5.Modelagem de tópicos. I.Título

CDU: 004.775

Elaborado por Cássia Diniz- CRB 13/910

Bárbara Adriana Pires Barata

Extração de conhecimento de anúncios de emprego: um estudo de caso em vagas de Ciência de dados.

Trabalho apresentado ao curso de Mestrado Profissional em Engenharia da Computação e Sistemas na Universidade Estadual do Maranhão como pré-requisito para obtenção do título de Mestre sob orientação do(a) Prof. Dr. Fábio Manoel França Lobato e Coorientação do(a) Prof. Dr. Antonio Fernando Lavareda Jacob Jr.

São Luís - MA, 27 de fevereiro de 2024:

Documento assinado digitalmente
gov.br
FABIO MANOEL FRANCA LOBATO
Data: 23/07/2024 12:22:42-0300
Verifique em https://validar.iti.gov.br

Profº. Dr. Fábio Manoel França Lobato
(Orientador - UFOPA)

Documento assinado digitalmente
gov.br
ANTONIO FERNANDO LAVAREDA JACOB JUNIOI
Data: 24/07/2024 11:59:22-0300
Verifique em https://validar.iti.gov.br

Profº. Dr. Antonio Fernando Lavareda Jacob Jr
(Coorientador - UEMA)

Documento assinado digitalmente
gov.br
OMAR ANDRES CARMONA CORTES
Data: 24/07/2024 13:54:12-0300
Verifique em https://validar.iti.gov.br

Profº. Dr. Omar Andres Carmona Cortes
(Examinador Interno - IFMA)

Documento assinado digitalmente
gov.br
RICARDO MARCONDES MARCACINI
Data: 24/07/2024 12:03:24-0300
Verifique em https://validar.iti.gov.br

Profº. Dr. Ricardo Marcondes Marcacini
(Examinador Externo - USP)

São Luís - MA

2024

AGRADECIMENTOS

Primeiramente agradecer a Deus por ter me sustentado até aqui e a Maria santíssima por cuidar sempre dos detalhes. A minha família, principalmente meus pais Brígida Pires e Adriano Barata, meu primo Fábio de Lima por serem a minha rede de apoio nos diversos momentos. Ao meu orientador, Fábio Lobato, pelos direcionamentos e todos os conhecimentos passados os quais foram fundamentais nessa etapa. Aos professores do PECS-UEMA pelos conhecimentos transmitidos e aos companheiros de mestrado Marcos Januário, Matheus Marinato e Fabrício Almeida pelas trocas. Ao Acordo de Cooperação Técnica N° 02/2021 (Processo N° 38328/2020 -TJ/MA) pela oportunidade de bolsa e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Aos colegas do LACA-UFOPA que contribuíram no desenvolvimento da pesquisa. Ao Marlon Coelho, Ana Portilho, Danielle Costa e as Manas Digitais que foram uma grande rede de apoio, motivação e por deixarem a trajetória mais divertida. A todas as pessoas que contribuíram direta ou indiretamente com o meu processo de formação e que rezaram por mim.

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 de emprego relacionadas a essa área. 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 na questão de estruturas curriculares e empresas para capacitações e atualizações. Neste contexto, o objetivo deste trabalho é coletar e analisar trabalhos relacionados à ciência de dados provenientes de diferentes fontes, utilizando técnicas de mineração de texto a fim de identificar as tendências requeridas pelos empregadores. Para tal, foram analisados mais de 14.000 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, permitindo para universidades e empresas uma base para a construção de atualizações alinhadas com as demandas do mercado de trabalho.

Palavras-chave: Ciência de Dados, Análise de vagas, Mineração de texto, Análise de dados, Modelagem de tópicos, Mercado de Trabalho.

ABSTRACT

The growing importance of data science in the business scenario is indisputable. Consequently, there has also been an increase in the number of job advertisements related to this area. The automatic analysis of this data can benefit professionals who wish to enter this area or seek opportunities in the job market, as well as universities in terms of curriculum structures and companies for training and updates of their employees. In this context, this work aims to collect and analyze works related to data science from different sources, using text-mining techniques to identify the trends required by employers. To this end, more than 14,000 job advertisements were analyzed on the leading platforms in the area, such as Indeed, NewScientist, Efinancialcareers, and Pharmiweb. The results provide valuable insights into key technologies, soft skills, ad sites, and general requirements. The findings of this study have the potential to direct possible updates for developing technical and interpersonal skills following job market trends, helping people in search of professional relocation, and providing universities and companies with a basis for building aligned updates with the demands of the job market.

Key-words: Data Science, Job Analysis, Text Mining, Data Analysis, topic modeling, Job market.

LISTA DE ILUSTRAÇÕES

Figura 1 – Infográfico anual *Data Never Sleeps 11.0* 20
Figura 2 – Tendência de buscas no Google Trends pelos termos "*big data*" e "*data science*". 21
Figura 3 – Habilidades procuradas globalmente. 24
Figura 4 – Maneiras pelas quais as habilidades interpessoais podem contribuir para o sucesso do negócio. 25

Figura 5 – Evolução temporal dos trabalhos relacionados. 32

Figura 6 – Fases do método CRISP-DM. 35
Figura 7 – Painel da plataforma *Diffbot*. 37
Figura 8 – Fluxo de trabalho para a coleta de dados na plataforma *Diffbot*. 38
Figura 9 – Exemplo ilustrativo de descrição de cargos. 40
Figura 10 – Demonstração do entendimento dos *N-gramas*. 43
Figura 11 – Etapas da construção do léxico. 44

Figura 12 – Distribuição do número de palavras nos textos. 48
Figura 13 – Representação das descrições dos anúncios em nuvem de palavras. . . . 49
Figura 14 – Plataformas com a maior quantidade de anúncios de vagas de ciência de dados. 50
Figura 15 – Tempo de experiência requisitado nas vagas. 50
Figura 16 – Distribuição de anúncios de vagas por país e frequência. 52
Figura 17 – Distribuição de anúncios de vagas nos Estados Unidos e frequência. . . 52
Figura 18 – Distribuição total dos elementos do léxico identificados 56
Figura 19 – Tecnologias frequentemente encontradas nas descrições das vagas. . . . 57
Figura 20 – Linguagens de programação mais mencionadas nos anúncios. 58
Figura 21 – Conhecimentos mais recorrentes nas descrições das vagas. 58
Figura 22 – Habilidades interpessoais mais requisitadas nas vagas. 59
Figura 23 – Indicação dos requisitos predominantes nas descrições de vagas. 59

Figura 24 – Infográfico anual *Data Never Sleeps 11.0* 107

Figura 25 – Maneiras pelas quais as habilidades interpessoais podem contribuir para o sucesso do negócio. 109

LISTA DE TABELAS

Tabela 1 – Panorama dos trabalhos relacionados desta pesquisa. 32

Tabela 2 – Campos do conjunto de dados coletados e suas informações. 38
Tabela 3 – Métodos aplicados ao pré-processamento de anúncios de emprego e seus resultados. 42

Tabela 4 – Campos do conjunto de dados e sua quantidade. 47
Tabela 5 – Representação dos *N-gramas* construídos baseado nos títulos das vagas. 51
Tabela 6 – Representação dos tópicos relacionados análise de dados e negócios. . . 53
Tabela 7 – Representação dos tópicos relacionados a área da saúde e pesquisa. . . 53
Tabela 8 – Representação dos tópicos relacionados a finanças e modelagem. 54
Tabela 9 – Representação dos tópicos ligados ao entretenimento. 54
Tabela 10 – Representação dos tópicos relacionados a tecnologia e inovação. 55

LISTA DE ABREVIATURAS E SIGLAS

AML	*Aggregate Modeling Language*
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
ENIAC	Encontro Nacional de Inteligência Artificial e Computacional
ISO	*International Organization for Standardization*
LDA	*Latent Dirichlet Allocation*
LSA	*Latent Semantic Analysis*
LSI	*Latent Semantic Indexing*
NLTK	*Natural Language Toolkit*
NMF	*Non-Negative Matrix Factorization*
POS	*Part of Speech*
SBSI	*Simpósio Brasileiro de Sistemas de Informação*
TF-IDF	*Term Frequency–Inverse Document Frequency*
URL	*Uniform Resource Locator*
WI-IAT	*Web Intelligence and Intelligent Agent Technology*

SUMÁRIO

1 Introdução 19
1.1 Contextualização 19
1.2 Justificativa 23
1.3 Objetivos 26
1.3.1 Objetivo Geral 26
1.3.2 Objetivo específicos 26
1.4 Contribuições e impactos do estudo 27
1.5 Organização do trabalho 27

2 Trabalhos Correlatos 29

3 Materiais e Métodos 35
3.1 Entendimento do negócio 36
3.2 Entendimento e preparação dos dados 37
3.3 Modelagem 42
3.4 Avaliação e implementação 45
3.5 Tecnologias Utilizadas 46

4 Resultados 47

5 Considerações finais 61
5.1 Contribuições técnicas 62
5.1.1 Ameaças a validade do estudo 62
5.1.2 Trabalhos Futuros 62

Referências 65

APÊNDICES 70

APÊNDICE A Artigo publicado no *SBSI* 2024 73

APÊNDICE B Artigo publicado no *WI-IAT* 2023 83

APÊNDICE C Artigo publicado no *ENIAC* 2023 91

APÊNDICE D Infográfico anual *Data Never Sleeps 11.0* - Imagem original . 107

APÊNDICE E Maneiras pelas quais as habilidades interpessoais podem contribuir para o sucesso do negócio - Imagem original 109

1 INTRODUÇÃO

Este capítulo visa apresentar os temas fundamentais que direcionaram a realização deste estudo, para isso o mesmo está estruturado da seguinte forma: contextualização de assuntos relevantes, justificativa, objetivos propostos, impactos esperados e por fim a organização deste trabalho.

1.1 CONTEXTUALIZAÇÃO

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; 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; FONSECA et al., 2020). A Figura 24 (apêndice D) apresenta uma perspectiva do grande número de dados gerados na internet a cada minuto, demostrando como os dados estão presentes no cotidiano da sociedade.

Na última década, tem havido um desejo crescente entre as organizações de formar áreas analíticas para terem colaboradores que tenham capacidades de exploração, interpretação e análise de dados (HERNÁNDEZ, 2021). Sendo assim, as mudanças dinâmicas nos mercados de trabalho impulsionadas pela digitalização, automatização, robotização e transição verde exigem uma resposta adequada dos principais intervenientes para mitigar a escassez de competências (NAPIERALA; KVETAN, 2023). Por isso, a demanda por uma nova geração de profissionais qualificados em dados, análises, aprendizado de máquina e inteligência artificial requer uma resposta necessária tanto do ensino superior quanto do desenvolvimento da força de trabalho (MILLER; HUGHES, 2017).

No que tange ao escopo deste estudo, dados obtidos por meio do *Google Trends*[1] que mostram o interesse de pesquisa por meio do valor de popularidade de um termo (em um intervalo de 0 a 100) 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

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

Figura 1 – Infográfico anual *Data Never Sleeps 11.0*

Fonte: Adaptado de (Domo, 2023).

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. Diante disso, educar e treinar especialistas em ciência de dados requer um novo modelo, refletindo em seu design todo o ciclo de vida dos dados em pesquisa, domínios da indústria e a exigência de habilidades (DEMCHENKO et al., 2016). Outro aspecto crítico é a dispersão das informações de vagas de emprego.

Neste contexto, a natureza e a organização do trabalho estão experimentando mudanças em escalas sem precedentes. A pós-pandemia afetou a economia, demografia, além de avanços tecnológicos (SCULLY-RUSS; TORRACO, 2020). Por conseguinte, os mercados de trabalho globais demandam novos perfis de negócios e força de trabalho (JAGANNATHAN; RA; MACLEAN, 2019). Logo, as vantagens introduzidas pelas inovações tecnológicas são inúmeras e prometem trazer benefícios efetivos para as organizações, com várias reconfigurações de negócios (CARVALHO; SORCI; FIGUEIREDO, 2020). Frente a

Figura 2 – Tendência de buscas no Google Trends pelos termos "*big data*" e "*data science*".

Fonte: *Google Trends.*

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).

A rápida adoção da Internet revolucionou muitos aspectos da vida, desde padrões de interação às oportunidades de emprego e à forma como as procuramos (ALJOHANI et al., 2022). Atualmente existem muitos sites dedicados ao recrutamento e divulgação de empregos como consequência da digitalização do mercado de trabalho (KHAOUJA; KASSOU; GHOGHO, 2021). McDonald, Damarin e Membrez-Weiler (2023) indicaram os intermediários no mercado de trabalho digital fazendo a seguinte classificação: conectores que são tecnologias que ligam trabalhadores e empregadores; curadores são plataformas de contratação automatizadas; e misturadores que são mídias sociais que têm como objetivo a conexão e curadoria ao mesmo tempo. Assim, o recrutamento e a seleção online podem ser uma ferramenta valiosa para empresas que buscam atingir um grupo maior de candidatos e agilizar o processo de recrutamento (ARMAN, 2023).

Os anúncios de emprego disponíveis on-line fornecem uma rica fonte de dados detalhados e em tempo real sobre as qualificações e habilidades buscadas pelos empregadores em todos os setores, ocupações e locais (BRÜNING; MANGEOL, 2020). Sendo assim, analisar dados baseados na web sobre a demanda e a oferta do mercado de trabalho pode enriquecer a compreensão ao nível micro de uma série de questões, como mudanças ocupacionais, salários, condições de trabalho e requisitos de habilidades (KUREKOVA; BEBLAVY; THUM, 2014). 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; GONZÁLEZ, 2021).

Nos últimos anos, temos testemunhado um aumento nas quantidades de dados textuais digitais disponíveis, abrindo oportunidades para investigação por meio de novos canais comunicacionais (HASSANI et al., 2020). Esses textos incluem conteúdos online de jornais e redes sociais, comunicados de imprensa de empresas, análises de utilizadores sobre experiências e produtos, bem como artigos e discursos científicos, dentre outros (ANTONS et al., 2020). 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. Destarte, técnicas de mineração de texto têm se tornado essenciais no apoio à descoberta de conhecimento (SINOARA; MARCACINI; REZENDE, 2021; GURCAN; CAGILTAY, 2019; TERNIKOV, 2022).

Conforme apontado acima, ocorreram diversos impactos no mercado de trabalho, incluindo avanços tecnológicos e o crescimento da quantidade de dados em diferentes contextos. Neste caso, é crucial ter conhecimento transparente sobre as competências necessárias para lidar com os desafios futuros (FORUM, 2020). Dessa forma, a análise de vagas levanta questões como:

- Quais são os requisitos mais frequentes nas vagas?
- Quais habilidades técnicas e comportamentais são mais demandadas?
- Quais métodos podem ser utilizados para obter resultados nessas análises?
- Quais são os interessados nessa pesquisa?

Visando responder a estas perguntas e dado a relevância do domínio de aplicação exposto, no presente estudo foram coletados em torno de 14.000 anúncios de vagas de diversas plataformas. Ao longo da pesquisa são utilizadas boas práticas de ciência de dados para se realizar a análise exploratória, modelagem e validação dos *insights* obtidos. Por meio dessa análise obteve-se resultados que identificaram tendências da área, como cargos mais prevalentes, requisitos de experiência, habilidades técnicas desejáveis (*hard skills*) e habilidades interpessoais (*soft skills)* em alta. 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 aquecido.

1.2 JUSTIFICATIVA

O Forum (2020) destaca que a pandemia do COVID-19 e a recessão global aceleraram as transformações no mercado de trabalho. A adoção de tecnologias como computação em nuvem, *big data*, e-commerce, criptografia e inteligência artificial tornaram-se prioritárias. Segundo Battista, Grayling e Hasselaar (2023), há uma demanda crescente por requalificação devido às lacunas de habilidades e competências 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. Se devidamente recuperado e analisado, o enorme número de ofertas de emprego disponíveis hoje em portais de emprego pode fornecer informações detalhadas e valiosas sobre a dinâmica e as tendências do mercado de trabalho na Web (BOSELLI et al., 2018). Ademais, Khaouja, Kassou e Ghogho (2021) pontuam que, ao disponibilizar essas informações, diversas organizações podem tomar decisões mais embasadas e implementar reformas orientadas por dados.

Neste sentido, ZIPRECRUITER (2023) entrevistou 2.000 empregadores sobre a contratação baseada em habilidades e concluiu que essa estratégia é amplamente adotada por empresas de todos os tamanhos e setores. Além disso, constatou-se que há uma carência de habilidades sociais importantes, tais como gestão de tempo (47%), profissionalismo (43%) e pensamento crítico (39%). À medida que as necessidades empresariais continuam a evoluir, a lacuna de competências só aumentará com o tempo, a menos que os empregadores e os membros das suas equipes se concentrem na melhoria de competências e na requalificação (LINKEDIN, 2023b). A contratação baseada em habilidades pode fornecer um mecanismo para as empresas fecharem importantes lacunas de competências e concorrência num mercado de trabalho apertado (FULLER et al., 2022). A Figura 3 ilustra as habilidades exigidas globalmente enfatizando as *soft skills* mais procuradas e as *hard skills* classificadas como difíceis, essas informações indicam meios para entender como andam as necessidades empresariais.

Habilidades interpessoais impulsionam resultados de negócios (ECONOMICS, 2017). Por exemplo, a Figura 4 apresenta as conexões entre o sucesso do negócio e algumas *soft skills* (apêndice E). Diante disso, as conclusões do inquérito da MCKINSEY & COMPANY (2020) apontam que 44% dos inquiridos afirmam que as suas organizações enfrentarão lacunas de competências nos próximos cinco anos e outros 43% reportam lacunas de competências existentes. No mais, os profissionais de RH relatam dificuldade em recrutar candidatos que tenham as competências interpessoais necessárias para um mundo automatizado (MCKINSEY & COMPANY, 2019). Em vista disso, a lacuna de competências interpessoais prejudica o desempenho geral e afeta a retenção e o crescimento de um profissional, afetando negativamente a organização (DUBEY; PAUL; TEWARI,

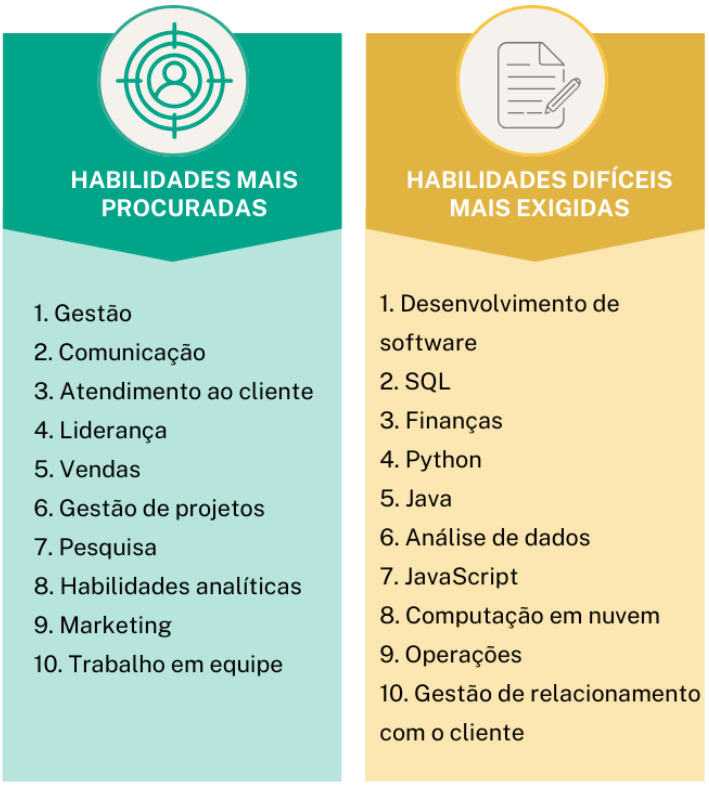


Figura 3 – Habilidades procuradas globalmente.

Fonte: Adaptado de (LINKEDIN, 2023b).

2022).

De acordo com LINKEDIN (2023c), os líderes de recrutamento enfrentam desafios significativos devido à economia instável e as mudanças no mercado de trabalho. As informações globais sobre o mercado de recrutamento SKYQUEST (2023) indicam que o mercado global de recrutamento online está em crescimento, impulsionado pela transformação digital e pela utilização de plataformas avançadas. As mudanças provocadas por tendências como a globalização, a evolução tecnológica, a economia baseada no conhecimento e a velocidade destas mudanças forçam as empresas a criar formas mais inovadoras de fazer negócios (DEDE, 2020). Diante disso, as organizações não podem existir sem pessoal para completar o ciclo de produção. Funcionários bem selecionados e bem colocados não apenas contribuem para o funcionamento eficiente da organização, mas também oferecem um potencial significativo para futuras substituições (GHAZZAWI; ACCOUMEH, 2014).

Em relação a construção de carreira, a empregabilidade é uma preocupação fundamental para os estudantes e é de importância crescente para as universidades com a inclusão de resultados de empregabilidade que afetam o desempenho nas tabelas classificativas e no quadro de excelência de ensino (BRADLEY; QUIGLEY; BAILEY, 2021). A progressão na carreira não se limita a um movimento linear entre um pequeno número de empregadores,

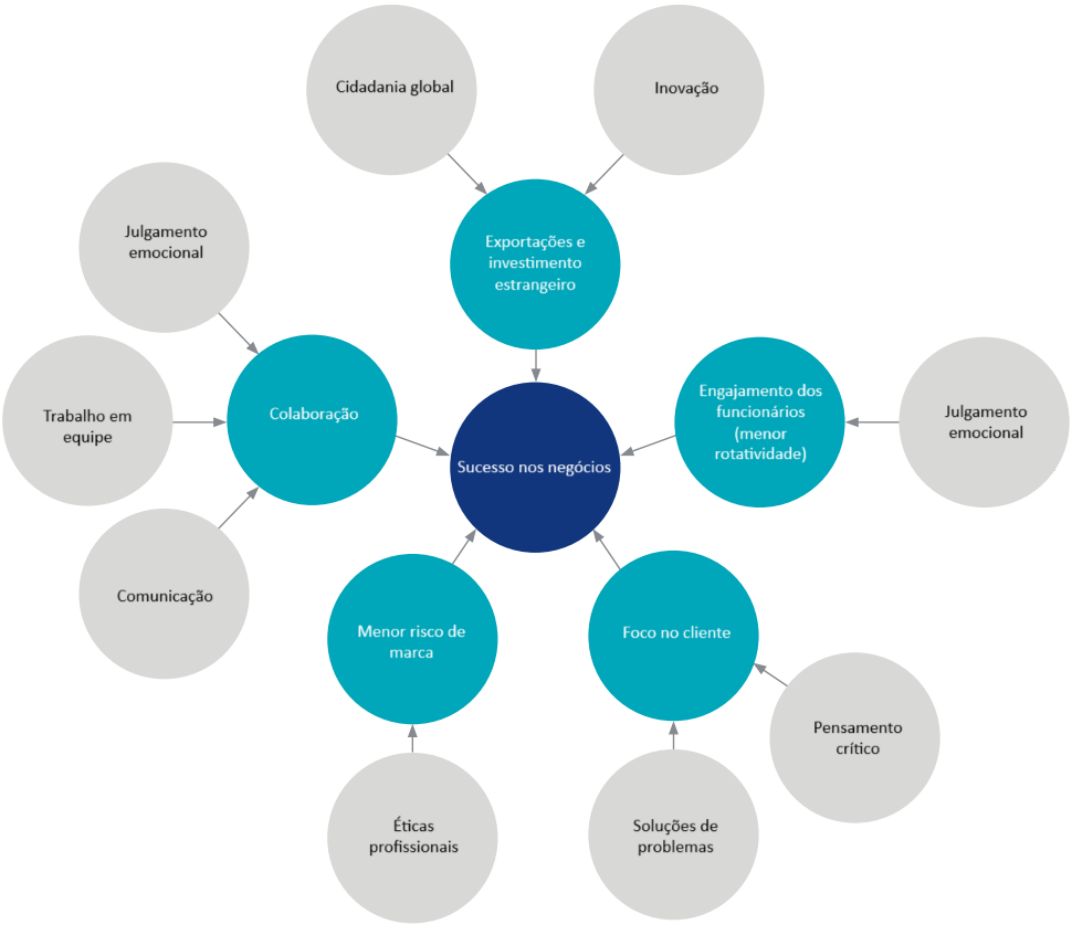


Figura 4 – Maneiras pelas quais as habilidades interpessoais podem contribuir para o sucesso do negócio.

Fonte: Adaptado de (ECONOMICS, 2017).

mas é definida por indivíduos que assumem a responsabilidade pela autogestão das suas carreiras no meio de uma riqueza de possibilidades, exigindo-lhes que sejam adaptáveis, proativos e planeiem o sucesso profissional (JACKSON; TOMLINSON, 2020). Além disso, a seleção de um caminho adequado deve ser calibrada de acordo com o potencial de desenvolvimento e os recursos disponíveis em cada etapa da carreira escolhida (HASSAN et al., 2022). Muitos Geração Z e *Millennials* tomam decisões de carreira com base em seus valores e desejam ter autonomia para impulsionar mudanças em suas organizações (DELOITTE, 2023). O relatório de trabalho assíncrono MIRO (2023) destaca que 64% da Geração Z expressa preocupação em causar incômodo aos colegas de trabalho ao formular perguntas de maneira assíncrona. Levando em consideração que em 2025, a Geração Z poderá compreender mais de um quarto da força de trabalho global MCKINSEY & COMPANY (2023), é de extrema importância preocupar-se com esta geração para assim atrair e reter talentos, além de trabalhar os desafios das habilidades comportamentais.

Frente às transformações do mercado de trabalho e o surgimento de novas profissões e reconfigurações de outras, sua análise é de extrema relevância. Pois facilita a identificação

e o estudo sistemático de noções relacionadas ao fator humano e seu papel no ambiente profissional, como a expertise, as competências, as habilidades sociais etc (PAPOUTSOGLOU et al., 2019). A análise de anúncios de emprego é uma tarefa relativamente nova e vem sendo recorrente para evidenciar as demandas do mercado de trabalho, avaliar a concordância ou discordância de habilidades e prever possíveis tendências a longo prazo.

Considerando o que foi levantado, fica evidente que o trabalho preenche uma lacuna patente, contribuindo efetivamente para a sociedade, especialmente diante da transformação tecnológica e das lacunas de competências. Sendo assim, espera-se que esta pesquisa contribua para uma compreensão mais aprofundada dos anúncios de vagas, gerando descobertas que possam auxiliar na tomada de decisões baseadas em dados.

1.3 OBJETIVOS

A luz do que foi discutido sobre a expansão da era dos dados, há uma grande demanda por profissionais capacitados, levando em conta a alta em dados relacionados às vagas de emprego. A seguir apresentam-se o objetivo geral do trabalho, seguido dos objetivos específicos.

1.3.1 OBJETIVO GERAL

O presente estudo tem como objetivo analisar ofertas de vagas relacionadas à Ciência de Dados utilizando métodos de mineração de textos a fim de identificar as tendências requeridas pelos empregadores quando à conhecimentos, tecnologias, linguagem de programação e habilidades comportamentais.

1.3.2 OBJETIVO ESPECÍFICOS

1. Realizar análise exploratória dos dados para obter conhecimento inicial para a manipulação dos dados de anúncios de emprego;
2. Construção do léxico do domínio de ciência de dados, permitindo a identificação e categorização das habilidades e competências requeridas;
3. Identificar quais são os requisitos mais comuns encontrados nas descrições de anúncios de vagas;
4. Identificar e visualizar as habilidades técnicas desejáveis (*hard skills*) e habilidades interpessoais (*soft skills)* em alta;
5. Avaliar a viabilidade de reprodutibilidade do estudo com outras áreas do conhecimento;

6. Documentar todo o processo das análises para a disponibilização dos códigos do *framework*.

1.4 CONTRIBUIÇÕES E IMPACTOS DO ESTUDO

A análise do mercado de trabalho suscita um grande interesse em diversas esferas, sobretudo em uma área em constante expansão como a ciência de dados. Dessa forma, esta pesquisa é direcionada para encontrar conhecimento em vagas de emprego especificamente voltadas para ciência de dados. Por fim, espera-se que os resultados obtidos colaborem com as seguintes vertentes:

- Geração de conhecimento para diversas esferas interessadas nessas análises como:
  - Os alunos e/ou pessoas em transição de carreira para o entendimento das tendências atuais visando a entrada no mercado de trabalho;
  - As instituições de ensino superior para oferecer treinamento ajustado as necessidades do mercado, principalmente para especializações e cursos de extensão considerando a curta duração das formações e para os docentes possíveis atualizações relacionadas tecnologias em alta;
  - Empresas para treinamentos e acompanhamento de tendências e planejamento da requalificação de seus colaboradores, e, por fim;
  - Formuladores de políticas, considerando que a inteligência artificial e ciência de dados são consideradas áreas estratégicas para o desenvolvimento nacional.
- Contribuir para a realização de análises com dados do mercado de trabalho, utilizando métodos e técnicas para extrair conhecimento e, dessa forma, potencializar as pesquisas relacionadas;
- Contribuição ao corpo do conhecimento relacionado a análise de vagas em ciência de dados, estimulando assim, pesquisas na temática abordada.

1.5 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado de forma que no Capítulo 2 são apresentados os trabalhos relacionados. No Capítulo 3 é descrito o processo de condução da pesquisa. No Capítulo 4 são apresentados os resultados obtidos e, por fim, o Capítulo 5 apresenta as considerações finais deste trabalho, bem como são discutidas as limitações do estudo e trabalhos futuros são propostos.

2 TRABALHOS CORRELATOS

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.

Diante da necessidade de acompanhar as tendências do mercado de trabalho, os pesquisadores começaram a utilizar conteúdos de fontes digitais, em Papoutsoglou et al. (2019) foi apresentado um estudo de mapeamento sistemático sobre fontes digitais voltadas para o mercado de trabalho de engenharia de *software* oferecendo uma classificação abrangente das questões relacionadas tendo como objetivo identificar fontes apropriadas para as análises dos interessados. Assim, foram selecionados 86 estudos visando identificar: fontes digitais utilizadas para análise do mercado de trabalho; tipos de habilidades que examinam; métodos usados para utilizar o conteúdo digital bruto; objetivos para os quais todo estudo primário é conduzido; beneficiários (partes interessadas) dos resultados; e tendências temporais para todos os itens acima. Por fim, é indicado cenários hipotéticos que poderiam contribuir para o planejamento de novos estudos pela academia e/ou pela indústria.

Seguindo a linha de mapeamento sistemático, Wu et al. (2022) tiveram como objetivo rever a investigação atual sobre a identificação de competências a partir de anúncios de emprego e discutir possíveis direções de investigação futura. Para isso, foram analisados 108 artigos sobre o tema, direcionando para avaliar e classificar os trabalhos com o intuito de identificar bases de competências utilizadas, tipos de competências extraídas e os métodos de identificação.

Silveira et al. (2020) posicionaram seus estudos em mapear e analisar as competências comportamentais e competências técnicas de cientistas de dados com uma abordagem descritiva, utilizando uma abordagem quanti-quali. A pesquisa descritiva teve como alvo retratar um fenômeno e seus aspectos. Uma pequena amostra de dados foi obtida por meio do *LinkedIn* para o entendimento do cenário das vagas, onde foi possível identificar o que as empresas buscam ao nível de escolaridade, indicação de perfil para candidatos a vagas e principalmente um direcionamento sobre essa profissão para a indústria quanto para a literatura.

Guimarães, Júnior e Freitas (2022) seguiram também com a análise de anúncios de emprego de ciência de dados no Brasil. No desenvolvimento do estudo, os dados utilizados foram extraídos de fontes diversas, a saber: *LinkedIn*, *Indeed*, *Infojobs*, Catho, Empregos e Vagas; adotaram-se métodos de mineração de texto como *N-grama* o qual serviu para

a representação dos termos frequentes e expressões composta por dois e três termos, modelagem de tópico *Latent Dirichlet Allocation* (LDA) para extrair os temas dominantes nos anúncios e agrupamento utilizando o algoritmo *K-means* que busca determinar grupos em uma coleção de documento pondo cada documento em um grupo. A intenção da análise é significativa principalmente para o entendimento dos requisitos recorrentes e validação para o uso de técnicas de mineração de texto em anúncios de emprego.

Gurcan e Cagiltay (2019) propuseram uma metodologia semi automática para analisar o conteúdo de anúncios de emprego *online* voltados para engenharia de *software* de *big data*. Os anúncios de emprego foram consultados pelo site *Indeed* e o método de modelagem de tópicos utilizado foi o LDA para apontar os domínios de conhecimento e conjuntos de habilidades. O método proposto pode ser melhorado para analisar automaticamente o conteúdo de outros anúncios de emprego focado em atualizações de acordo com as novas exigências do setor.

Debao, Yinxia e Min (2021) buscaram entender as características dos requisitos para vagas de *big data* a partir dos aspectos da cidade, salário, histórico educacional e experiência. 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 em comparar vários métodos usados na literatura para extrair requisitos de habilidade do texto de anúncios de emprego. Os autores utilizaram três abordagens de cima para baixo (*top-down)*, 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 mostra o desempenho mais elevado, 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 da modelagem de 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 técnicas de mineração 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*. A pesquisa forneceu ainda *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.

Em Alibasic et al. (2022) os autores 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.

Em Akkol, Olucoglu e Dogan (2023) o estudo foi direcionado as oportunidades de emprego na Turquia para profissionais de engenharia de computação e sistemas e sistemas de informação de gerenciamento. O objetivo foi identificar competências em vagas de emprego extraídas do *LinkedIn*, para isso a técnica aplicada foi o *Latent Semantic Analysis* (LSA). Os resultados revelaram que as ofertas de emprego refletem uma variedade de características em termos de competências técnicas e sociais. Por fim, os autores destacaram que para avaliar padrões globais a expansão da escala de busca das vagas é um caminho promissor levando em consideração que este estudo de caso se concentrou em apenas um local.

Em Umamaheswaran et al. (2023) os autores examinaram as habilidades que os empregadores valorizam, analisando as descrições de cargos para desvendar as expectativas da indústria em relação ao perfil de *Business Analytics*. Foi empregada uma abordagem de mineração de texto para avaliar a importância de diferentes habilidades e investigar agrupamentos de habilidades nas funções de análise de negócios. Neste estudo de caso a fonte de dados foi o *Indeed* com raio de busca sendo a Índia e a técnica utilizada foi o Bertopic. A principal contribuição desta análise foi a identificação de competências e os conceitos que constituem as capacidades fundamentais dos profissionais da área. Os autores destacaram que pelo foco da análise ser em uma região específica não se pode fazer generalizações globais sobre os achados.

A Figura 5 ilustra a cobertura temporal dos métodos utilizados nos trabalhos correlatos. Em toda a linha temporal observou-se que a modelagem de tópicos foi amplamente utilizada, destaca-se o uso da abordagem mais tradicional no caso o LDA. Nos estágios iniciais eram aplicadas apenas um método para atingir os objetivos, no entanto, à medida que foi passando, observou-se o uso de duas ou mais formas para extrair conhecimento dos anúncios de emprego, solidificando o entendimento de técnicas complementares para uma melhor interpretação dos resultados além de também haver o acompanhamento do estado

Tabela 1 – Panorama dos trabalhos relacionados desta pesquisa.

Trabalho	**Métodos**	**Fonte dos dados**	**Área alvo**
(GURCAN; CAGILTAY, 2019)	LDA	Indeed	Engenharia de software de big data
(GURCAN, 2019)	LDA	Sem informação	Big data
(SILVEIRA et al., 2020)	Abordagem quanti-quali	LinkedIn	Ciência de dados
(DEBAO; YINXIA; MIN, 2021)	K-mens	zhaopin	Big data
(ALIBASIC et al., 2022)	LSI, LDA e NMF	Sites diversos	Petroléo e gás
(AO et al., 2022)	Dicionário e LDA	Sem informação	Sem informação
(TERNIKOV, 2022)	NLP, Agrupamento Hierárquico e mineração por associação	HeadHunter	Tecnologia da Informação
(GUIMARÃES; JÚNIOR; FREITAS, 2022)	N-grama, LDA, K-mens	LinkedIn, Indeed e Infojobs,Catho, Empregos e Vagas	Ciência de dados
(AKKOL; OLUCOGLU; DOGAN, 2023)	LSA	LinkedIn	Engenharia de Computação e Sistemas de Informação de Gerenciamento
(UMAMAHESWARAN et al., 2023)	Bertopic	Indeed	Business Analytics
(PAPOUTSOGLOU et al., 2019)	Mapeamento Sistemático	Bibliotecas digitais	Engenharia de software
(WU et al., 2022)	Mapeamento Sistemático	Bibliotecas digitais	Sem informação

Fonte: Própria autora.

da arte para a modelagem de tópicos como o uso do Bertopic.

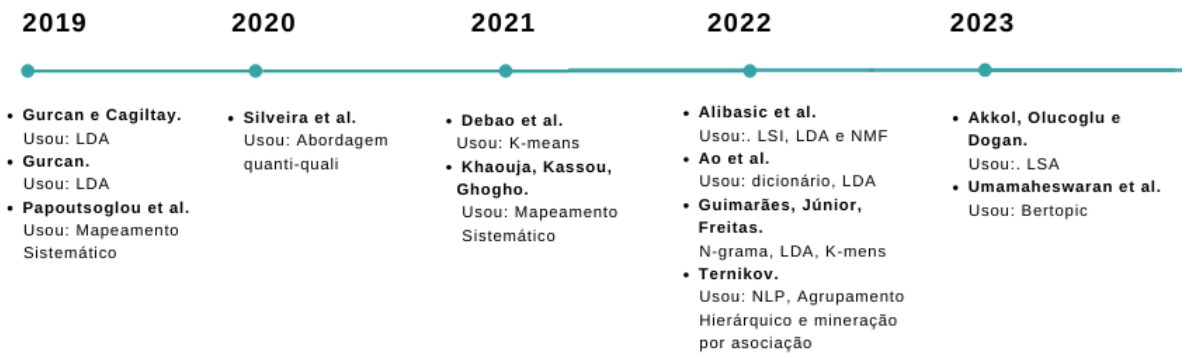


Figura 5 – Evolução temporal dos trabalhos relacionados.

Fonte: Própria autora.

A Tabela 1 sintetiza as informações relacionadas aos trabalhos correlatos indicando os métodos, fonte dos dados e área-alvo. Por meio da avaliação dos trabalhos verificou-se que a aplicação da Mineração de Texto em anúncios de emprego tem o potencial de extrair 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 e técnicas adicionais 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, este trabalho 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 utilizando o Bertopic visando uma melhor interpretabilidade dos resultados; e iii) a construção de um léxico específico para a área para a contagem de habilidades. 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 o tamanho da amostra em relação à literatura e a adoção de uma estratégia de validação dos dados seguindo o princípio *human-in-the-loop*. Para a amostra, foi considerada a utilização de diferentes plataformas de vagas para uma análise mais ampla. Em relação à forma de validação dos resultados foi usada a abordagem *grounded-theory*, seguindo uma estratégia baseada no *human-in-the-loop* (WU et al., 2022).

3 MATERIAIS E MÉTODOS

Este capítulo relata os procedimentos metodológicos adotados, o detalhamento sobre a obtenção dos dados, os processos e métodos utilizados para a construção das análises e por fim as tecnologias empregadas.

Para esta pesquisa o processo *Cross-Industry Standard Process for Data Mining* (CRISP-DM) foi escolhido, pois é amplamente reconhecido e valorizado na área. O CRISP-DM ainda é um padrão de fator na mineração de dados (SCHRÖER; KRUSE; GÓMEZ, 2021). O método permite um controle eficiente das etapas de análise (WIRTH; HIPP, 2000). Este processo é cíclico, o que permite maior flexibilidade para mudanças no processo do desenvolvimento. Guimarães, Júnior e Freitas (2022) utilizaram esse processo em sua análise de anúncios de vagas de emprego.

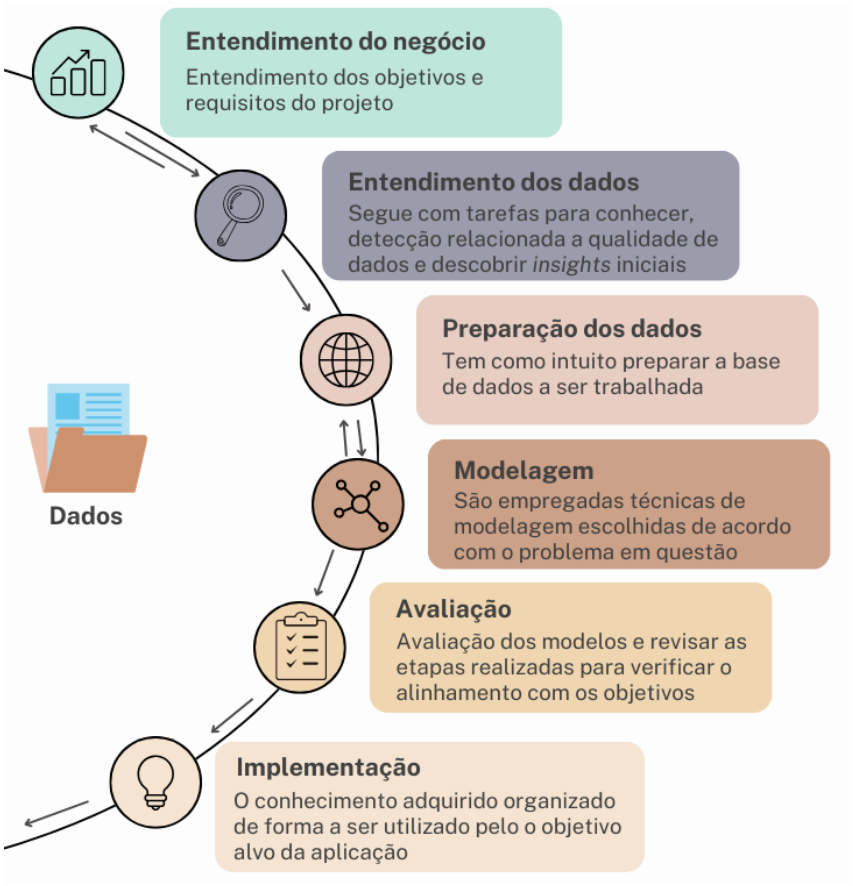


Figura 6 – Fases do método CRISP-DM.

Fonte: Própria autora.

O CRISP-DM é composto pelas etapas apresentadas na Figura 6, a saber: entendimento do negócio, direcionada para a compreensão dos objetivos e as necessidades do projeto; entendimento dos dados, esta etapa concentra-se em conhecer e explorar

inicialmente os dados para entender sua qualidade e características; preparação dos dados, tem como finalidade aperfeiçoar a base de dados a ser trabalhada e pode envolver diversos processos como limpeza dos dados, transformação e outros; modelagem, são empregadas técnicas de modelagem escolhidas conforme o problema em questão; avaliação, onde avaliam-se os modelos aplicados para a verificação do alinhamento com os objetivos; e implementação, orientada para que o conhecimento adquirido seja organizado para ser utilizado pelo alvo da aplicação. As próximas subseções detalham as etapas do CRIPS-DM instanciadas para o presente trabalho e que foram apresentadas na Figura 6.

3.1 ENTENDIMENTO DO NEGÓCIO

No estágio inicial, realizou-se uma investigação preliminar para obter informações sobre a área-alvo do estudo de caso (ciência de dados), o papel dos profissionais dessa área, a importância da análise de anúncios de emprego e sua conexão com os objetivos desta pesquisa. Posteriormente, realizamos uma revisão abrangente na literatura tradicional para identificar estudos relacionados à extração de informações de vagas de emprego.

Além disso, foram feitas pesquisas na literatura cinza que abrange uma ampla gama de formatos, como relatórios, documentos de trabalho, documentos governamentais, entre outros, para obter compreensões e informações valiosas que podem enriquecer a pesquisa, auxiliar na tomada de decisão e ampliar o conhecimento em diversas áreas. As partes que conduziram esta etapa do desenvolvimento do trabalho, são:

1. Investigação preliminar foi conduzida tendo em vista os *roadmaps* de estudo da área, tendências, palavras-chave correlatas que dão indício das tecnologias, atividades realizadas, conhecimentos solicitados, habilidades técnicas e comportamentais voltadas para a área em questão;

2. Por literatura tradicional entendem-se os artigos técnico-científicos, os quais foram discutidos no Capítulo 2. Como resumo desta etapa, foram identificados aplicações em diferentes vertentes como *Big data*, petróleo e gás, Tecnologia da Informação e outros; além disso, utilizaram-se métodos nas análises como modelagem de tópicos, *N-grama*, agrupamento e processamento de linguagem natural;

3. Na literatura cinza, o Forum (2020) ressaltou mudanças na questão do trabalho relacionada a pandemia. Em Battista, Grayling e Hasselaar (2023), LINKEDIN (2023b), ZIPRECRUITER (2023) são destacadas as dimensões das habilidades como lacunas e a crescente contratação baseada em habilidades. Na questão do mercado para a área de dados, o *State of Data 2022*[1] foi de grande contribuição, pois é

[1] https://www.bain.com/pt-br/insights/state-of-data-2022_profissionais_dados/

um raio-x dos profissionais de dados do Brasil e apresenta resultados baseados nas respostas de profissionais da área;

4. Por meio da literatura cinza foi possível identificar que existem diversas plataformas que divulgam dados de vagas de emprego, sendo assim houve a necessidade de encontrar uma forma para a coletar essas informações de diferentes plataformas de vagas e a partir de pesquisas a plataforma selecionado foi o *Diffbot*[2]. A plataforma oferece serviços de extração e análise de dados da web. É possível coletar informações de diversas fontes na internet, incluindo notícias, artigos, produtos, mídias sociais e vagas de emprego. A Figura 7 apresenta o painel da plataforma utilizada.

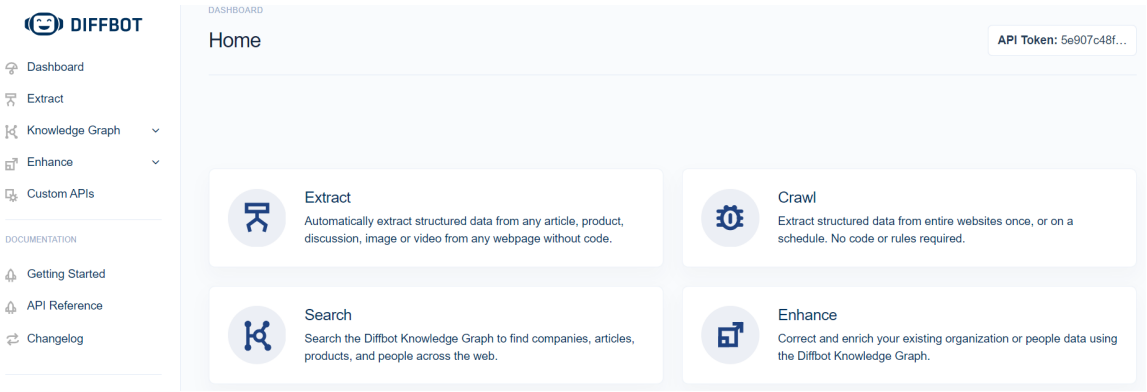


Figura 7 – Painel da plataforma *Diffbot*.

Fonte: Página do *Diffbot*.

A compreensão do negócio é de extrema importância para as conexões de todas as etapas da metodologia. Após uma análise das questões norteadoras da pesquisa, é passado para a próxima etapa de entendimento e preparação dos dados.

3.2 ENTENDIMENTO E PREPARAÇÃO DOS DADOS

A obtenção dos dados no *Diffbot* foi realizada consoante o fluxo apresentado na Figura 3.2.

Sendo assim, iniciou-se a seleção da opção *Search*, seguida pela escolha da categoria alvo para a pesquisa. Em seguida, foi realizado o ajuste do filtro (com opções disponíveis como título, texto e outros). Para a coleta das vagas neste trabalho, o filtro escolhido foi o título. Posteriormente, indicou-se a palavra-chave utilizada, que neste caso foi *"Data Scientist"*. Os dados foram coletados durante os meses de setembro de 2022, quando foram registrados 6.000 anúncios, e outubro de 2023, quando foram registrados 8.000

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

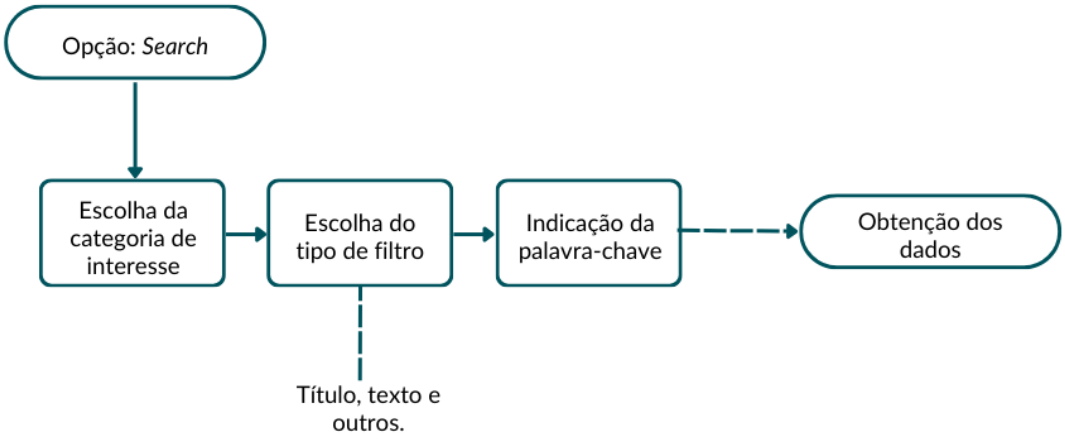


Figura 8 – Fluxo de trabalho para a coleta de dados na plataforma *Diffbot*.

Fonte: Própria autora.

Tabela 2 – Campos do conjunto de dados coletados e suas informações.

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

Fonte: Própria autora.

anúncios, resultando em um total de 14.000 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 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 2 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 de modo a identificar as informações potencialmente extraíveis. Diversos recursos

online, como o site *Indeed*[3], o blog da *Gupy*[4], o *LinkedIn*[5] e o *Glassdoor*[6], fornecem orientações e diretrizes para a criação de anúncios de emprego eficazes. Esses recursos 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 e compreender melhor as tendências e demandas do mercado de trabalho. A Figura 9 apresenta uma descrição fornecida pelo *LinkedIn* de uma descrição genérica de anúncio de emprego e seus elementos, o qual fez parte deste processo de entendimento da estrutura dos anúncios.

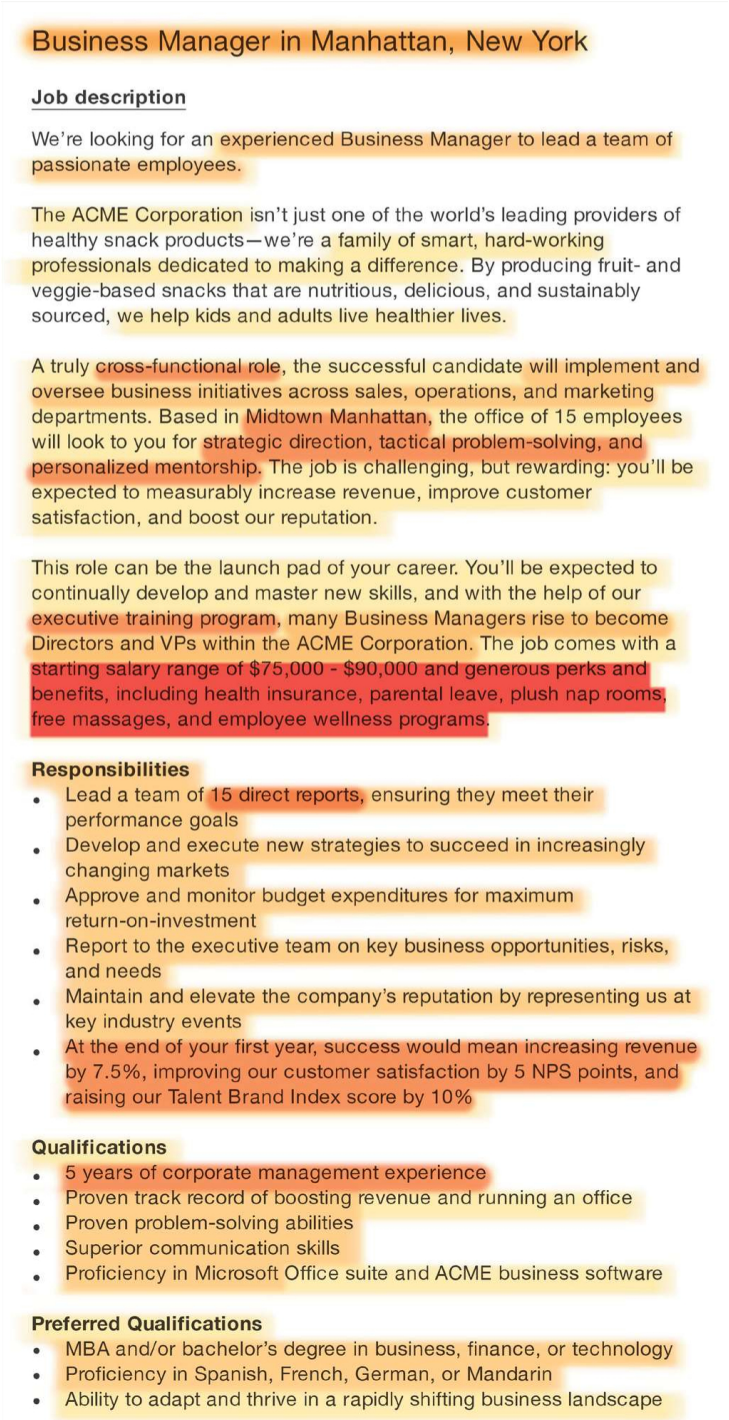
A partir do entendimento dos dados, foram selecionados os campos principais para análise, como o título e o texto dos anúncios. Outros campos, como URL, idioma e requisitos, foram manipulados de forma secundária. No caso das URLs, foi realizada uma operação de limpeza para remover o prefixo e o sufixo do caminho e assim obter a fonte do anúncio utilizando os pacotes *Urllib* [7] *e Publicsuffix2* [8]. Além disso, foram utilizadas expressões regulares para extrair informações sobre anos de experiência mencionados nos anúncios, tendo como recurso a biblioteca *Re*[9]. A expressão regular utilizada foi a seguinte: (?i)\b(?:\d+(?:\.\d+)?\s∗[−+]?\s∗\d∗(?:\.\d+)\s∗(?:year|yr|yrs|years)|\ d+(?:\.\d+)?\s∗[−+]\s(?:year|yr|yrs|years))\s∗(?:of)?\s∗(?:relevant)?\s∗experience\b.

Assim como, foi feito o reconhecimento de entidades nomeadas para identificar e classificar entidades que aparecem em textos em linguagem natural. O intuito foi identificar países e estados, utilizando a biblioteca *Spacy*[10] onde é passada a função *en_core_web_sm* para lidar com texto em inglês, assim para capturar essas informações é especificado o seguinte rótulo GPE para países e/ou estados. Após esta captura o passo seguinte é contabilizar a frequência das entidades para visualizar essas informações.

Além da frequência das entidades extraídas, houve a manipulação dessas informações para realizar a análise geográfica. É importante destacar que antes de contabilizar a frequência é feita a padronização de termos de siglas de países para os nomes dos mesmos. Para compor essa análise é necessário fazer a geocodificação que é a transformação de informações em texto de um local para coordenadas geográficas, no caso deste estudo países e estados.

Para ser possível visualizar as informações geográficas foi realizado um mapeamento

[3] https://uk.indeed.com/career-advice/career-development/job-advertisement
[4] https://www.gupy.io/blog/divulgacao-de-vagas
[5] https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/resources/pdfs/job-descriptions-ebook.pdf
[6] https://www.glassdoor.co.uk/employers/resources/9-things-every-job-ad-should-have/)
[7] https://docs.python.org/pt-br/3/library/urllib.html
[8] https://pypi.org/project/publicsuffix2/
[9] https://docs.python.org/pt-br/3/library/re.html
[10] https://spacy.io/



Business Manager in Manhattan, New York

Job description

We're looking for an experienced Business Manager to lead a team of passionate employees.

The ACME Corporation isn't just one of the world's leading providers of healthy snack products—we're a family of smart, hard-working professionals dedicated to making a difference. By producing fruit- and veggie-based snacks that are nutritious, delicious, and sustainably sourced, we help kids and adults live healthier lives.

A truly cross-functional role, the successful candidate will implement and oversee business initiatives across sales, operations, and marketing departments. Based in Midtown Manhattan, the office of 15 employees will look to you for strategic direction, tactical problem-solving, and personalized mentorship. The job is challenging, but rewarding: you'll be expected to measurably increase revenue, improve customer satisfaction, and boost our reputation.

This role can be the launch pad of your career. You'll be expected to continually develop and master new skills, and with the help of our executive training program, many Business Managers rise to become Directors and VPs within the ACME Corporation. The job comes with a starting salary range of $75,000 - $90,000 and generous perks and benefits, including health insurance, parental leave, plush nap rooms, free massages, and employee wellness programs.

Responsibilities

- Lead a team of 15 direct reports, ensuring they meet their performance goals
- Develop and execute new strategies to succeed in increasingly changing markets
- Approve and monitor budget expenditures for maximum return-on-investment
- Report to the executive team on key business opportunities, risks, and needs
- Maintain and elevate the company's reputation by representing us at key industry events
- At the end of your first year, success would mean increasing revenue by 7.5%, improving our customer satisfaction by 5 NPS points, and raising our Talent Brand Index score by 10%

Qualifications

- 5 years of corporate management experience
- Proven track record of boosting revenue and running an office
- Proven problem-solving abilities
- Superior communication skills
- Proficiency in Microsoft Office suite and ACME business software

Preferred Qualifications

- MBA and/or bachelor's degree in business, finance, or technology
- Proficiency in Spanish, French, German, or Mandarin
- Ability to adapt and thrive in a rapidly shifting business landscape

Figura 9 – Exemplo ilustrativo de descrição de cargos.

Fonte: (LINKEDIN, 2023a).

dos nomes (países e estados) para obter o código da *International Organization for Standardization* (ISO) para essa tarefa foram utilizados os seguintes recursos *pycountry*[11] e

[11] https://pypi.org/project/pycountry/

us[12]. Tendo essas informações é feita a visualização do mapa de frequência com *Plotly* [13].

O pré-processamento do conteúdo textual foi realizado baseando-se nos trabalhos Cirqueira et al. (2018), Gurcan e Cagiltay (2019), Debao, Yinxia e Min (2021) e consistiu nos seguintes passos:

- Remoção de duplicatas, esta é considerada uma tarefa trivial, que contou com o *id* e a *pageURL* para a verificar se o anúncio era único, onde se utilizou a função da biblioteca *Pandas*;
- Transformação do texto para caixa baixa, visando a uniformidade sem divergência pelas letras maiúsculas e minúsculas;
- Filtragem do idioma, na qual se optou por se trabalhar apenas com os textos em inglês, dado a sua maior frequência;
- Remoção de pontuação, acentuação, caracteres especiais e espaços excessivos. A remoção destes elementos que não contribuem diretamente tem uma melhora significativa na eficácia e eficiência para as etapas que procedem;
- Remoção de caracteres numéricos, essa parte foi empregada após a captura de elementos relacionados aos anos de experiência e também necessária após testes com a modelagem de tópicos, eliminando assim ruídos em seus resultados;
- Remoção de *stopwords*, no caso palavras consideradas irrelevantes, que, ao serem retiradas de um texto, não sofrem alterações significativas. O *Natural Language Toolkit*[14](NLTK) foi utilizado para a remoção das (*stopwords*) além de usar as palavras do NLTK, outras palavras foram adicionadas para serem retiradas como *"Job"*, *"tab"*, *"send"* e outras;
- Tokenização, o qual é o processo de transformar a *string* inteira (texto de entrada) em palavras singulares.

Para a filtragem do idioma, o campo *humanlanguage* da base de dados foi considerado para identificar quais os idiomas estavam presentes e a frequência dos mesmos. É importante mencionar que a filtragem do idioma considerou o entendimento dos dados, pois o inglês era o idioma mais frequente. Além do idioma predominante, outros foram identificados como francês, português, italiano, espanhol, russo, dinamarquês, coreano, entre outros. Sendo assim, os anúncios classificados com outros idiomas foram excluídos da base de dados, pois divergem dos critérios definidos. Essa decisão metodológica foi

[12] https://pypi.org/project/us/
[13] https://plotly.com/
[14] https://www.nltk.org/

Tabela 3 – Métodos aplicados ao pré-processamento de anúncios de emprego e seus resultados.

Método	**saída**
Texto original	*Minimum 6 years experience in a data science or advanced analytics role (with Masters' degree, minimum experience is 4 years)*
Trasformação para caixa baixa	*minimum 6 years experience in a data science or advanced analytics role (with masters' degree, minimum experience is 4 years)*
Remoção de pontuação, acentuação, caracteres especiais e espaços excessivos	*minimum 6 years experience in a data science or advanced analytics role with masters degree minimum experience is 4 years*
Remoção de caracteres numéricos	*minimum years experience in a data science or advanced analytics role with masters degree minimum experience is years*
Remoção de *stopwords*	*minimum years experience data science advanced analytics role masters degree minimum experience years*
Tokenização	*"minimum", "years", "experience", "data", "science", "advanced", "analytics", "role", "masters", "degree", "minimum", "experience", "years"*

Fonte: Própria autora.

tomada após testes com a tradução, entretanto, o tradutor acabava por traduzir também tecnologias, potencialmente enviesando os resultados.

Com os dados sem duplicações e apenas em inglês, foram conduzidos os passos do pré-processamento textual, na Tabela 3 são apresentados os métodos utilizados para o pré-processamento do conteúdo textual aplicado em um trecho extraído da base de dados.

3.3 MODELAGEM

Para identificar padrões nos anúncios de emprego utilizamos diferentes técnicas de análise como *N-grama*, Modelagem de tópicos e construção de um léxico do domínio. Foi adotada a abordagem de geração de *N-grama* para apresentar sequências de palavras formadas por um conjunto específico de n palavras. Na Figura 10 é demostrado um exemplo de *N-grama* aplicado na frase *"Natural Language Processing"*. Sendo assim, considerando o entendimento do domínio de aplicação descrito na Seção 3.1, a técnica se concentrou em bigramas e trigramas. As análises tiveram como enfoque os títulos das vagas, onde foi utilizada a representação baseada em *Term Frequency–Inverse Document Frequency* (TF-IDF), a qual é representada pela equação 3.1, para mostrar o quão importante são as palavras nos documentos, para essa representação utilizou-se a biblioteca *scikit-learn*[15].

[15] https://scikit-learn.org/stable/

$$W_{t,d} = t_{f,t,d} \cdot \log\left(\frac{N}{t_{ft}}\right) \tag{3.1}$$

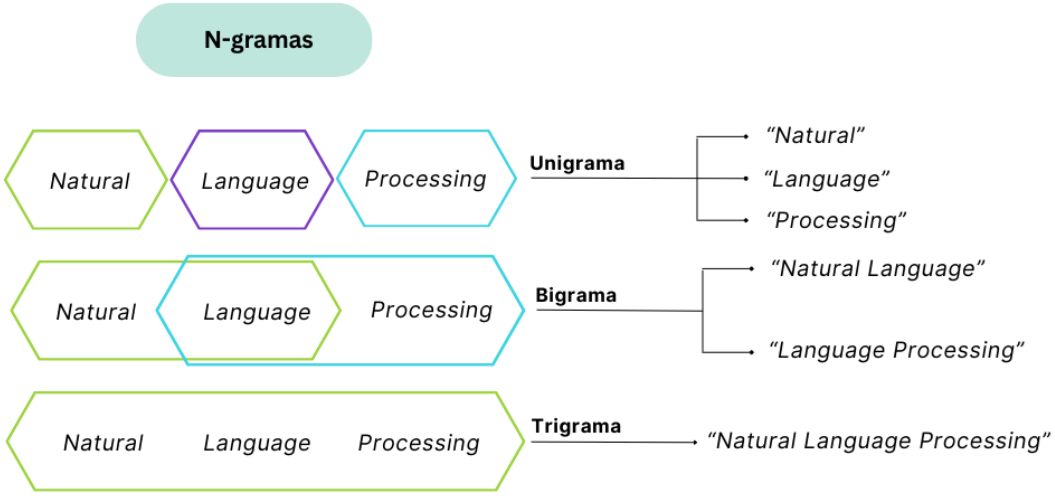


Figura 10 – Demonstração do entendimento dos *N-gramas*.

Fonte: Própria autora.

Este estudo 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; XIA; CHENG, 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). O TF-IDF baseado em classe tem o intuito de extrair as palavras mais importantes por cluster, o mesmo é representado da seguinte maneira pela equação 3.2, cada agrupamento é convertido em um único documento em vez de um conjunto de documentos. Em seguida, a palavra x é extraída de c, onde c é o agrupamento (*cluster*) criado anteriormente. O tf é uma representação baseada em classes, normalizada em $L1$ para levar em conta as diferenças nos tamanhos dos tópicos. Em seguida, utilizamos o logaritmo de um número médio de palavras por classe A dividido pela frequência da palavra x em todas as classes. Adicionamos mais um valor ao logaritmo para forçar os valores a serem positivos, o que resulta na representação em classe.

$$W_{x,c} = t_{f,x,c} \cdot \log\left(1 + \frac{A}{t_{fx}}\right) \tag{3.2}$$

No desenvolvimento da modelagem de tópicos, é realizado o treinamento, seguido pela geração de tópicos para análises. Os tópicos são compostos por palavras-chave que

refletem os principais temas nos documentos. A etapa seguinte implicou em uma análise qualitativa, na qual os tópicos foram avaliados e ajustados. O objetivo foi determinar um número de tópicos realista e facilmente interpretável por um especialista humano, seguindo o conceito de *human-in-the-loop*. Para a redução de tópicos optou-se por fazer esse processo após o treinamento, pois assim se tem noção dos tópicos criados automaticamente e quais são realistas, para isso o recurso utilizado do BerTopic foi o *topic_model.reduce_topics*.

Figura 11 – Etapas da construção do léxico.

Fonte: Própria autora.

Outra análise empregada foi a criação de um léxico de vagas de emprego, seguindo o fluxo apresentado na Figura 11, composta por quatro etapas, a saber: categorização, identificação de termos relevantes, refinamento e quantificação. Cada fase encontra-se detalhada a seguir.

1. **Categorização:** É inicialmente feita a separação dos eixos para a construção do dicionário. Esses eixos incluem *conhecimentos*, *tecnologias*, *linguagem de programação* e *habilidades comportamentais*.

2. **Identificação de termos relevantes:** 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, as palavras-chave utilizadas foram: *"platforms"*, *"platform"*, *"tools"*, *"software"*, *"knowledge"*, *"certification"*, *"plus"*, *"ability"*, *"required"*, *"programs"* e *"experience"*. Sendo que para as anotações das *habilidades comportamentais* houve um recurso adicional como a utilização da base de dados disponibilizada no trabalho de Calanca et al. (2019) para ampliar a formação do léxico.

3. **Refinamento:** Nesta etapa, é realizada uma padronização dos termos onde termos semelhantes são identificados em suas formas expressadas ao longo do texto. Por exemplo, em *"ai"* é feita uma substituição para *"artificial intelligence"* em todas as suas ocorrências, assim como em *"c++"* é feita a seguinte modificação *"c_plus_plus"*.

4. **Quantificação:** Após a construção do dicionário o mesmo é utilizado como recurso para identificação de informações, a qual segue a abordagem direcionada por CHAIBATE et al. (2019) que as palavras contidas no dicionário são relacionados a variáveis dicotômicas com dois valores distintos (0 e 1), sendo o valor 1 a indicação que a palavra se encontra no anúncio e o 0 indica o não aparecimento da palavra.

3.4 AVALIAÇÃO E IMPLEMENTAÇÃO

Para a avaliação optou-se pela utilização da abordagem embasada em *grounded theory*, pois esta abordagem qualitativa permite que os pesquisadores identifiquem processos explícitos e implícitos em seus dados (CHARMAZ; THORNBERG, 2021). No presente trabalho a abordagem foi conduzida de maneira que a(os) autora(es) fazem a leitura de uma amostragem relevante dos dados e cruzam com os resultados obtidos. Destaca-se também a realização de um cálculo amostral para determinar a amostra para ser feita a leitura dos anúncios, levando em consideração o tamanho da população que foi delimitado, o valor de 6.000 da primeira coleta de dados, um 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. Convém pontuar que a leitura dos 1.691 anúncios de vagas de emprego é complementar à leitura de entendimento de domínio de aplicação, que não seguiu metodologia estrita. Ou seja, a(os) autora(es) fizeram uma análise exaustiva das vagas visando uma maior confiabilidade do estudo. Outro aspecto pertinente foi o contato durante o processo com *tech recruiters*, mais especificamente, uma psicóloga especialista em recrutamento envolvendo tecnologia e um gestor responsável por gerenciar times. Esse contato possibilitou a validação de informações para direcionar partes das análises, a qual foram as seguintes:

- Contextualização da dinâmica da contratação e a pertinência da temática para pessoas de recursos humanos voltadas para área tecnológica;
- Validação dos eixos para a construção do léxico;
- Validação de partes das anotações do dicionário, e, por fim;
- Validação de *insights* relacionados a habilidades comportamentais.

Na fase de implementação destaca-se a contribuição veiculada por meio de artigos científicos, apresentações em eventos, relatório técnico para apresentar o *framework* experimental, além do presente manuscrito, provendo uma visão detalhada das etapas para a elaboração das análises e *insights* obtidos.

3.5 TECNOLOGIAS UTILIZADAS

Esta subseção é direcionada para apresentar as tecnologias utilizadas neste trabalho, abrangendo ambiente de implementação, linguagem de programação adotada e bibliotecas utilizadas.

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. Além disso, para a execução dos *notebooks* de análise, optou-se pelo ambiente de execução *Google Colaboratory*, devido às suas vantagens, como a facilidade de configuração e acesso aos recursos de processamento e armazenamento em nuvem. A versão utilizada foi o Colab Pro+, que oferece benefícios adicionais, como maior capacidade de memória e recursos computacionais aprimorados, possibilitando a análise eficiente de grandes volumes de dados.

No processo de preparação e modelagem dos dados, foram utilizadas várias bibliotecas para realizar diferentes tarefas. O *pandas* foi empregado para a manipulação dos dados, o NLTK foi utilizado para o pré-processamento de textos, em especial, para a remoção de palavras irrelevantes; para a manipulação das URLs foram utilizadas *Urllib Publicsuffix2*; nesta etapa também foram utilizadas expressões regulares implementadas na biblioteca (Re) para lidar com padrões específicos nos dados textuais; e para lidar com entidades nomeadas utilizou-se *Spacy*.

A representação dos dados textuais foi realizada utilizando a biblioteca *scikit-learn* que possibilitou a utilização da técnica TF-IDF para destacar a importância das palavras nos documentos e foi responsável pela geração dos bigramas e trigramas. Para a Modelagem de tópicos usou-se e *BerTopic*[16]. Para a visualização de dados utilizaram-se o matplotlib/pyplot[17], o *Seaborn*[18] ambos para a geração de gráficos, o *Plotly* que foi usado para a geração dos mapas e o *WordCloud*[19] que permitiu a criação de nuvem de palavras, um recurso baseado na frequência em que determinada palavra esta presente no texto, indicando uma visualização proporcional ao aparecimento dos termos.

[16] https://maartengr.github.io/BERTopic/index.html
[17] https://matplotlib.org/stable/tutorials/introductory/pyplot.html
[18] https://seaborn.pydata.org/
[19] https://pypi.org/project/wordcloud/

4 RESULTADOS

Este capítulo discorre sobre os resultados desta pesquisa mostrando a visualização dos dados, as plataformas em destaque das fontes de informação, visualização dos títulos das vagas, questão dos anos de experiência e modelagem de tópicos, conforme os procedimentos metodológicos apresentados no capítulo anterior.

Os dados de anúncios de empregos que compõem esta pesquisa são apresentados na Tabela 4 e apresentam a quantidade de cada campo. Conforme descrito na metodologia, os principais campos de interesse eram *Title* e *Text*, sendo assim, *requirements* e *tasks* foram considerados dados complementares. Das 13.785 instâncias que tinham o texto da vaga, foram removidos os dados duplicados e os que não estavam escritos em inglês, sendo assim, após a filtragem dos dados, foi realizada a análise de um total de 13.060 anúncios de emprego.

Baseado na exploração dos dados, o tamanho médio dos anúncios é de 481 palavras. A Figura 12, apresenta um histograma que demonstra a distribuição do número de palavras nos textos analisados. A visualização mostra a frequência com que as palavras aparecem, dando uma perspectiva clara sobre a variação no comprimento dos textos.

Tabela 4 – Campos do conjunto de dados e sua quantidade.

Campos	**Quantidade**
Id	14.000
Title	14.000
Text	13.785
humanlanguage	14.000
pageUrl	14.000
requirements	6.783
tasks	8.682

Fonte: Própria autora.

A Figura 13 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 13 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"*. Esses aspectos encontrados estão alinhados com as decisões de carreiras da Geração Z e *Millennials* como foi mencionado em (DELOITTE, 2023).

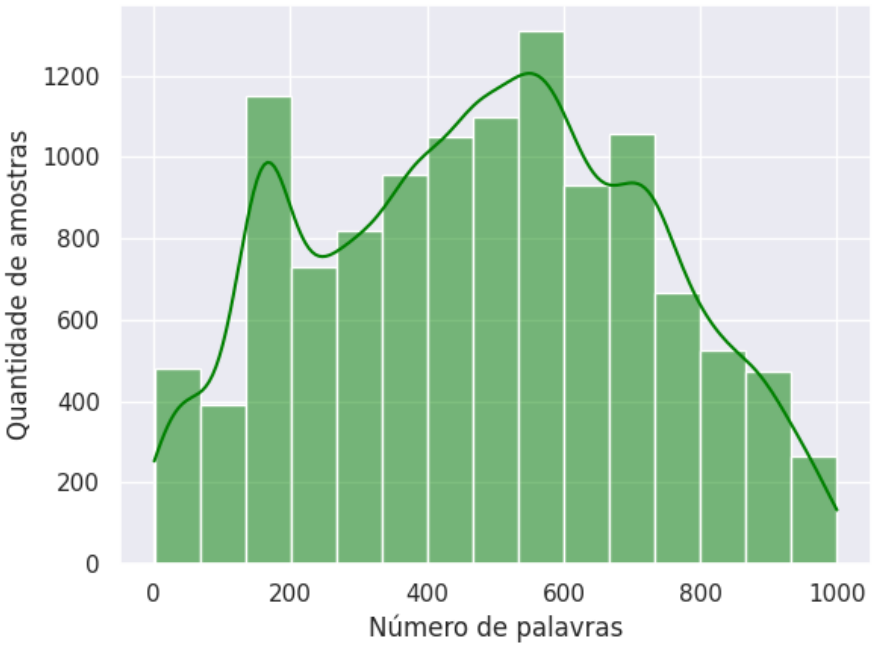


Figura 12 – Distribuição do número de palavras nos textos.

Fonte: Própria autora.

Para examinar as principais plataformas de anúncios de vagas, realizou-se a contagem da frequência dos *sites* com base na coluna *url*. Foram identificadas um total de 1.068 plataformas, dentre essas foi possível fazer a representação das dez plataformas mais frequentes observável na Figura 14. 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 INFOJOBS (2023), como o *LinkedIn*, *99jobs* e *Infojobs*, não estão entre as opções predominantes, apenas o *Indeed* encontra-se presente. Destarte, este achado de pesquisa tem o potencial de guiar profissionais que estão buscando colocação na área de ciência de dados no cenário internacional.

Os resultados da análise dos *N-gramas* nos títulos dos anúncios são apresentados na Tabela 5, a qual exibe 18 expressões comuns nos documentos. Pode-se observar diferentes posições destacadas como pela senioridade como *senior data scientist* indicando a necessidade de um alto nível de experiência para a posição e *junior data scientist* direcionada para profissionais iniciantes. Observam-se expressões direcionadas à coordenação na área a qual requer uma sólida experiência, como *Lead data scientist* e *Manager data scientist*. 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.

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

Figura 13 – Representação das descrições dos anúncios em nuvem de palavras.

Fonte: Própria autora.

na Figura 15, 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 773 vagas. Em seguida, temos as faixas de 3 anos, com 706 vagas, e 2 anos, com 615 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.

A análise geográfica dos anúncios de vagas (Figura 16) revelou uma distribuição significativa por diferentes lugares. Substancialmente, os Estados Unidos são os principais anunciantes com um total de 1.536 anúncios, indicando uma alta demanda por profissionais, isso também se deve ao fato do idioma prevalente da análise ser o inglês. Em segundo lugar, a Índia também apresenta uma presença significativa, com 586 anúncios. Outros

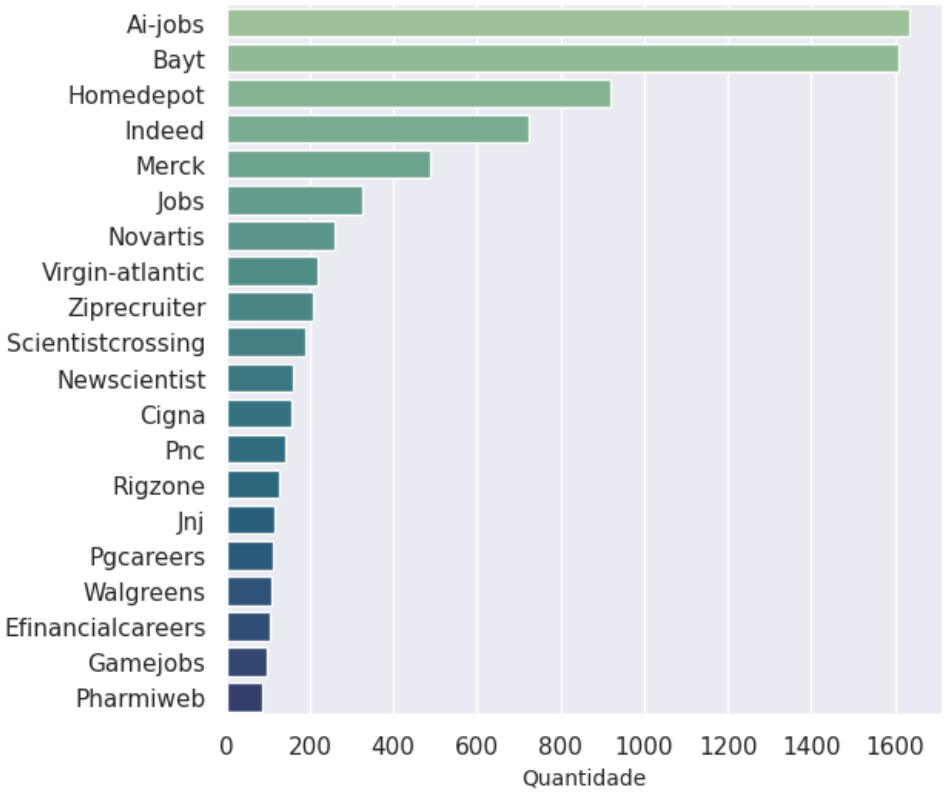


Figura 14 – Plataformas com a maior quantidade de anúncios de vagas de ciência de dados.

Fonte: Própria autora.

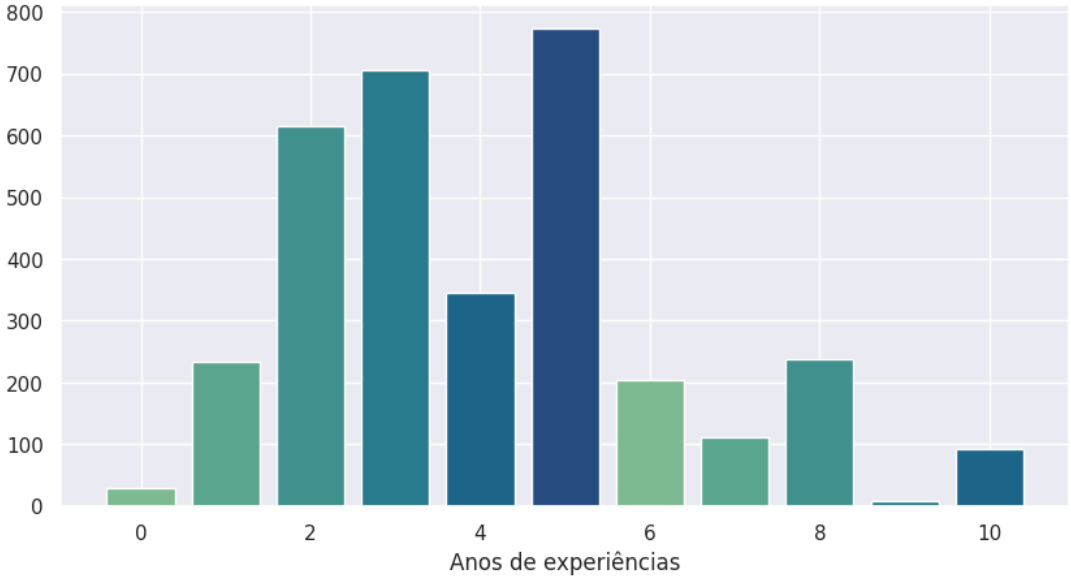


Figura 15 – Tempo de experiência requisitado nas vagas.

Fonte: Própria autora.

países exibem uma presença menor, sugerindo variações nas oportunidades de emprego em diferentes regiões. Esses dados ajudam a entender como as pessoas estão recrutando

Tabela 5 – Representação dos *N-gramas* construídos baseado nos títulos das vagas.

N-gramas	valor correspondente ao TF-IDF
Data scientist	4711.158049
Senior data scientist	1277.040439
Lead data scientist	233.725158
data scientist jobs	213.414673
Principal data scientist	206.062451
Associate data scientist	150.797253
Machine learning	93.339243
Junior data scientist	91.107934
Staff data scientist	90.405591
Data scientist intern	83.806842
Data scientist analyst	63.850644
Clinical data scientist	41.114523
Data analytics	34.253779
Data scientist ai	31.683058
Director data scientist	31.101681
Manager data scientist	30.692433
Data scientist engineer	29.273071
Scientist nlp	29.173354

Fonte: Própria autora.

no mundo todo e podem ajudar a encontrar emprego e crescer profissionalmente em um contexto internacional.

Levando em consideração a predominância dos Estados Unidos em relação aos demais países, é apresentado o mapa do mesmo com a distribuição de vagas de emprego pelo território (Figura 17), revela uma concentração significativa em dois estados da costa oeste: Califórnia e Washington. Convém pontuar que o estado de Washington é diferente da cidade Washington, DC, que é o distrito federal americano e é localizado na costa leste. Do estado de Washington, o mais ao norte da costa oeste no mapa da Figura 17, a capital é Seattle. Já a capital da Califórnia é Sacramento. Por fim, os outros estados apresentaram uma proporção significativa de vagas.

Em relação à modelagem de tópicos, inicialmente o BerTopic gerou 320 tópicos. Visando melhorar a legibilidade e seguindo a estratégia *human-in-the-loop*, foram testados diversos valores para o número de tópicos (250, 200, 150 100, 50, 25). Após a leitura dos tópicos e considerando o conhecimento das descrições das vagas, foi selecionado o valor de 50 tópicos para serem analisados. Durante a análise para o entendimento do sentido dos tópicos e a curadoria para rotulá-los, 21 tópicos foram omitidos, pois continham palavras irrelevantes para as análises como posições preenchidas, vagas fechadas, problemas na coleta (edição de vaga em aberto), etc. Esses tópicos abrangem uma variedade de assuntos relacionados a diferentes meios que o profissional de ciência de dados pode estar inserido

Figura 16 – Distribuição de anúncios de vagas por país e frequência.

Fonte: Própria autora.

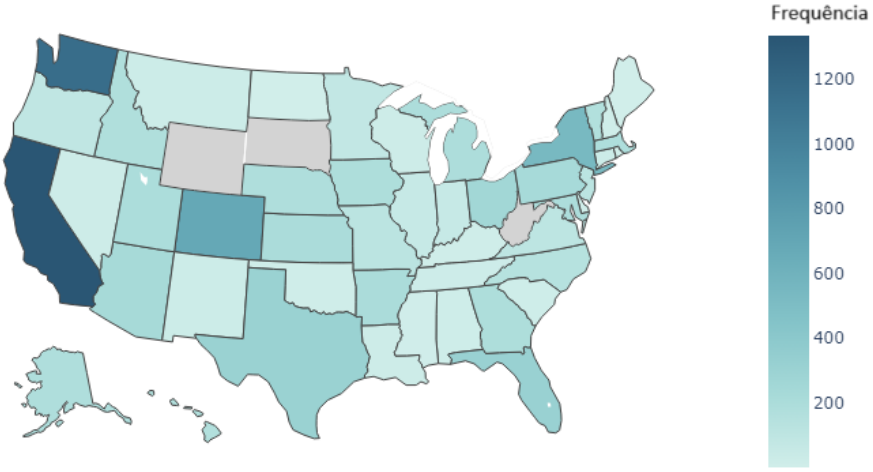


Figura 17 – Distribuição de anúncios de vagas nos Estados Unidos e frequência.

Fonte: Própria autora.

no mercado.

Tabela 6 – Representação dos tópicos relacionados análise de dados e negócios.

Tópicos	**Palavras-chave**
Análise e Dados de Negócios	*data, business, experience, work, science, learning, team, company, analytics, machine*
Insights de negócios	*data, business, work, team, experience, johnson, adobe, science, insights, world*
Dados e aprendizado de máquina	*civis, media, learning, science, applied data, machine, machine learning, new, work, data*
Análise de petróleo e gás	*ge, development, analytics, oil, industrial, oil gas, gas, ge oil, work, characteristics*
Dados comércio	*etsy, shopify, commerce, data, experience, product, role, team, merchants, shopifys*
Gerente Sênior de Varejo	*walmart, lowe, science, data science, data, lowes, senior, retail, manager, customer*

Fonte: Própria autora.

Tabela 7 – Representação dos tópicos relacionados a área da saúde e pesquisa.

Tópicos	**Palavras-chave**
Estágio de Pesquisa em Saúde	*data, internship, work, must, health, research, program, intern, students, school*
Dispositivos Médicos	*data, medical, help people, device, people, chronic pain, pain movement, movement disorders, disorders, chronic*
Análise de Vida	*abbott, data, experience, analytics, analytical, business, life, stakeholders, analysis, compliant handling*
Ciência comportamental	*behavioral, sandoz, behavioral science, behavioral data, science, ethics, ethics risk, compliance, risk, data*

Fonte: Própria autora.

No contexto de análise de dados e negócios alguns tópicos foram inseridos na Tabela 6, uma vez que diferentes negócios utilizam abordagens analíticas, e assim é possível entender o quanto é importante um perfil mais interdisciplinar nesta área, estes são: tópico 0: Análise e dados de negócios, tópico 5: *Insights* de negócios, tópico 11: Dados e aprendizado de máquina, tópico 30: Análise de petróleo e gás, tópico 13: Dados comércio, tópico 22: Gerente sênior de varejo. Sendo assim, é possível observar o impacto significativo da ciência de dados na sociedade, na ciência e nos negócios o qual foi mencionado por (GROSSI et al., 2021).

Dentre os anúncios é possível observar por meio dessas informações que existem oportunidades relacionadas a ciência de dados aplicada à área da saúde e pesquisa em diferentes meios, os tópicos que formam esse contexto estão na Tabela 7 e são: tópico 3: Estágio de pesquisa em saúde, tópico 7: Dispositivos médicos, tópico 38: Análise de vida e tópico 33: Ciência comportamental. Esses resultados corroboram o que o já foi mencionado por Demchenko et al. (2016) que o campo de ciência de dados requer uma abordagem multidisciplinar onde os campos das oportunidades são variados.

Alguns tópicos seguiram para o ramo das finanças e modelagem (Tabela 8), estes

Tabela 8 – Representação dos tópicos relacionados a finanças e modelagem.

Tópicos	**Palavras-chave**
Modelagem AML	*pnc, aml, modeling, business, data, usaa, management, employees, risk*
Salários de cientistas de dados	*average, salary, scientist, data scientist, salaries, national average, average annual, pay, data, scientist salaries*
Dados e negócios	*travelers, expedia, business, actuarial, grand, grand junction, expedia group, junction, statistical, data*
Análise de Risco	*binance, binance us, risk, readable files, machine readable, health plans, readable, files, us, cryptocurrency*

Fonte: Própria autora.

Tabela 9 – Representação dos tópicos ligados ao entretenimento.

Tópicos	**Palavras-chave**
Experiências de Entretenimento	*disney, entertainment, experiences, walt, walt disney, disney company, data, stories, espn, streaming*
Insights de dados musicais	*music, spotify, product, artists, spotifys, insights, data, audio, user, work*
Gerenciamento de preferências do usuário	*manage preferences, provide content, navigation improve, content third, navigation, third parties, manage, improve products, manage settings, select manage*

Fonte: Própria autora.

são: tópico 6: Modelagem *Aggregate Modeling Language* (AML), tópico 8: Salários de cientistas de dados, tópico 26: Dados, negócios e tópico 48: Análise de risco. Também foi possível notar tópicos mais voltados para o entretenimento(Tabela 9), isso já revela a demanda por profissionais que trabalham neste mercado relacionado aos serviços de *streaming*, que são: tópico 24: Experiências de entretenimento, tópico 29: Análise de dados musicais e o tópico 23: Gerenciamento de preferências do usuário, focado mais no gerenciamento e questão de conteúdo;

No âmbito tecnológico, tais como governança, bibliotecas específicas; e outros, os seguintes tópicos estão relacionados à tecnologia e inovação (Tabela 10): tópico 16: Design e dados, tópico 18: Técnicas modernas, tópico 43: Engenharia de Pesquisa, Tópico 28: Modelo de Governança, tópico 46: Métricas de Projeto Experimental, tópico 41: Proficiência em Bibliotecas, Tópico 12: Software de Análise Estatística e tópico 20: Processamento de linguagem natural.

É importante mencionar que alguns tópicos não foram enquadrados nestes contextos destacados, por não se assemelharem aos demais, como o tópico 27: Pesquisa em Geociências o qual é mais direcionado para análise das ciências da terra; o Tópico 10: Ciência de Dados e trabalho e por fim o tópico 37: Técnicas e habilidades, que são mais gerais relacionados a empregabilidade.

Levando em conta as informações apresentadas por ZIPRECRUITER (2023), Fuller

Tabela 10 – Representação dos tópicos relacionados a tecnologia e inovação.

Tópicos	**Palavras-chave**
Design e dados	*wolt, yammer, product, team, microsoft, get, wolts, people, work, designers data*
Técnicas Modernas	*cerner, applying modern, modern, discover, mathematical, techniques, intelligent experiences, cerner intelligence, cerner focusing, scientifically valid*
Engenharia de Pesquisa	*siri, apple, search, siri search, data, engineering teams, evaluation, search engineering, measurement evaluation, analysis*
Modelo de Governança	*governance, model governance, model, cigna, health, tobacco, assumptions performance, governance needs, governance framework, fields mathematics*
Métricas de Projeto Experimental	*experimental design, protected, metricsexperimental, qualificationsqualificationsmasters, metric like, approaches learn, demand dimensions, improvements qualificationsqualificationsmasters, engineers areas*
Proficiência em Bibliotecas	*entire, libraries excellence, one mandatory, awareness entire, numpy libraries, mandatory proficiency, apis services, good awareness, java good, using either*
Software de Análise Estatística	*data, learning, software, etc, techniques, analysis, systems, statistical, products, machine*
Processamento de Linguagem Natural	*nlp, language, natural, natural language, experience, processing, text, nlp data, data, team*

Fonte: Própria autora.

et al. (2022), percebe-se que a contratação por habilidades está sendo adotada com grande intensidade. Dessa forma, ao elaborar o léxico, foi possível realizar uma análise meticulosa das 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. Sendo assim, na Figura 18 é apresentada a distribuição total dos eixos do léxico, oferecendo uma visão global desta análise. Em seguida, é feito o detalhamento de cada parte individualmente.

Os resultados da análise das tecnologias mais frequentes são apresentados na Figura 19. 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 *MongoDB*, sistema operacional *Linux*, visualização de dados como *Tableau*, controle de versão e bibliotecas para manipulação de dados como *Pandas* e *Numpy*. Essas constatações indicam a relevância e demanda por domínio relacionadas a essas tecnologias no mercado de trabalho. Na Figura 20 são apresentadas as linguagens de programação mais recomendadas, revelando o predomínio do *Python*, seguido pela linguagem *R*. Alguns destes conhecimentos estão entre as habilidades mais difíceis procuradas globalmente como *Python*, *SQL*, *Java*. Essas constatações indicam a relevância e demanda por domínios relacionadas a essas

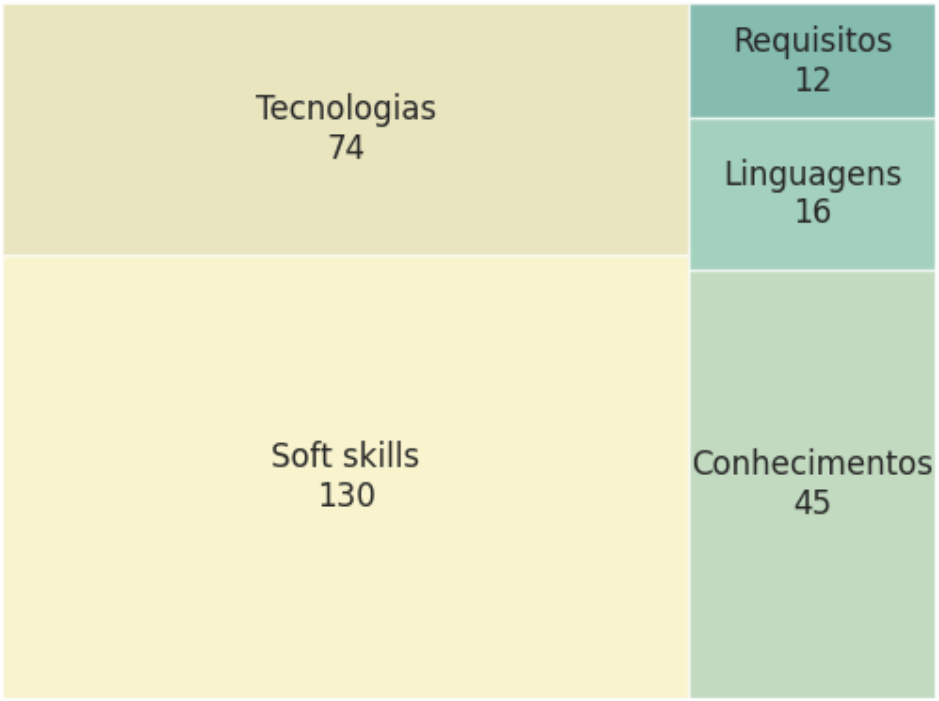


Figura 18 – Distribuição total dos elementos do léxico identificados

Fonte: Própria autora.

tecnologias no mercado de trabalho, como em Akkol, Olucoglu e Dogan (2023) que seus achados para profissionais listam tecnologias como *Linux*, *Python*, *SQL*, *Java* e *cloud computing*; assim como em Umamaheswaran et al. (2023) é possível inferir sua proximidade com ciência de dados, pois seus resultados de conceitos, ferramentas e habilidades para *Business Analytics* tem proximidade.

Ao abordar o campo multidisciplinar da Ciência de Dados, é evidente que existem diversos conhecimentos associados à área. A Figura 21 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. Esses conhecimentos estão conectados às tecnologias mais frequentes para um bom desempenho em suas aplicações.

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). Além dos conhecimentos destacados para a área em questão, como já foi mencionado por Economics (2017), as habilidades trazem um diferencial para o mercado de trabalho. A Figura 22 apresenta as habilidades mais mencionadas, juntamente com sua frequência de ocorrência, segundos os resultados, um Cientista de dados precisa ser *responsible*, além disso, outras habilidades que estão conec-

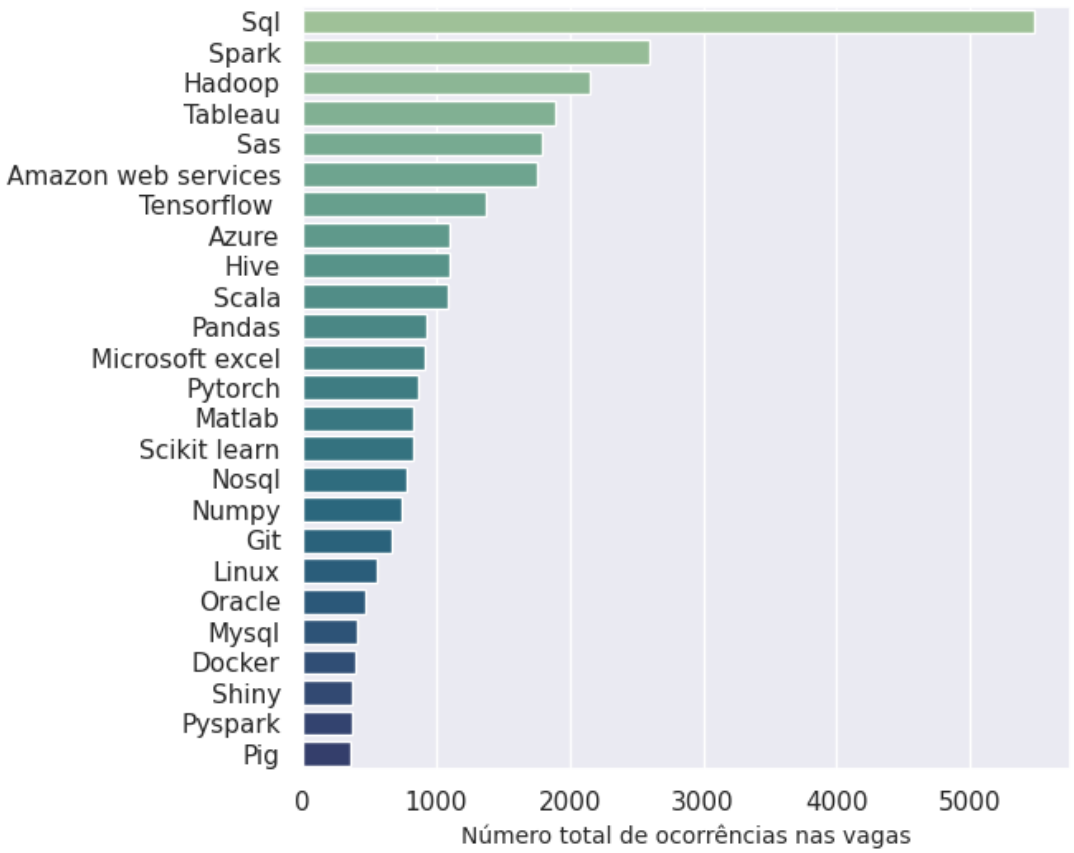


Figura 19 – Tecnologias frequentemente encontradas nas descrições das vagas.

Fonte: Própria autora.

tadas com um ambiente em equipe são *communication skill*, *team player* e *empathy* isso já demonstra que além de conhecimentos técnicos é necessário trabalhar as habilidades comportamentais.

A Figura 23 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, Yinxia e Min (2021), que analisou vagas relacionadas a *big data*, percebe-se que a exigência de formação é menor nessa área em comparação à ciência de dados.

Neste sentido, hipotetizamos que as vagas de cientistas de dados juniors 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.

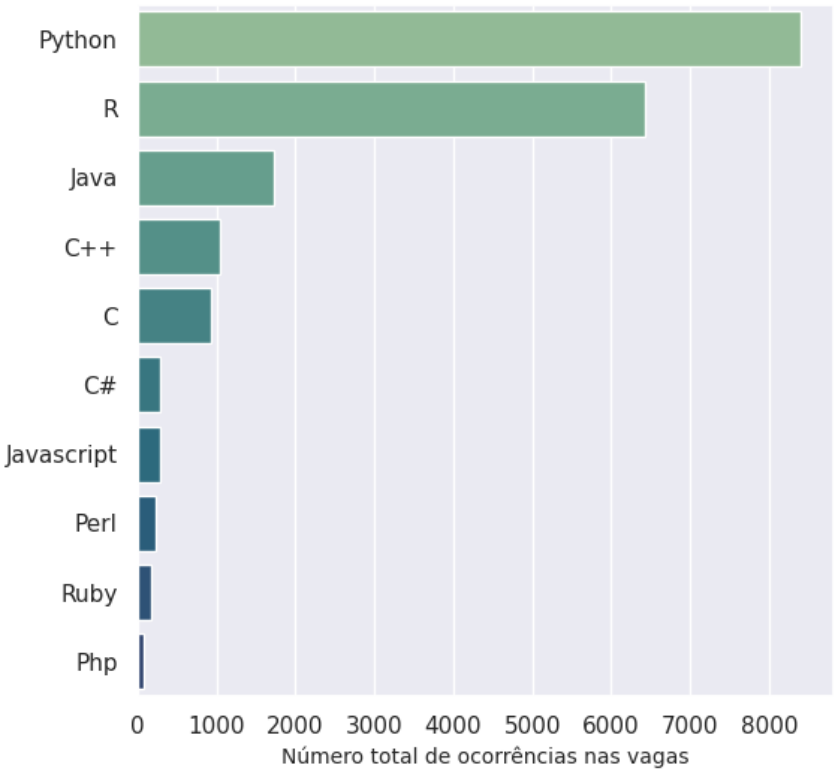


Figura 20 – Linguagens de programação mais mencionadas nos anúncios.

Fonte: Própria autora.

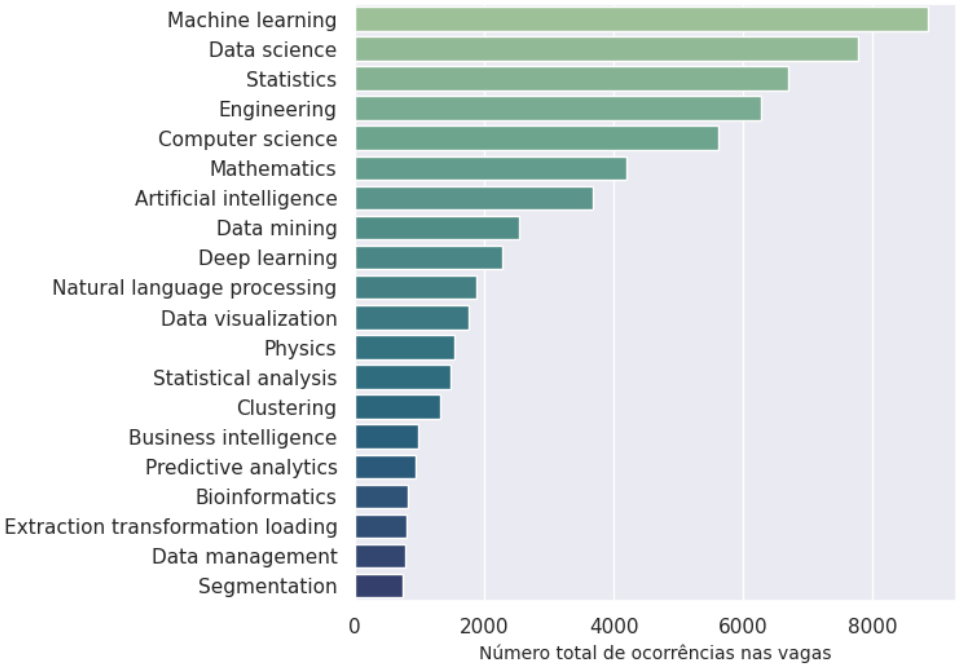


Figura 21 – Conhecimentos mais recorrentes nas descrições das vagas.

Fonte: Própria autora.

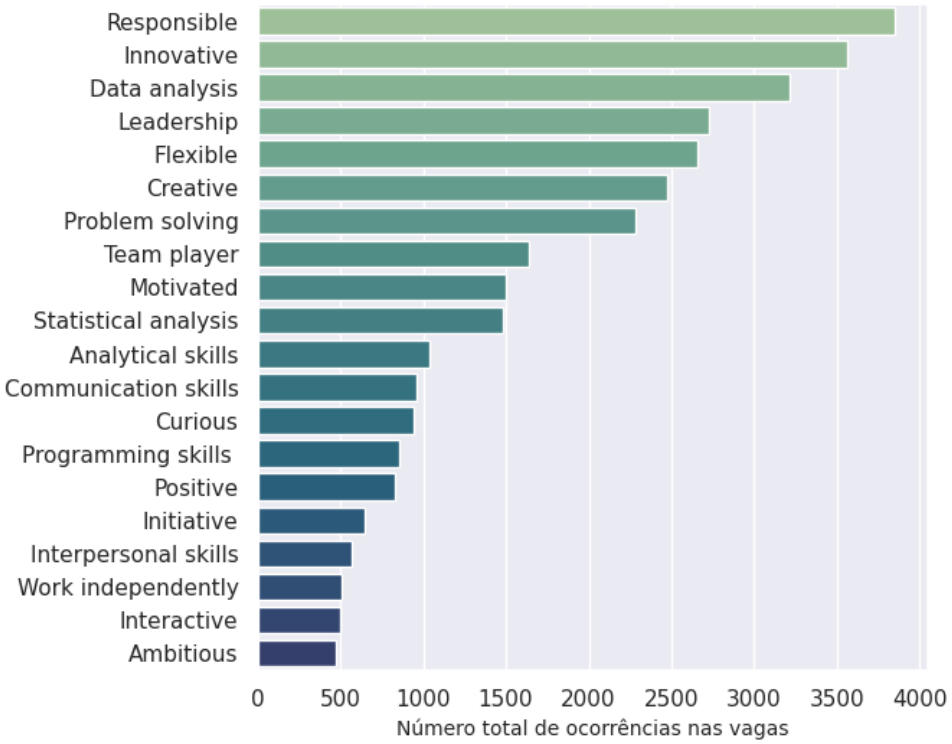


Figura 22 – Habilidades interpessoais mais requisitadas nas vagas.

Fonte: Própria autora.

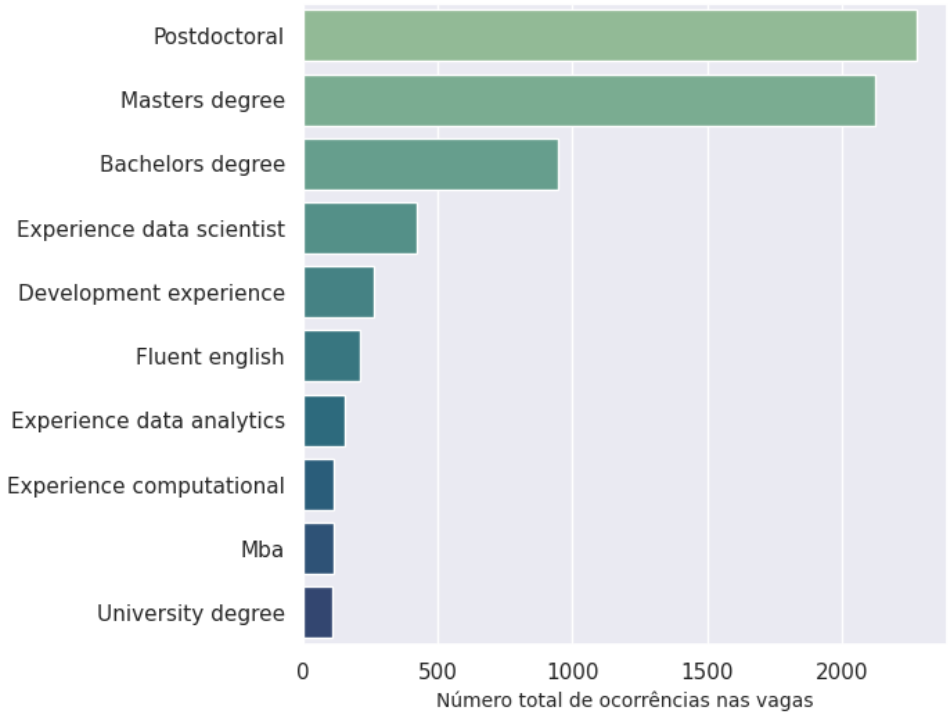


Figura 23 – Indicação dos requisitos predominantes nas descrições de vagas.

Fonte: Própria autora.

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 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, a presente pesquisa 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 e a validação dos dados seguindo o princípio *human-in-the-loop*. Para tal, foi realizado uma coleta de 14.000 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; um diferencial do estudo que merece destaque foi que na validação dos resultados foi utilizada a abordagem *grounded-theory*, seguindo uma estratégia 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 serão publicamente disponibilizados em um repositório aberto.

Os resultados alcançados permitiram identificar que existe uma preocupação relacionada inclusão e equidade em uma parte dos anúncios de vagas, assim como há oportunidades em diversos locais do mundo para trabalhar com ciência de dados. Além disso, foi possível identificar os diversos campos para os quais as vagas são direcionadas, o que reforça a ideia de que a ciência de dados é multidisciplinar. Em termos de tendências da área, foi possível ter compreensão dos cargos mais prevalentes, habilidades, conhecimentos, 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.

5.1 CONTRIBUIÇÕES TÉCNICAS

Das principais contribuições técnicas do trabalho, destacam-se:

- O dicionário construído que pode ser utilizados não somente para vagas de emprego em ciência de dados, mas para outros domínios como educação, tecnologia da informação, saúde, marketing, finanças, recursos humanos, etc.
- Um *framework* experimental para analisar vagas de emprego que utiliza diferentes técnicas para extrair valor dos dados, com foco na identificação de requisitos, habilidades comportamentais, conhecimentos, linguagens de programação e outras informações.
- Um relatório técnico com a análise do panorama atual do mercado baseado em um corte transversal de anúncios de emprego de ciência de dados.
- A partir da consolidação do conhecimento e a expertise adquirida foi possível a publicação de artigos relacionados a esta pesquisa no Simpósio Brasileiro de Sistemas de Informação (SBSI), *International Conference on Web Intelligence and Intelligent Agent Technology* (WI-IAT) e Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), os artigos encontram-se respectivamente nos apêndices A, B e C.

5.1.1 AMEAÇAS A VALIDADE DO ESTUDO

No contexto da avaliação, por se tratar de meios de avaliação qualitativa, houve influência de pontos de vista específicos dos investigadores, o que leva a que para futuras replicações do estudo novos métodos de validação complementares sejam necessários. Nesta pesquisa, ao analisar apenas anúncios em inglês, pode ser notado um viés linguístico, pois acaba sendo excluídas outras regiões com idiomas diferentes, não trazendo uma visão tão abrangente. No mais, questões relacionadas ao tamanho da amostra, pois não se pode afirmar ser um recorte do mercado de trabalho na totalidade.

5.1.2 TRABALHOS FUTUROS

Como trabalhos futuros, pretende-se agregar diferentes técnicas de avaliação dos métodos utilizados, além de investigar o uso de *Large Language Model* na identificação de habilidades. Em relação ao léxico do domínio, pretende-se trabalhar com diferentes idiomas na construção do mesmo, além de refinar o mesmo com o intuito de construir uma base de competências que englobasse uma descrição completa da profissão, pois essas informações podem ser valiosas na detecção de competências para empregos emergentes.

Para identificar os anos e localidades pretende-se utilizar o *Part of Speech* (POS). Por fim, a ampliação da base de dados para diferentes vertentes relacionados a dados como engenharia de dados, análise de dados, entre outras.

Com relação a trabalhos futuros, vislumbra-se analisar as vagas específicas de sites de emprego internacionais e nacionais, delimitando o campo de busca no Brasil para o entendimento do mercado de trabalho nacional na área de dados. Além disso, é necessário relacionar as variáveis do *framework* experimental com informações de cursos acadêmicos para uma maior compreensão e também a inserção de outras frentes de análises, como a questão da inclusão, modalidade de trabalho, faixa salarial, benefícios e outras.

REFERÊNCIAS

AGARWAL, R.; DHAR, V. *Big data, data science, and analytics: The opportunity and challenge for IS research*. [S.l.]: INFORMS, 2014. 443–448 p. Citado na página 19.

AKKOL, E.; OLUCOGLU, M.; DOGAN, O. Topic modeling for skill extraction from job postings. In: SPRINGER. *Iberoamerican Knowledge Graphs and Semantic Web Conference*. [S.l.], 2023. p. 277–289. Citado 3 vezes nas páginas 31, 32 e 56.

ALIBASIC, A. et al. Evaluation of the trends in jobs and skill-sets using data analytics: a case study. *Journal of Big Data*, Springer, v. 9, n. 1, p. 32, 2022. Citado 2 vezes nas páginas 31 e 32.

ALJOHANI, N. R. et al. Bridging the skill gap between the acquired university curriculum and the requirements of the job market: A data-driven analysis of scientific literature. *Journal of Innovation & Knowledge*, Elsevier, v. 7, n. 3, p. 100190, 2022. Citado na página 21.

ANTONS, D. et al. The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. *R&D Management*, Wiley Online Library, v. 50, n. 3, p. 329–351, 2020. Citado na página 22.

AO, Z. et al. Skill requirements in job advertisements: A comparison of skill-categorization methods based on explanatory power in wage regressions. *arXiv preprint arXiv:2207.12834*, 2022. Citado 2 vezes nas páginas 30 e 32.

ARMAN, M. The advantages of online recruitment and selection: A systematic review of cost and time efficiency. *Business Management and Strategy*, v. 14, n. 2, p. 220–240, 2023. Citado na página 21.

BAIRD, A.; XIA, Y.; CHENG, Y. Consumer perceptions of telehealth for mental health or substance abuse: a twitter-based topic modeling analysis. *JAMIA open*, Oxford University Press, 2022. Citado na página 43.

BATTISTA, A. D.; GRAYLING, S.; HASSELAAR, E. *Future of jobs report 2023*. [S.l.], 2023. Citado 2 vezes nas páginas 23 e 36.

BOSELLI, R. et al. Wolmis: A labor market intelligence system for classifying web job vacancies. *Journal of intelligent information systems*, Springer, v. 51, p. 477–502, 2018. Citado na página 23.

BRADLEY, A.; QUIGLEY, M.; BAILEY, K. How well are students engaging with the careers services at university? *Studies in Higher Education*, Taylor & Francis, v. 46, n. 4, p. 663–676, 2021. Citado na página 24.

BRADY, H. E. The challenge of big data and data science. *Annual Review of Political Science*, Annual Reviews, v. 22, p. 297–323, 2019. Citado na página 19.

BRÜNING, N.; MANGEOL, P. What skills do employers seek in graduates?: Using online job posting data to support policy and practice in higher education. OECD, 2020. Citado na página 21.

CALANCA, F. et al. Responsible team players wanted: an analysis of soft skill requirements in job advertisements. *EPJ Data Science*, SpringerOpen, v. 8, n. 1, p. 1–20, 2019. Citado na página 44.

CAO, L. Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 50, n. 3, p. 1–42, 2017. Citado na página 19.

CARVALHO, F. S. de; SORCI, P. A. B. da S.; FIGUEIREDO, G. L. A. de S. Os desafios do administrador frente às novas tendências. *Facit Business and Technology Journal*, v. 1, n. 21, 2020. Citado na página 20.

CHAIBATE, H. et al. Analyzing the engineering soft skills required by moroccan job market. In: IEEE. *2019 5th international conference on optimization and applications (ICOA)*. [S.l.], 2019. p. 1–6. Citado na página 45.

CHARMAZ, K.; THORNBERG, R. The pursuit of quality in grounded theory. *Qualitative research in psychology*, Taylor & Francis, v. 18, n. 3, p. 305–327, 2021. Citado na página 45.

CIRQUEIRA, D. et al. A literature review in preprocessing for sentiment analysis for brazilian portuguese social media. In: *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. [S.l.: s.n.], 2018. Citado na página 41.

DEBAO, D.; YINXIA, M.; MIN, Z. Analysis of big data job requirements based on k-means text clustering in china. *PloS one*, Public Library of Science San Francisco, CA USA, v. 16, n. 8, p. e0255419, 2021. Citado 4 vezes nas páginas 30, 32, 41 e 57.

DEDE, N. P. The role of e-hrm practices on digital era. In: *Tools and techniques for implementing international e-trading tactics for competitive advantage*. [S.l.]: IGI Global, 2020. p. 1–20. Citado na página 24.

DELOITTE. *2023 Gen Z and Millennial Survey*. 2023. Disponível em:https://www.deloitte.com/global/en/issues/work/content/genzmillennialsurvey.html. Acesso em: 27 de Novembro 2023. Citado 2 vezes nas páginas 25 e 47.

DEMCHENKO, Y. et al. Edison data science framework: a foundation for building data science profession for research and industry. In: IEEE. *2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*. [S.l.], 2016. p. 620–626. Citado 2 vezes nas páginas 20 e 53.

Domo. *Data Never Sleeps 11.0*. 2023. Https://www.domo.com/learn/infographic/data-never-sleeps-11?utm$_s ource$ = $oktopostutm_m edium$ = $prutm_c ampaign$ = $OP_D ata_N ever_S leeps_1 1_1 21423campid$ = $7015w000000vniPAAQ.Citado2vezesnaspáginas20e$ 107.

DUBEY, R. S.; PAUL, J.; TEWARI, V. The soft skills gap: a bottleneck in the talent supply in emerging economies. *The International Journal of Human Resource Management*, Taylor & Francis, v. 33, n. 13, p. 2630–2661, 2022. Citado na página 24.

ECONOMICS, D. A. Soft skills for business success. *DeakinCo*, 2017. Citado 4 vezes nas páginas 23, 25, 56 e 109.

FORUM, V. W. E. The future of jobs report 2020. *Retrieved from Geneva*, 2020. Citado 3 vezes nas páginas 22, 23 e 36.

FULLER, J. B. et al. The emerging degree reset. AEI: American Enterprise Institute for Public Policy Research, 2022. Citado 3 vezes nas páginas 23, 54 e 55.

GHAZZAWI, K.; ACCOUMEH, A. Critical success factors of the e-recruitment system. *Journal of Human Resources Management and Labor Studies*, v. 2, n. 2, p. 159–170, 2014. Citado na página 24.

GROOTENDORST, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022. Citado na página 43.

GROSSI, V. et al. Data science: a game changer for science and innovation. *International Journal of Data Science and Analytics*, Springer, v. 11, p. 263–278, 2021. Citado 2 vezes nas páginas 20 e 53.

GUIMARÃES, A. J. R.; JÚNIOR, R. M.; FREITAS, M. d. C. D. Requisitos para a ciência de dados: analisando anúncios de vagas de emprego com mineração de texto. *Revista Ibérica de Sistemas e Tecnologias de Informação*, Associação Ibérica de Sistemas e Tecnologias de Informacao, n. 46, p. 54–70, 2022. Citado 3 vezes nas páginas 29, 32 e 35.

GURCAN, F. Extraction of core competencies for big data: Implications for competency-based engineering education. *International Journal of Engineering Education*, v. 35, n. 4, p. 1110–1115, 2019. Citado 3 vezes nas páginas 30, 32 e 56.

GURCAN, F.; CAGILTAY, N. E. Big data software engineering: Analysis of knowledge domains and skill sets using lda-based topic modeling. *IEEE access*, IEEE, v. 7, p. 82541–82552, 2019. Citado 4 vezes nas páginas 22, 30, 32 e 41.

HASSAN, H. et al. Career path decisions and sustainable options. *Sustainability*, MDPI, v. 14, n. 17, p. 10501, 2022. Citado na página 25.

HASSANI, H. et al. Text mining in big data analytics. *Big Data and Cognitive Computing*, MDPI, v. 4, n. 1, p. 1, 2020. Citado na página 22.

HERNÁNDEZ, E. J. M. The training of analytical talent: A shared task between university and industry. *Dimensión Empresarial*, Universidad Autonoma del Caribe, v. 19, n. 1, p. 92, 2021. Citado na página 19.

INFOJOBS. *Sites de emprego: afinal, qual o melhor?* 2023. Disponível em:https://blog.infojobs.com.br/candidatos/sites-de-emprego-qual-o-melhor/. Acesso em: 02 de Fevereiro 2024. Citado na página 48.

JACKSON, D.; TOMLINSON, M. Investigating the relationship between career planning, proactivity and employability perceptions among higher education students in uncertain labour market conditions. *Higher education*, Springer, v. 80, n. 3, p. 435–455, 2020. Citado na página 25.

JAGANNATHAN, S.; RA, S.; MACLEAN, R. Dominant recent trends impacting on jobs and labor markets-an overview. *International Journal of Training Research*, Taylor & Francis, v. 17, n. sup1, p. 1–11, 2019. Citado na página 20.

KHAOUJA, I.; KASSOU, I.; GHOGHO, M. A survey on skill identification from online job ads. *IEEE Access*, IEEE, v. 9, p. 118134–118153, 2021. Citado 2 vezes nas páginas 21 e 23.

KUREKOVA, L. M.; BEBLAVY, M.; THUM, A.-E. Using internet data to analyse the labour market: a methodological enquiry. IZA Discussion Paper, 2014. Citado na página 21.

LINKEDIN. *7 Job Post Tips*. 2023. Disponível em:https://business.linkedin.com/talent-solutions/recruiting-tips/7-job-post-tips. Acesso em: 02 de Fevereiro 2024. Citado na página 40.

LINKEDIN. *The Most In-Demand Skills for 2023*. 2023. Disponível em:https://www.linkedin.com/business/talent/blog/talent-strategy/linkedin-most-in-demand-hard-and-soft-skills. Acesso em: 29 de Novembro 2023. Citado 3 vezes nas páginas 23, 24 e 36.

LINKEDIN. *O futuro do recrutamento 2023*. 2023. Disponível em:https://business.linkedin.com/pt-br/talentsolutions/resources/future-of-recruiting. Acesso em: 27 de Abril 2023. Citado na página 24.

MCDONALD, S.; DAMARIN, A. K.; MEMBREZ-WEILER, N. J. Organizational perspectives on digital labor market intermediaries. *Sociology Compass*, Wiley Online Library, v. 17, n. 4, p. e13061, 2023. Citado na página 21.

MCKINSEY & COMPANY. *Soft skills for a hard world As workers interact with ever-smarter machines, the demand for soft skills is beginning to surge*. 2019. Disponível em:https://www.mckinsey.com/featured-insights/future-of-work/five-fifty-soft-skills-for-a-hard-world. Acesso em: 13 de Novembro 2023. Citado na página 23.

MCKINSEY & COMPANY. *Beyond hiring: How companies are reskilling to address talent gaps*. 2020. Disponível em:https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/beyond-hiring-how-companies-are-reskilling-to-address-talent-gaps. Acesso em: 13 de Novembro 2023. Citado na página 23.

MCKINSEY & COMPANY. *The Gen Z equation*. 2023. Disponível em:https://www.mckinsey.com/quarterly/the-five-fifty/five-fifty-the-gen-z-equation. Acesso em: 13 de Novembro 2023. Citado na página 25.

MILLER, S.; HUGHES, D. The quant crunch: How the demand for data science skills is disrupting the job market. *Burning Glass Technologies*, 2017. Citado na página 19.

MIRO. *Asynchronous Work Report: What knowledge workers want and what's working*. 2023. Disponível em:https://miro.com/blog/asynchronous-work-report/. Acesso em: 27 de Novembro 2023. Citado na página 25.

NAPIERALA, J.; KVETAN, V. Changing job skills in a changing world. In: *Handbook of Computational Social Science for Policy*. [S.l.]: Springer International Publishing Cham, 2023. p. 243–259. Citado na página 19.

PALETTA, F. C.; GONZÁLEZ, J. A. M. 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*, v. 26, n. 3, 2021. Citado na página 22.

PAPOUTSOGLOU, M. et al. Extracting knowledge from on-line sources for software engineering labor market: A mapping study. *IEEE Access*, IEEE, v. 7, p. 157595–157613, 2019. Citado 4 vezes nas páginas 21, 26, 29 e 32.

REIS, L. C. R.; FONSECA, M. I. da et al. Big data: Um novo campo de atuação para bibliotecários. *Prisma. Com*, v. 2020, n. 41, p. 231–250, 2020. Citado na página 19.

SCHRÖER, C.; KRUSE, F.; GÓMEZ, J. M. A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, Elsevier, v. 181, p. 526–534, 2021. Citado na página 35.

SCULLY-RUSS, E.; TORRACO, R. The changing nature and organization of work: An integrative review of the literature. *Human Resource Development Review*, Sage Publications Sage CA: Los Angeles, CA, v. 19, n. 1, p. 66–93, 2020. Citado na página 20.

SILVEIRA, C. C. da et al. What is a data scientist? analysis of core soft and technical competencies in job postings. *Revista Inovação, Projetos e Tecnologias*, v. 8, n. 1, p. 25–39, 2020. Citado 2 vezes nas páginas 29 e 32.

SINOARA, R. A.; MARCACINI, R. M.; REZENDE, S. O. Mineraçao de textos e semântica: desafios, abordagens e aplicaçoes. *Revista de Sistemas de Informaçao da FSMA*, v. 27, n. ja-ju 2021, p. 41–53, 2021. Citado na página 22.

SKYQUEST. *The Most In-Demand Skills for 2023*. 2023. Disponível em:https://www.skyquestt.com/report/online-recruitment-market. Acesso em: 29 de Novembro 2023. Citado na página 24.

TERNIKOV, A. Soft and hard skills identification: insights from it job advertisements in the cis region. *PeerJ Computer Science*, PeerJ, Inc, v. 8, 2022. Citado 4 vezes nas páginas 22, 30, 32 e 56.

UMAMAHESWARAN, S. et al. What do employers look for in "business analytics" roles?–a skill mining analysis. *Information Systems Frontiers*, Springer, p. 1–17, 2023. Citado 3 vezes nas páginas 31, 32 e 56.

WIRTH, R.; HIPP, J. Crisp-dm: Towards a standard process model for data mining. In: MANCHESTER. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. [S.l.], 2000. v. 1, p. 29–39. Citado na página 35.

WU, X. et al. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, Elsevier, v. 135, p. 364–381, 2022. Citado 3 vezes nas páginas 29, 32 e 33.

ZIPRECRUITER. *Beyond Degrees: Rethinking Qualifications in the Skills-First Era*. 2023. Disponível em:https://www.ziprecruiter.com/blog/skills-hiring-report/. Acesso em: 29 de Novembro 2023. Citado 4 vezes nas páginas 23, 36, 54 e 55.

APÊNDICES

APÊNDICE A – ARTIGO PUBLICADO NO *SBSI* 2024

CRM Market Overview: A Case Study of Job Vacancies

Fernando A. do Carmo
Universidade Federal do Oeste do Pará
Santarém, Pará, Brasil

Pedro H. C. Menezes
Universidade Federal do Oeste do Pará
Santarém, Pará, Brasil

Bárbara A. P. Barata
Universidade Estadual do Maranhão
São Luis, Maranhão, Brasil

Antonio F. L. Jacob Junior
Universidade Estadual do Maranhão
São Luis, Maranhão, Brasil
antoniojunior@professor.uema.br

Fabio M. F. Lobato
Universidade Federal do Oeste do Pará
Santarém, Pará, Brasil
fabio.lobato@ufopa.edu.br

RESUMO

Context: Social media has revolutionized how companies relate to their customers, directly impacting Customer Relationship Management (CRM) systems. Furthermore, such media have drastically changed how companies advertise job vacancies and recruit employees. **Problem:** Considering the recruitment platforms multitude and the increased job post data volume, extracting knowledge to keep up with market developments represents a prominent research challenge. **Solution:** We present a textual analysis of job advertisements related to CRM using text mining methods and a lexical construction. We identified the most common vacancies, as well as knowledge, technical skills (hard skills), and interpersonal skills (soft skills). **IS Theory:** This work was conceived from the perspective of Design Theory, mainly concerning the construction of artifacts that allow textual analysis. **Method:** The analyses were carried out using a methodology based on Design Science Research, where the problem was identified, the objectives were outlined, and moving on to the design stage, where we applied text mining techniques including N-gram analysis of job descriptions, topic modeling and the lexicon construction oriented to the CRM area. **Summary of Results:** The research findings reveal the main knowledge areas, experience requirements, skills, training, and technologies that stand out in this sector, as well as characteristics such as preference for diverse teams, concerned with gender equity, inclusion of LGBQIAP+ community and race/color diversity. The constructed lexicon can be used for a more precise and structured view of the CRM employment scenario. **Contributions and Impact in the IS area:** This information provides good opportunities for candidates and recruiters in the hiring process, allowing both to identify such aspects more precisely and supporting decision-making.

CCS CONCEPTS

• **Information systems** → **Data mining**; **Web mining**; • **Computing methodologies** → **Natural language processing**.

Publication rights licensed to ACM. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of a national government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.
SBSI '24, May 20–23, 2024, Juiz de Fora, Brazil
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0996-8/24/05…$15.00
https://doi.org/10.1145/3658321.3658362

KEYWORDS

CRM, Text Mining, Data Analysis, Job Posts, Market Analysis

ACM Reference Format:
Fernando A. do Carmo, Pedro H. C. Menezes, Bárbara A. P. Barata, Antonio F. L. Jacob Junior, and Fabio M. F. Lobato. 2024. CRM Market Overview: A Case Study of Job Vacancies. In *XX Brazilian Symposium on Information Systems (SBSI '24), May 20–23, 2024, Juiz de Fora, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3658321.3658362

1 INTRODUÇÃO

As mídias sociais revolucionaram a forma com a qual as empresas se relacionam com os seus clientes, impactando diretamente nos sistemas de gestão de relacionamento com clientes, mais conhecido pelo termo em inglês, *Customer Relationship Management* (CRM). Ademais, tais mídias também mudaram drasticamente a forma com a qual as empresas divulgam vagas de emprego e recrutam colaboradores [30]. Neste seguimento, é de grande importância para este setor explorar novas estratégias em CRM, considerando os aspectos que o atual mercado exige [34]. Sendo assim, o CRM surgiu como uma nova ferramenta para gerenciar e otimizar a automação da força de vendas nas empresas, se tornando uma das principais estratégias para gerenciamento de informações empresariais, não apenas para fins de vendas e *marketing*, mas também para uma interação mais eficaz com o cliente [23].

No contexto atual, as redes sociais proporcionam um meio de fortalecer o relacionamento entre clientes e prestadores de serviços. A adoção das mídias sociais no CRM é conhecida como *Social CRM* ou uma segunda geração de CRM (CRM 2.0) que permite aos clientes expressarem suas opiniões e expectativas sobre produtos ou serviços [3]. Outro fenômeno importante advindo da pervasividade das mídias sociais é o processo de recrutamento *online*, na qual empresas utilizam destes meios para analisar o perfil de candidatos, buscando por perfis qualificados e que melhor se adéquem aos requisitos que a vaga exige [4, 7]. As ferramentas atuais permitem que os recrutadores publiquem automaticamente oportunidades, acessem os perfis dos interessados para explorar mais informações, rastreiem quais canais sociais geram mais *leads* e que resultam nos melhores candidatos [26]. Neste cenário, áreas como *Big Data* têm sido bastante exploradas por empresas, haja vista que o grande volume de dados gerados por usuários nas mídias sociais pode ajudar as empresas a retratar seu comportamento para ganhar

valor, especialmente em vendas, atendimento ao cliente, *marketing* e promoção [18, 36].

Considerando a multitude de plataformas de recrutamento e o aumento do volume de dados gerados, a extração de conhecimento para acompanhar a evolução mercadológica representa um desafio de pesquisa importante, uma vez que as análises manuais desses dados tornam-se um processo custoso e ineficaz para as empresas [15, 25]. Dessa forma, técnicas baseadas em Inteligência Artificial (IA) podem ser extremamente úteis para analisar dados internos disponíveis em Sistemas de Informação (SI) da indústria CRM [35]. Uma revisão da literatura, do tipo *ad hoc* [31], conduzida no presente estudo, evidenciou que ainda existe uma lacuna de pesquisa quanto a trabalhos com foco em análises do setor CRM. Embora a revisão tenha mostrado trabalhos que utilizam técnicas de IA e Aprendizado de Máquina (AM) na busca por melhores perfis profissionais, muitos utilizam técnicas consideradas ultrapassadas e não levam em consideração o contexto de SI, que é o foco deste estudo. Além disso, o presente trabalho tem como diferencial, além do uso de técnicas mais atuais de AM consolidadas tanto no estado da arte quanto da prática, o uso de uma base de dados mais amplas em detrimento a estudos anteriores. Essa amplitude da base de dados diz respeito tanto ao volume (quantidade de instâncias), quanto à consideração de múltiplas fontes.

Sendo assim, o objetivo deste trabalho é analisar dados textuais de vagas de empregos relacionados ao CRM para identificar os perfis de profissionais mais requisitados por empresas, suas habilidades, conhecimentos e experiências, por meio de técnicas de inteligência artificial e mineração de texto. Para alcançá-lo, foram estudadas e definidas as seguintes Perguntas de Pesquisa (PP):

- PP-1: Como as empresas selecionam profissionais no setor de CRM?
- PP-2: Quais aspectos as empresas consideram para selecionar os melhores perfis de profissionais?
- PP-3: Quais as habilidades e conhecimentos mais requisitados no cenário atual do mercado em CRM?

Para responder as perguntas de pesquisa, o estudo foi conduzido embasado no método de pesquisa *Design Science Research* (DSR), o qual é bastante utilizado em SI, sobretudo, para o desenvolvimento de soluções para problemas organizacionais de relevância em SI [11]. Além disso, o estudo foi realizado sob a ótica *Design Theory*, por ser uma teoria bastante utilizada ao se tratar de um artefato inovador construído sob a perspectiva do DSR [24]. Vale destacar ainda, que o presente trabalho possui contribuições relevantes relacionadas aos grandes desafios de pesquisa em sistemas de informação no Brasil [6], principalmente no que tange ao Desafio 2 - *Information Systems and the Open World Challenges*, dado as novas discussões sobre acesso à informação, dados abertos e a evolução da análise de redes sociais [16].

Os resultados alcançados contribuem para orientar empresas nos processos de seleção, bem como para interessados em ingressar neste mercado, informando quais aspectos precisam levar em consideração antes de realizar o processo de seleção, além de promover políticas públicas voltadas a este mercado, tornando-o mais inclusivo e igualitário.

O restante deste artigo está organizado como segue. Na Seção 2, descrevem-se os trabalhos relacionados a esta pesquisa. Na Seção 3 descrevem-se os métodos utilizados para o desenvolvimento das análises. Na Seção 4 apresentam-se os resultados alcançados. Por fim, a Seção 5 as considerações finais são apresentadas e discutidas.

2 TRABALHOS RELACIONADOS

Diversos estudos voltados para a análise de vagas de emprego têm explorado técnicas de mineração de texto. Charcon *et al.* [7] utilizaram o algoritmo *k-nearest neighbor* (kNN) para extrair conhecimentos, bem como identificar padrões específicos em currículos, este processo pode auxiliar os interessados na escolha por melhores perfis de candidatos, como tempo de experiência, escolaridade e idiomas. No contexto do presente trabalho, busca-se a ampliação do trabalho de Charcon *et al.* no que tange à utilização de técnicas de *Natural Language Processing* (NLP) mais avançadas, consoantes ao estado da arte atual, tais como abordagem de modelagem de tópicos e construção de um léxico do domínio.

Gurcan e Cagiltay [21] focaram em revelar o conjunto de habilidades e os conhecimentos predominantes na área de *Big Data*. O algoritmo de modelagem de tópicos *Latent Dirichlet Allocation* (LDA) foi utilizado com o objetivo de analisar as estruturas semânticas no corpo do texto. Com tais resultados, foi possível criar um mapa sistemático das competências para promover maior clareza dos conhecimentos essenciais, habilidades desejáveis e as principais ferramentas para este setor. Os resultados podem ajudar na avaliação e melhora das seleções dos profissionais da área, identificação dos requisitos para processos de recrutamento e auxiliar a promoção de programas educacionais focados nas necessidades do mercado de trabalho. Visando atualizar o *framework* experimental de Gurcan e Cagiltay, o presente trabalho incluiu técnicas mais recentes de modelagem de tópicos visando uma análise mais robusta dos anúncios de vagas de emprego.

Cunha *et al.* [14] realizaram uma pesquisa qualitativa baseada em *Survey* com participantes de diferentes países acerca das principais habilidades requeridas por grandes empresas para equipes de desenvolvedores de *software*, os resultados mostram que dentre as habilidades se destacam tanto as pessoais quanto técnicas bem como comunicação, trabalho em equipe, programação, *marketing* e desenvolvimento de *software*. Os resultados podem ser fundamentais para orientar empresas quanto ao oferecimento de treinamentos internos e também para guiar seus processos de seleção, bem como para interessados em ingressar neste mercado, informando quais as habilidades que precisam ser adquiridas antes de prestar o processo de seleção.

Fonseca e Digiampietri [12] exploraram o uso de técnicas de AM com o objetivo de identificar as principais áreas de atuação dos pesquisadores cadastrados na Plataforma Lattes[1]. Em seguida, utilizando do título das produções científicas, foram aplicadas as técnicas de representação de N-gramas de caracteres *Term Frequency-Inverse Document Frequency* (TF-IDF) e *Word2Vec*. Os achados revelaram uma acurácia de 95,91% ao empregar a técnica de N-gramas de caracteres TF-IDF para representação de texto no idioma inglês. Para dados traduzidos, notou-se 95,78% de precisão, superando outras técnicas atuais, demonstrando a eficácia da abordagem de N-gramas e sua consolidação no estado da arte.

[1] https://lattes.cnpq.br/

No estudo de Machado e Diirr [27], o objetivo foi compreender a influência e o impacto da experiência profissional na sinergia entre os membros das equipes de Sistemas de Informação nas empresas. Por meio de uma revisão da literatura, 31 estudos foram selecionados utilizando uma *string* base: *("collaboration*") AND ("professional experience*" OR "professional background" OR "professional know-how" OR "professional knowhow") AND ("benefit*" OR "challenge*" OR "limitation*" OR "difficult*" OR "problem*" OR "opportunit*" OR "influence*" OR "impact*").* As análises realizadas revelaram que a experiência profissional e habilidades heterogêneas ampliam a capacidade de produção de uma equipe, promovendo resultados de alta qualidade, mediante a diversidade de conhecimentos entre os membros. O trabalho de Machado e Diirr auxiliou no desenho do presente estudo no que tange à definição dos termos de busca, considerando também experiência profissional, entretanto, diferencia-se por ter como fonte primária de dados os anúncios de emprego.

Barata *et al.* [4] utilizaram algoritmos de modelagem de tópicos para a extração de conhecimento de dados textuais de ofertas de vagas, com foco na área de ciência de dados. Com as análises realizadas foi possível identificar os domínios de conhecimento e conjuntos de habilidades necessárias aos profissionais desta área. Além disso, os autores propuseram uma abordagem de construção do léxico das vagas de emprego, a fim de identificar a frequência das sentenças relacionadas ao domínio no conteúdo textual das vagas de emprego. O *pipeline* de análise de vagas de emprego proposto no estudo é extensível, podendo ser adaptado para diversas áreas de interesse. Desse modo, o presente estudo busca ampliar a gama de uso do *pipeline* desenvolvido voltado a atender as necessidades presentes na esfera do CRM.

A partir da análise dos trabalhos relacionados, ficou evidente o potencial da aplicação de técnicas de mineração de texto como alternativa para identificar informações valiosas em vagas de emprego *online*. Ademais, os estudos enfatizam a abordagens de Processamento de Linguagem Natural (PLN) como análise dos N-gramas, modelagem de tópicos, algoritmos de aprendizado de máquina e construção do léxico do domínio, como alternativas para compreender as principais necessidades e características da área de domínio especificado. No entanto, nota-se uma escassez de trabalhos com foco na área de análise de vagas de emprego relacionadas ao CRM, além do que, embora os estudos utilizem metodologias para obter conhecimentos dos dados, percebe-se a falta do uso de técnicas baseadas em IA. Diversos estudos utilizam algoritmos de modelagem de tópicos como LDA, LSA e NMF em detrimento a métodos mais modernos, como o BERTopic, utilizado neste estudo e que é capaz de fornecer tópicos mais contextualmente específicos. Dessa forma, o diferencial desse trabalho consiste em preencher essas lacunas por meio de uma análise de vagas de emprego *online* utilizando uma base de dados de larga escala com dados de diferentes sites de empregos, com destaque para as principais necessidades do mercado de trabalho no setor do CRM. Para tanto, utilizaram-se abordagens baseadas em IA e mineração de texto identificadas na literatura, bem como a análise de N-gramas, utilização do algoritmo de modelagem de tópicos BERTopic e a construção de um léxico orientado para a área de interesse.

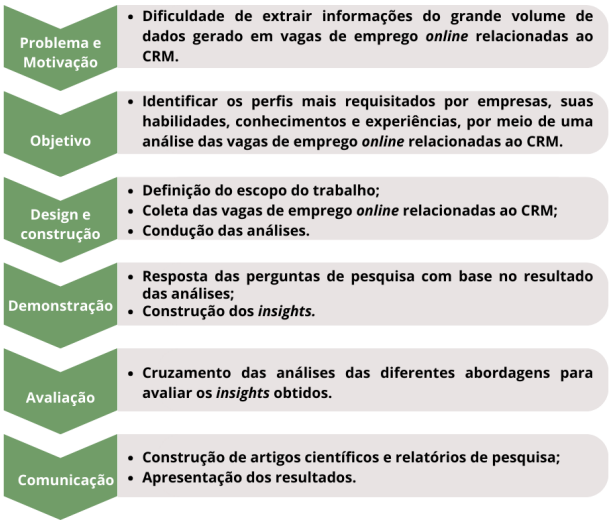


Figura 1: Etapas do *Design Science Research*.

3 MATERIAIS E MÉTODOS

No presente estudo, utilizou-se o método de pesquisa *Design Science Research* por melhor se adequar ao escopo da pesquisa. O DSR consiste em uma metodologia bastante utilizada, sobretudo, na área de Sistemas de Informação e seu processo envolve o desenvolvimento de soluções para problemas organizacionais relevantes, o artefato produzido deve ser avaliado rigorosamente quanto à sua utilidade, qualidade e eficiência [11, 17]. Conceitualmente, um artefato produzido sob a DSR pode ser qualquer objeto no qual uma contribuição de pesquisa está nele embutido. Neste contexto, este estudo foi conduzido sob a perspectiva de um artefato inovador que permite a análise de dados textuais de vagas de empregos *online*.

Para analisar e avaliar o artefato construído sob a ótica de SI, utilizou-se o *Design Theory* em consonância ao DSR, assim como descrito no trabalho de [24]. A ideia de aplicar a *Design Theory* no contexto deste estudo foi definida por ser uma teoria que não apenas informa o design de um artefato inovador, mas também é um resultado central do DSR. Todos os passos utilizados seguindo essa abordagem são apresentados na Figura 1.

3.1 Problema e Motivação

Conforme apresentado na Figura 1, a primeira etapa consiste na identificação do problema e motivação, onde delimitamos o estudo em um ou mais problemas a serem enfrentados, apresentando a respectiva motivação da ação. Neste estudo, a identificação do problema se deu por meio de uma busca na literatura correlata para identificar o atual cenário de vagas trabalhos relacionadas a CRM. Considerando a multitude de plataformas de recrutamento e o grande volume de dados gerados, a extração de conhecimento para acompanhar as necessidades e tendências mercadológicas mostrou-se um desafio de pesquisa relevante. A investigação também foi motivada por uma solicitação de um instituto de pesquisa Alemão,

o *Social CRM Research Center*[2], que busca atualizar os cursos disponibilizados, como também visa ampliar o portfólio de serviços oferecidos.

3.2 Objetivo

No âmbito desta pesquisa, os objetivos da solução foram delineados a partir da identificação do problema e do conhecimento do que é possível e viável de ser implementado. Neste sentido, percebeu-se que técnicas baseadas em mineração de texto podem ser de grande benefício para os interessados. A construção de um léxico orientado pode ser utilizado para facilitar na identificação e estudo sistemático de noções relacionadas ao fator humano e seu papel no ambiente profissional, como a *expertise*, as competências, as habilidades sociais. Frente ao exposto, a realização desta análise está relacionada às mudanças nas organizações, que indicam um aumento da importância das competências no processo de recrutamento.

3.3 *Design* e Construção

A coleta dos dados foi realizada por meio da plataforma *Diffbot* [20], uma vez que essa plataforma é capaz de coletar dados de diferentes sites de empregos. Como *string* de busca foram utilizadas as palavras-chave *"Customer Relationship Management"* e *"CRM"*. Neste sentido, foram coletados 6.400 anúncios de empregos de 929 plataformas diferentes. Os dados foram baixados em formato de arquivo do tipo *Comma-Separated Values* (CSV). Para entender a composição da base de dados, vários aspectos foram avaliados, incluindo o volume de dados e dados ausentes, para obter uma noção precisa de sua dimensionalidade. A Tabela 1 exibe os campos presentes na base de dados, incluindo identificador, título, texto, idioma, *Uniform Resource Locator* (URL), requisitos e tarefas. Vale destacar que alguns destes campos possuíam valores faltantes, sendo este um aspecto importante para determinar a dimensionalidade dos dados.

Convém destacar que o *Id* é um código gerado automaticamente para cada entrada. Algumas vezes a saída da ferramenta duplica o mesmo registro - em inspeção, identificamos que isso ocorre quando a vaga possui mais de uma das palavras-chave utilizadas na busca. Visando assegurar a confiabilidade dos resultados, tal fenômeno foi tratado na etapa de pré-processamento, a qual é detalhada na subseção a seguir.

3.3.1 Preparação dos dados. Uma das etapas mais importantes para a construção das análises do projeto foi o processo de limpeza dos dados. Para tal, foi utilizado um *pipeline* de pré-processamento já consolidado na literatura e adaptado de [9, 10, 13]. As principais etapas de pré-processamento aplicadas aos dados textuais foram:

- **i) Remoção de duplicatas:** a remoção de anúncios duplicados baseou-se na verificação repetida do identificador (*id*) e do endereço da página (*PageUrl*) para garantir que cada anúncio aparecesse apenas uma única vez no conjunto de dados. Houve casos onde o anúncio aparece múltiplas vezes, pois a empresa anuncia em diversas plataformas, estes casos não foram considerados como duplicidade e foram mantidos para as análises subsequentes;
- **ii) Padronização:** O texto foi convertido para letras minúsculas para evitar discrepâncias devido a diferenças nas letras maiúsculas e para reduzir a dimensionalidade, prática comum em mineração de textos. Os anúncios em idiomas diferentes do inglês também foram removidos - esta decisão visou a não imputação de potencial viés decorrente da tradução automática;
- **iii) Remoção de acentos, pontuação, caracteres especiais, URLs e espaços excessivos:** Estas etapas foram aplicadas para eliminar elementos que não contribuem para a análise;
- **iv) Remoção de *stopwords*:** Foram utilizadas as *stopwords* da biblioteca *Natural Language Toolkit* (NLTK), o que permitiu o acesso a uma lista predefinida de *stopwords* em inglês removidas do texto. Além de usar o conjunto de palavras irrelevantes do NLTK, outras palavras foram adicionadas manualmente à lista. Destaca-se que palavras como *"Customer"*, *"Relationships"*, *"Management"* e *"Job"* foram removidas dado sua frequência esperada nas análises.
- **v) Remoção de caracteres numéricos:** Esta etapa foi aplicada somente após avaliar os anos de experiência. A remoção dos caracteres numéricos se mostrou necessária para obter resultados mais precisos na modelagem de tópicos.

Reitera-se que para algumas análises específicas os caracteres especiais e números foram mantidos. Por exemplo, na busca por tempo de experiência foi importante os números. Estes também eram importantes para determinar versões de *software* etc. No entanto, conforme mencionado acima, números e caracteres especiais impactavam negativamente na modelagem de tópicos, por este motivo eles foram removidos, em específico, para esta análise. A Tabela 2 mostra as principais etapas de pré-processamento aplicada a uma amostra de texto de anúncio retirada da base de dados.

3.3.2 Modelagem. No âmbito geral das análises, as quais envolveram a extração de termos e expressões que representam de forma concisa e adequada o conteúdo do *post*, em um primeiro momento, todas as palavras podem ser utilizadas para representar a informação textual. No entanto, para reduzir a ambiguidade e manter um conjunto conciso, foram aplicadas técnicas como radicalização e lematização [1]. Além disso, outra técnica aplicada é a extração de termos compostos (expressões), como N-gramas, que identificam quando dois ou mais termos podem ser unidos [10, 22]. Também vale destacar o uso de modelos de linguagem treinados em grandes conjuntos de informação externa como *Wikipedia* e notícias [28]. Exemplos desses modelos são o *Word2Vec* e *FastText*, disponibilizados pelo *Google* e *Facebook*, respectivamente, bem como modelos mais recentes, como o BERT [19]. Neste estudo, além da técnica de N-gramas, foi utilizado o algoritmo de modelagem de tópicos baseado em *Word Embeddings*, BERTopic.

O BERTopic foi considerado para as análises por ser o mais eficiente para descrições de vagas de empregos que possuem dados não estruturados. Essa abordagem oferece vários benefícios quando usadas em algoritmos de agrupamentos, incluindo desempenho aprimorado, captura de informações contextuais, flexibilidade, personalização e suporte multilíngue. Essas vantagens tornam o

[2]https://scrc-leipzig.de/

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

Campos	Descrição	Quantidade
Id	Identificador numérico único gerado automaticamente pela ferramenta de coleta	6.525
Title	Título do anúncio de emprego correspondente	5.662
Text	Corpo da descrição do anúncio, com todas as informações da vaga	6.525
Humanlanguage	Idioma do texto	6.525
PageUrl	URL de origem do anúncio	6.525
Requirements	Informações relacionadas aos requisitos da vaga	2.794
Tasks	Informações sobre as tarefas relacionadas a vaga	3.699

Tabela 2: Exemplo das etapas de pré-processamento.

Método	Saída
Texto original	*On average, Public Relations Managers earn $57,262 per year.*
Padronização para caixa baixa	*on average, public relations managers earn $57,262 per year.*
Remoção de pontuação, acentuação e espaços excessivos	*on average public relations managers earn $57,262 per year*
Remoção de números e caracteres especiais	*on average public relations managers earn per year*
Remoção de *stopwords*	*average public relations managers earn per year*

BERTopic uma ferramenta poderosa para agrupamento em tarefas de PLN [32]. Este modelo, por ser baseado em *Word Embeddings*, representa uma estratégia para aprendizado de representações na área de PLN, no qual palavras (ou frases) de um conjunto de textos são mapeadas para vetores numéricos *(word vectors*). A partir dos vetores numéricos das palavras (ou frases), é possível realizar operações de álgebra vetorial para identificar termos correlacionados. Por exemplo, é possível identificar que os termos *"smartphones"* e "celulares" são muito próximos, algo que não é possível por comparação literal. Do ponto de vista de mineração de *posts* em mídias sociais de anúncios de empregos, *word embeddings* podem ser aplicados para expansão de vocabulário do domínio, bem como identificar relacionamentos entre termos e expressões, como gírias, apelidos e abreviações. A terceira análise conduzida consistiu na construção de um léxico relacionado às vagas de emprego, o qual envolveu uma abordagem semi-automática, conforme mostra a Figura 2.

Esta etapa foi desenvolvida considerando os dois principais tipos de competências identificadas no mercado de trabalho digital: competências interpessoais *(soft skills)* e competências técnicas *(hards skills).* As habilidades técnicas são aquelas que uma pessoa pode adquirir por meio de processos formais de aprendizagem, como por exemplo, como usar uma linguagem de programação [29]. Já as competências interpessoais, como o trabalho em equipe ou comunicação podem desempenhar um papel decisivo na determinação da correspondência qualitativa entre um candidato e uma vaga de emprego aberta, ou ainda a quantidade de tempo que o trabalhador permanece no emprego [5, 33].

Inicialmente foram definidos os eixos, levando em consideração os elementos existentes na estrutura do anúncio, que resultou em vagas, tecnologias, habilidades (técnicas e comportamentais) e conhecimentos. Em seguida foi realizada a leitura em um conjunto representativo das vagas. Essa etapa foi combinada com um *script* de pesquisa de termos relacionados com palavras-chave observadas durante a leitura, que são: *"platforms", "platform", "tools", "software", "knowledge", "certification", "plus", "ability", "required", "programs" e "experience"*. Esse processo culminou na criação de um dicionário que facilitou a identificação e visualização dos temas abordados nos anúncios de emprego. Por fim, os termos classificados no dicionário foram utilizados para identificar os mais recorrentes nos anúncios.

Para a implementação dos *scripts* para as análises destacadas foi utilizada a linguagem Python e os experimentos foram conduzidos na plataforma *Google Colab*. Na etapa de pré-processamento dos dados foram utilizadas algumas bibliotecas disponibilizadas pela linguagem, bem como Pandas na sua versão (1.4.4); NLTK (3.7); Expressões Regulares foram implementadas utilizando a biblioteca *RE* (2022.7.9). Para a modelagem de Tópicos, conforme mencionado anteriormente, foi utilizado o BerTopic[3] com sua versão (0.15.0). Para visualização e entendimento dos dados, matplotlib/pyplot (3.5.2) e Seaborn(0.11.2) para criação de gráficos e nuvem de palavras[4] (1.8.2.2). Convém destacar que a avaliação dos resultados foi feito em momento posterior, o qual está descrita na Subseção a seguir.

3.4 Demonstração e Avaliação

Na demonstração, é onde retorna-se aos objetivos iniciais e reavalia-se as perguntas de pesquisa inicialmente propostas. Com a definição do escopo, a lista de *insights* dos dados e critérios de exclusão foi definida, a saber:

- Quantidade de plataformas de empregos consideradas relevantes e irrelevantes para o estudo;
- Profissionais mais requisitados apenas no mercado CRM;
- Habilidades pessoais e interpessoais dos candidatos;
- Anos de experiências do candidato;
- Área de atuação do candidato.

Esta etapa foi dividida entre a análise exploratória, modelagem de tópicos e N-gramas visando responder a *PP-1* e *PP-2* e a construção do léxico aplicado à amostra dos dados a fim de responder a *PP-3*.

[3]https://maartengr.github.io/BERTopic/index.html
[4]https://pypi.org/project/wordcloud/

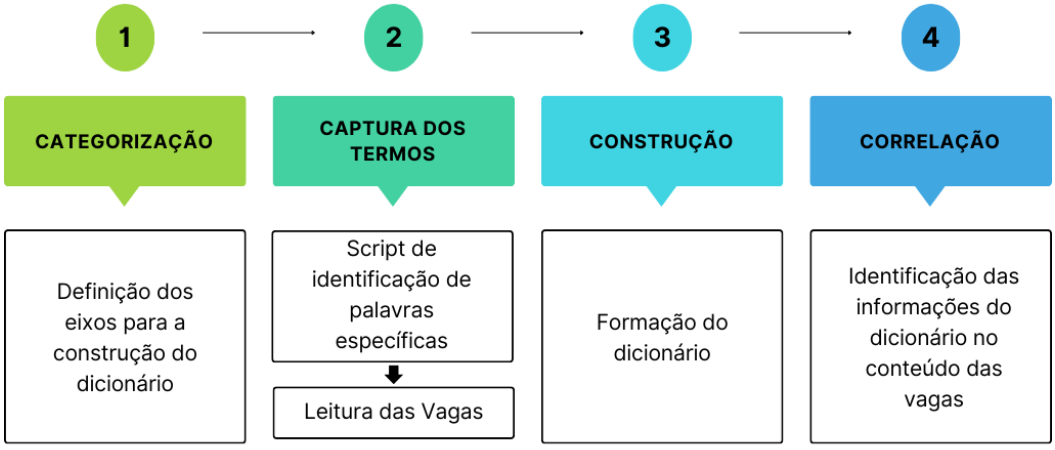


Figura 2: Etapas para a construção do léxico.

Após estes passos, foi construído um relatório preliminar, do tipo *slidument*, que foi apresentado para o grupo de pesquisa, condensando os resultados. Nele foram identificados as inconsistências dos resultados, o que motivou a revisão do *pipeline* de análises, que resultou na consolidação de um *pipeline* mais robusto e que foi descrito nas seções anteriores. Com os novos resultados, foi realizada uma segunda apresentação, que validou os resultados obtidos e os *insights* foram considerados significativamente relevantes, passando-se para a etapa de avaliação.

A etapa de avaliação se baseou na abordagem da teoria fundamentada, que consiste em um método de pesquisa qualitativa que permite aos pesquisadores discernir processos explícitos e implícitos em seus dados [2, 8]. Neste estudo, a abordagem foi utilizada para que os especialistas do domínio, baseando-se nesta teoria, lessem uma amostra representativa de dados e cruzassem com os resultados obtidos. Vale ressaltar também que foi feito um cálculo amostral, considerando o tamanho da população, um nível de confiança de 95% e uma margem de erro de 5.

3.5 Comunicação

Por fim, tem-se a comunicação. Esta fase visa dar publicidade ao problema e sua importância, ao artefato, sua utilidade e inovação, o rigor de seu projeto e efetividade de uso. Os principais documentos construídos nesta etapa são artigos científicos, relatórios técnicos, documentos de registro de programa de computador, apresentações em conferências e entrevistas.

4 RESULTADOS E DISCUSSÕES

Após aplicar a etapa de pré-processamento nos dados coletados, obteve-se um total de 6.243 amostras de vagas de empregos relacionadas CRM. Foram utilizadas técnicas de visualização de dados para uma melhor compreensão dos principais termos presentes nas descrições das vagas, bem como mostra a Figura 3.

Analisando a Figura 3, é possível identificar a predominância de palavras relacionadas ao tempo de experiência, habilidades pessoais e qualificação do candidato. Destaca-se termos como *"full time"*, *"hours credit"*, *"credits course"* que sugerem requisitos relacionados ao tipo de posição e formação acadêmica do candidato. Além disso, tais termos podem indicar a tendência de posições que demandam profissionais com sólida experiência e reconhecimento no mercado, bem como horas de cursos, formações e outros tipos de atividades relacionadas ao cargo. Por outro lado, os termos *"equal opportunity"* *"sexual orientation"*, *"gender identity"* e *"origin"* sugerem características como preferência por equipes diversas, preocupação com equidade de gênero, inclusão da comunidade LGBQIAP+ e diversidade de raça/cor. Esses aspectos mostram que as empresas estão adotando um ambiente cada vez mais inclusivo com oportunidades iguais para todos os candidatos.

A Figura 4, apresenta os anos de experiência mais requisitados nas vagas, as análises mostram que as empresas buscam por candidatos com mais tempo de experiência, embora também se possa perceber a prevalência de candidatos com dois anos de experiência, as hipóteses iniciais são de que esses candidatos surgem com as novas profissões e demandas do mercado mais atual, bem como na área de tecnologias que por serem profissões mais recentes no mercado, requerem profissionais com menos tempo de experiência.

A Figura 5 apresenta os N-gramas que tiveram mais ocorrência nos títulos das vagas. Tal análise destaca os profissionais mais requisitados, bem como *"account manager"*, *"manager jobs"*, *"project manager"* e *"success account manager"*. Com base nas especificações de tais vagas, pode-se inferir que essas vagas estão diretamente ligadas ao CRM ou ao Social CRM.

Com base na hipótese da presença de áreas especificas descritas nos títulos das vagas, bem como mostra a Figura 5, foi desenvolvido o léxico para categorizar essas áreas e identificar sua natureza. A Figura 6 mostra os principais termos que denotam áreas no léxico construído, com ênfase para áreas de tecnologias, como *"technology specialist"*, *"marketing technologist"*, *"data analyst"* e *"technology*

Figura 3: Nuvem de palavras das descrições das vagas.

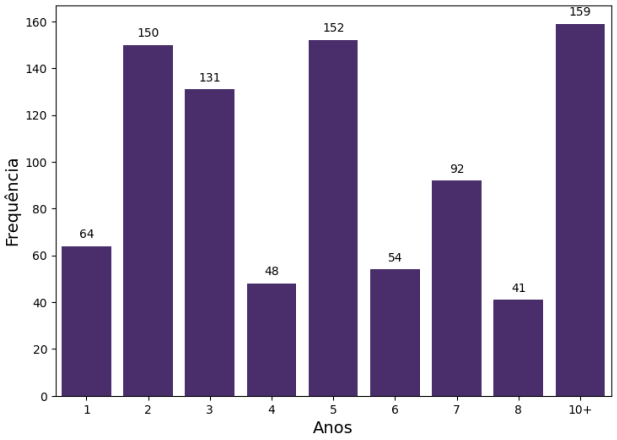


Figura 4: Tempo de experiência exigido nas vagas envolvendo CRM.

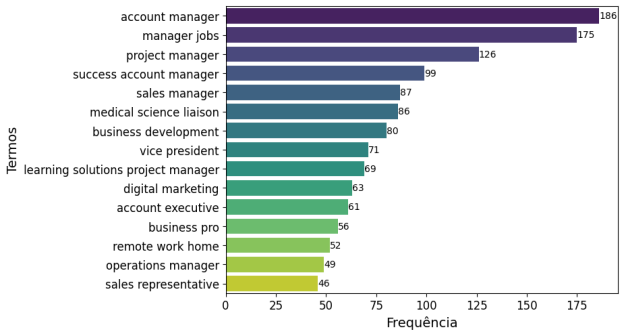


Figura 5: N-gramas com base nos títulos das vagas.

manager". Fazendo um paralelo com os termos da Figura 5, percebe-se que além de profissionais ligados diretamente ao CRM, também há uma grande procura por profissionais de outras áreas. Esses *insights* estão em consonância com as análises exploratórias inicias e respondem a *PP-1* sobre como as empresas selecionam profissionais neste setor, que levam em consideração não apenas aspectos como inclusão social; orientação sexual; identidade de gênero como também fatores relacionados a formação, tempo de experiência e área do candidato. Desse modo, estes resultados também respondem diretamente a *PP-2* acerca dos aspectos que as empresas consideram para selecionar os melhores perfis de profissionais.

Neste ensejo, a grande tendência mercadológica está voltada aos profissionais de áreas específicas de tecnologias, bem como na grande área de análise de dados, o que pode ser corroborado em trabalhos vistos anteriormente [36]. A grande procura por profissionais desta área está relacionada diretamente ao crescimento exponencial das plataformas, culminando em grande volume de dados gerados por clientes, os quais podem ser de grande ajuda para as empresas no entendimento do comportamento do cliente

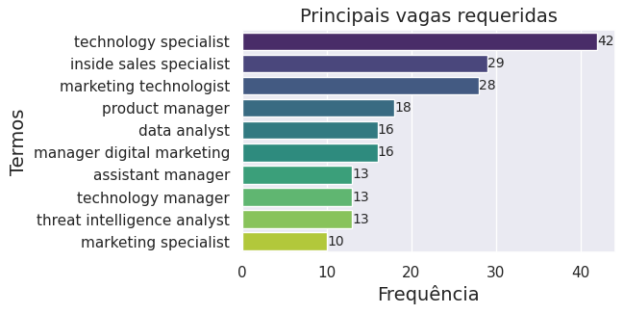


Figura 6: Principais vagas relacionadas ao CRM.

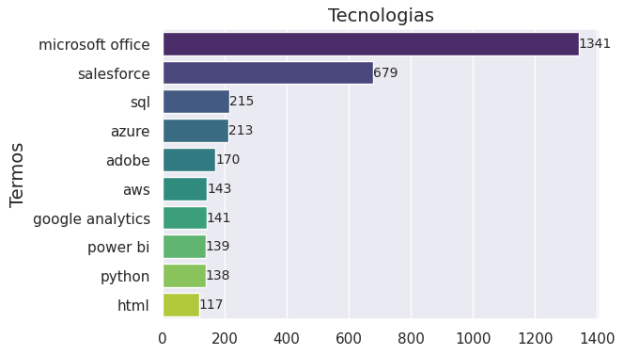


Figura 7: Principais tecnologias requeridas em vagas relacionadas ao CRM.

e ganhar valor, especialmente em vendas, atendimento ao cliente, *marketing* e gestão de promoções.

Além das áreas específicas identificadas, ainda foi possível expandir o dicionário construído para identificar e categorizar as tecnologias mais aparentes no processo de recrutamento. A Figura 7 destaca as tecnologias mais requisitadas descritas nas descrições das vagas. Ferramentas como *"microsoft office"* ainda são predominantes, sugerindo a preocupação por profissionais que possuem conhecimentos para processamento de texto, planilhas, apresentações e comunicação, sendo este o conjunto de ferramentas padrão em diversas empresas.

A presença do Termo *"salesforce"* refere-se a plataforma online voltada ao atendimento ao cliente, gestão de comunidade, *marketing*, IA entre outros no setor CRM. Além disso, destaca-se fortemente a presença de tecnologias ligadas a grande área de Análise de Dados e desenvolvimento de Software citadas em análises anteriores, bem como *"sql"*, *"google analytics"*, *"power bi"*, *"python"* e *"html"*. A grande procura por profissionais que possuem estes conhecimentos sugere a escassez destes no mercado atual. Esses *insights* respondem a *PP-3* sobre as habilidades e conhecimentos de destaque e são fundamentais para direcionar profissionais sobre as tecnologias mais requisitadas no mercado de CRM e Social CRM.

A Figura 8 mostra as principais habilidades identificadas na construção do léxico. Essa análise foi dividida entre as categorias de habilidades interpessoais *(soft skills)* e habilidades técnicas *(hard skills)*. Destacam-se, em *hard skills*, as habilidades de *marketing* e *design*, além de habilidades relacionadas a área de IA e estatística. Pode-se fazer um cruzamento com as análises da Figura 6, onde mostra a prevalência de profissionais com conhecimentos voltados para ferramentas destas áreas. Além disso, como habilidades pessoais (*Soft Skills)*, percebe-se a prevalência dos termos *"communication"* *"organization"* e *"written"*, fazendo um paralelo também com as análises mostradas na Figura 6, onde ficam evidentes as habilidades e conhecimentos relacionados a escrita e organização. Além disso, a busca por profissionais com capacidade de uma boa comunicação é altamente valorizada no setor de CRM, dado que o setor este mercado está intrinsecamente ligado à comunicação direta e efetiva entre empresas e seus clientes.

Na etapa de modelagem de tópicos, utilizou-se o algoritmo BERTopic para identificar os cinco tópicos mais frequentes nas descrições das vagas. Como padrão o modelo gera um grande número de tópicos, dado a repetição de alguns destes, fez-se a redução para os mais frequentes, vale destacar que as palavras de cada tópico estão em ordem de frequência, portanto, os nomes dos tópicos foram definidos com base nas primeiras palavras dos tópicos. A Tabela 3 mostra os os tópicos reduzidos com 10 termos por tópicos. Os tópicos são descritos a seguir:

- **Tópico 1: Gestão e controle de equipes**: - Este tópico indica que empresas podem estar em busca de profissionais com habilidades de gerenciamento e análise de dados. Além disso. Os termos *"sales"*, *"experience"* e *"skills"* podem indicar a presença de termos relacionados a experiência, habilidades e salário, que são importantes fatores para melhores perfis de profissionais que denotam a procura por profissionais as melhores habilidades e conhecimentos para o setor de vendas;
- **Tópico 2: CRM e i** - Este tópico está relacionado diretamente tanto com o CRM quanto com Social CRM, onde percebe-se a preocupação em contratar profissionais que possuem experiência perante a análise de informação de clientes, bem como gestão de vendas e *marketing*. A área do CRM pode estar relacionada ao gerenciamento de dados do cliente, análise de métricas de *marketing* e implementação de processos eficientes para melhorar o relacionamento com os clientes;
- **Tópico 3: Formação e Academia** - este tópico sugere elementos relacionados ao processo de formação acadêmica, com destaque para *"humboldt"*, *"campus"*, *"state"* e *"position"*. Que demonstram a preocupação, vista em análises anteriores, onde as corporações favorecem profissionais com boas qualificações, horas de cursos e certificados no processo de recrutamento;
- **Tópico 4: Equipes e grande demanda** - Este tópico selecionado pode também ser cruzado com análises anteriores, principalmente com relação aos *insights* obtidos com a construção do léxico mostrado na Figura 6, onde há a predominância de ferramentas voltadas para área de programação e soluções computacionais. Dado este cenário, este tópico mostra termos referentes a contração de profissionais na área de Tecnologias. Tais análises são cruciais para identificar as áreas de maior impacto neste setor;

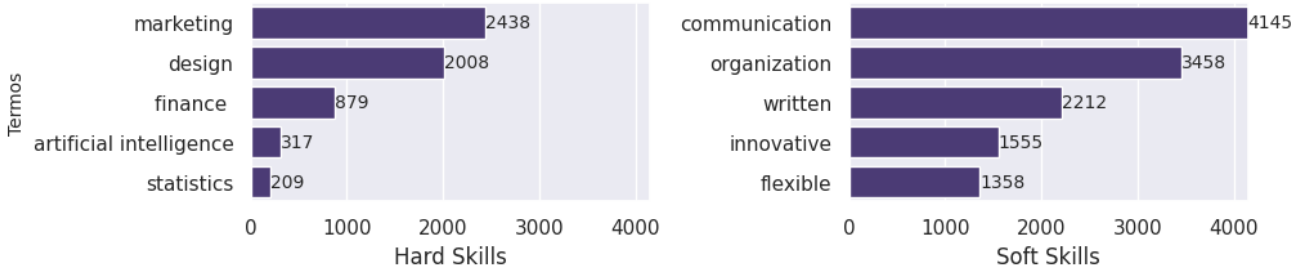


Figura 8: *Hard* e *Soft skills* identificadas nos anúncios de vagas de emprego.

Tabela 3: Representação dos tópicos com os termos selecionados

Título	Termos
Gestão e controle de equipes	*business,* ***sales****,* ***experience****, work, team,* ***skills****, support, marketing, new, data*
CRM e *Marketing*	***crm****, sales,* ***marketing****, customers, information,* ***business****, software, contact, systems*
Formação e Academia	***humboldt****, poly, cal, recruitment, staff, demonstrated,* ***campus****, outreach,* ***state****,* ***position***
Equipes e grande demanda	ciff, srhr, **programmes**, team, reproductive, support, **programme**, health, **portfolios**, teams
Geral	*service, representative, job, ago, hour, call, agodescription, jobs, ca, results*

- **Tópico 5: Geral** - O último tópico possui termos que denotam contexto mais genérico, bem como hora, serviço e trabalho. O qual pode ser termos esperados em grandes quantidades ao considerarmos o contexto dos dados.

Em cômputo geral, os resultados obtidos respondem as perguntas de pesquisa, evidenciando aspectos ligados às áreas correlatas ao CRM, bem como tecnologias e habilidades mais requisitadas pelas empresas. Foi possível fazer uma avaliação por meio do cruzamento das diferentes abordagens utilizadas neste estudo para obter melhores *insights* e, com a análise baseada em Teoria Fundamentada, foi possível avaliar a ferramenta desenvolvida para construção do artefato que permita a automação de análise de dados textuais voltadas para o setor CRM. Por meio da análise de N-gramas e modelagem de tópicos, foi possível vislumbrar a diversidade e as nuances das oportunidades profissionais neste campo. As revisões acima fornecem uma visão geral dos requisitos e áreas de foco, que vão desde gestão de negócios e *marketing* digital até análise de dados. Dessa forma, foi construído um léxico para categorizar e organizar essas informações complexas para obter uma visão mais precisa e estruturada do cenário de emprego no CRM. Além disso, neste estudo utilizou-se uma abordagem baseada na Teoria de Sistema de Informação *(Design Theory)* em consonância ao DSR, a qual foi adaptada para este estudo, principalmente no que tange a construção do artefato supracitado, o qual norteou todos os passos seguidos neste estudo.

5 CONSIDERAÇÕES FINAIS

O presente trabalho apresentou a análise de 6.400 anúncios de vagas de emprego ligadas à gestão de relacionamento com clientes - o CRM. As análises conduzidas visavam identificar e distinguir habilidades, conhecimentos, cargos e tecnologias presentes nas chamadas de trabalho. Foram utilizadas boas práticas de ciência de dados, que permitiram a análise dos dados centrada na abordagem de N-gramas, modelagem de tópicos e a construção de um léxico orientado para o domínio do CRM. Além disso, são destacados os conceitos relacionados à teoria fundamentada nos dados para validar os resultados. Essas etapas foram cruciais para compreender a dinâmica, os padrões e as tendências do mercado de CRM.

Em suma, os resultados convergem para identificar uma demanda crescente por profissionais altamente especializados em CRM, que seriam capazes de liderar estratégias digitais e de CRM - respondendo a PP-1 de "Como as empresas selecionam profissionais no setor de CRM?". A modelagem de tópicos também sugere uma demanda crescente por profissionais com habilidades em estratégias digitais, análise de dados e gestão de relacionamento com o cliente - respondendo as PP-2 e PP-3, respectivamente: "Quais aspectos as empresas consideram para selecionar os melhores perfis de profissionais?" e "Quais as habilidades e conhecimentos mais requisitados no cenário atual do mercado em CRM?".

Os resultados obtidos têm o potencial de orientar as pessoas que desejam ingressar nesta área. Também facilita a promoção de políticas públicas voltadas para esse mercado de trabalho. A mineração de textos e a análise de dados tornaram-se ferramentas poderosas para compreender a dinâmica do mercado de trabalho e tomar decisões importantes sobre contratação e qualificação de profissionais nesta área em constante mudança. Neste ensejo, o presente trabalho contribui para os novos desafios na área de SI, fornecendo uma perspectiva inovadora para lidar com tais mudanças [16]. Em trabalhos futuros pretende-se ampliar a coleta de dados, incluindo outros termos de busca. Além disso, também de refinar o léxico convidando especialistas em Recursos Humanos para validar o dicionário. Finalmente, pretende-se aplicar mais métodos de mineração de texto, como *k-means* para agrupar os dados, e grandes modelos de linguagem (por exemplo, GPT) para testar uma abordagem de perguntas e respostas.

AGRADECIMENTOS

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)-DT-303031/2023-9, PIBIC - 161082/2023-8; e pela Fundação Amazônia de Amparo a Estudos e Pesquisas (FAPESPA) PRONEM-FAPESPA/CNPq nº 045/2021.

REFERÊNCIAS

[1] Charu C Aggarwal and Charu C Aggarwal. 2015. *Mining text data*. Springer.
[2] Antonio de Lucas Ancillo, María Teresa del Val Núñez, and Sorin Gavrila Gavrila. 2021. Workplace change within the COVID-19 context: a grounded theory approach. *Economic Research-Ekonomska Istraživanja* 34, 1 (2021), 2297–2316.
[3] Muhammad Anshari, Mohammad Nabil Almunawar, Syamimi Ariff Lim, and Abdullah Al-Mudimigh. 2019. Customer relationship management and big data enabled: Personalization & customization of services. *Applied Computing and Informatics* 15, 2 (2019), 94–101. https://doi.org/10.1016/j.aci.2018.05.004
[4] Barbara AP Barata, Adrielson F Justino, Antonio FL Jacob Junior, and Fábio MF Lobato. 2023. What about data science? An analysis of the market based on Job posts. In *Anais do XX Encontro Nacional de Inteligência Artificial e Computacional*. SBC, 824–838.
[5] Maite Blázquez. 2014. Skills-based profiling and matching in pes. *Publications Office of the European Union, Luxembourg* (2014).
[6] Clodis Boscarioli, Renata Mendes de Araujo, Rita Suzana Maciel, Valdemar Vicente Graciano Neto, Flavio Oquendo, Elisa Yumi Nakagawa, Flavia Cristina Berrnardini, José Viterbo, Dalessandro Vianna, Carlos Bazilio Martins, et al. 2017. I GranDSI-BR: Grand Research Challenges in Information Systems in Brazil 2016-2026. (2017).
[7] Douglas Charcon, Nizam Omar, and Luiz Henrique Alves Monteiro. 2022. On Using Artificial Intelligence in the Search of the Best Professional Resumes. In *XVIII Brazilian Symposium on Information Systems*. 1–7.
[8] Kathy Charmaz and Robert Thornberg. 2021. The pursuit of quality in grounded theory. *Qualitative research in psychology* 18, 3 (2021).
[9] Douglas Cirqueira, Fernando Almeida, Gültekin Cakir, Antonio Jacob, Fabio Lobato, Marija Bezbradica, and Markus Helfert. 2020. Explainable sentiment analysis application for social media crisis management in retail. (2020).
[10] Douglas Cirqueira, Márcia Fontes Pinheiro, Antonio Jacob, Fábio Lobato, and Ádamo Santana. 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)*. IEEE, 746–749.
[11] Tadeu Classe, Sean Siqueira, Renata Araujo, and Geraldo Xexéo. 2020. Play Your Process - Uma Método de Design de Jogos Digitais Baseados em Modelos de Processos de Negócio. In *Anais Estendidos do XVI Simpósio Brasileiro de Sistemas de Informação* (Evento Online). SBC, Porto Alegre, RS, Brasil, 142–157. https://doi.org/10.5753/sbsi.2020.13136
[12] Felipe Penhorate Carvalho da Fonseca and Luciano Antonio Digiampietri. 2021. Improving researcher's area of expertise identification using TF-IDF Characters N-grams. In *XVII Brazilian Symposium on Information Systems*. 1–7.
[13] Marcos VJ da Silva, Ewaldo E Santana, Fábio MF Lobato, and Antonio FL Jacob Jr. 2023. Preprocessing Applied to Legal Text Mining: analysis and evaluation of the main techniques used. In *Anais do XX Encontro Nacional de Inteligência Artificial e Computacional*. SBC, 1010–1021.
[14] Ana Elisa da Silva Cunha, Bruno BP Cafeo, Davi Viana, and Awdren Fontão. 2022. A Survey on Skills of DevRel professionals. In *Anais do XVIII Simpósio Brasileiro de Sistemas de Informação*. SBC.
[15] Carolina Coelho da Silveira, Carla Bonato Marcolin, Matheus da Silva, and Jean Carlos Domingos. 2020. What is a Data Scientist? Analysis of core soft and technical competencies in job postings. *Revista Inovação, Projetos e Tecnologias* 8, 1 (2020), 25–39.
[16] Renata Mendes de Araujo. 2017. Information systems and the open world challenges. *Sociedade Brasileira de Computação* (2017).
[17] Gustavo Nogueira de Sousa, Luan Vinícius Huppes, Antônio Fernando Lavareda Jacob Jr, and Fábio Manoel França Lobato. 2018. Ado\c {c}\~ao de Social CRM em Micro e Pequenas Empresas: Uma An\'alise do Mercado Santareno. *arXiv preprint arXiv:1811.11821* (2018).
[18] Dai Debao, Ma Yinxia, and Zhao Min. 2021. Analysis of big data job requirements based on K-means text clustering in China. *PloS one* 16, 8 (2021), e0255419.
[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[20] Diffbot. 2023. *Diffbot - Web Data for your AI*. https://www.diffbot.com/
[21] Fatih Gurcan and Nergiz Ercil Cagiltay. 2019. Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling. *IEEE access* 7 (2019), 82541–82552.
[22] Fatih Gurcan and Nergiz Ercil Cagiltay. 2023. Research trends on distance learning: a text mining-based literature review from 2008 to 2018. *Interactive Learning Environments* 31, 2 (2023), 1007–1028. https://doi.org/10.1080/10494820.2020.1815795 arXiv:https://doi.org/10.1080/10494820.2020.1815795
[23] Raul Oltra-Badenes Hermenegildo Gil-Gomez, Vicente Guerola-Navarro and José Antonio Lozano-Quilis. 2020. Customer relationship management: digital transformation and sustainable business model innovation. *Economic Research-Ekonomska Istraživanja* 33, 1 (2020), 2733–2750. https://doi.org/10.1080/1331677X.2019.1676283 arXiv:https://doi.org/10.1080/1331677X.2019.1676283
[24] Juhani Iivari. 2020. A critical look at theories in design science research. *Journal of the Association for Information Systems* 21, 3 (2020), 10.
[25] Imane Khaouja, Ismail Kassou, and Mounir Ghogho. 2021. A Survey on Skill Identification From Online Job Ads. *IEEE Access* 9 (2021), 118134–118153. https://doi.org/10.1109/ACCESS.2021.3106120
[26] Hari G Krishna, Vyshak Mohan, and N Maithreyi. 2016. Social media recruitment from employers perspective. *International Journal of Applied Business and Economic Research* 14, 14 (2016), 153–166.
[27] Sofia Maria Bouzon Machado and Bruna Diirr. 2023. Professional experience characterization on information systems teams collaboration. In *Proceedings of the XIX Brazilian Symposium on Information Systems*. 468–475.
[28] Sara Meftah and Nasredine Semmar. 2018. A neural network model for part-of-speech tagging of social media texts. In *Proceedings of the eleventh international Conference on Language Resources and Evaluation (LREC 2018)*.
[29] Maria Papoutsoglou, Apostolos Ampatzoglou, Nikolaos Mittas, and Lefteris Angelis. 2019. Extracting knowledge from on-line sources for software engineering labor market: A mapping study. *IEEE Access* 7 (2019), 157595–157613.
[30] Thiago P. Pimentel and Ronaldo R. Goldschmidt. 2019. Sequential Sentiment Pattern Mining to Predict Churn in CRM Systems: A Case Study with Telecom Data. In *Proceedings of the XV Brazilian Symposium on Information Systems* (Aracaju, Brazil) *(SBSI '19)*. Association for Computing Machinery, New York, NY, USA, Article 11, 8 pages. https://doi.org/10.1145/3330204.3330220
[31] Paul Ralph and Sebastian Baltes. 2022. Paving the way for mature secondary research: the seven types of literature review. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 1632–1636.
[32] Samsir Samsir, Reagan Surbakti Saragih, Selamat Subagio, Rahmad Aditiya, and Ronal Watrianthos. 2023. BERTopic Modeling of Natural Language Processing Abstracts: Thematic Structure and Trajectory. *JURNAL MEDIA INFORMATIKA BUDIDARMA* 7, 3 (2023), 1514–1520.
[33] Kai Johannes Schleutker, Valeria Caggiano, Fabiana Coluzzi, and Jose Luis Poza Luján. 2019. Soft skills and European labour market: Interviews with Finnish and Italian managers. *Journal of Educational, Cultural and Psychological Studies (ECPS Journal)* 19 (2019), 123–144.
[34] Narendra Singh, Pushpa Singh, and Mukul Gupta. 2020. An inclusive survey on machine learning for CRM: a paradigm shift. *Decision* 47, 4 (2020), 447–457.
[35] Pilar Talón-Ballestero, Lydia González-Serrano, Cristina Soguero-Ruiz, Sergio Muñoz-Romero, and José Luis Rojo-Álvarez. 2018. Using big data from Customer Relationship Management information systems to determine the client profile in the hotel sector. *Tourism Management* 68 (2018), 187–197. https://doi.org/10.1016/j.tourman.2018.03.017
[36] Peter C Verhoef, Edwin Kooge, and Natasha Walk. 2016. *Creating value with big data analytics: Making smarter marketing decisions*. Routledge.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

APÊNDICE B – ARTIGO PUBLICADO NO *WI-IAT* 2023

2023 IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)

Market Overview in Social CRM: An analysis of job advertisements

Pedro C. Menezes
Federal University of Western Pará
Santarém, PA, Brazil

Adrielson F. Justino
State University of Maranhão
São Luís, MA, Brazil

Barbara A. P. Barata
State University of Maranhão
São Luís, MA, Brazil

Antonio F. L. Jacob Junior
State University of Maranhão
São Luís, MA, Brazil
antoniojunior@professor.uema.br

Fábio M. F. Lobato
Federal University of Western Pará
Santarém, PA, Brazil
fabio.lobato@ufopa.edu.br

2023 IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) | 979-8-3503-0918-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/WI-IAT59888.2023.00092

***Abstract*—Technological advances and constant changes in the market have significantly driven the evolution of Social Customer Relationship Management (Social CRM). As digital platforms and social media gain prominence, companies seek to strengthen their relationship with their customers, making it essential for industry professionals to keep up with trends and adapt to new demands. Given this phenomenon, there is a high frequency of job postings related to Social CRM. Evaluating these advertisements provides an excellent opportunity to keep up with the fast pace of the job market and identify the critical elements for success in this area. In this context, this work aims to collect and analyze job advertisements related to Social CRM and apply text-mining techniques to identify the knowledge domains and skill sets needed by professionals in this field. Through this approach, the intention is to develop a taxonomy that systematically and comprehensively maps the competencies required in this constantly changing market. We analyzed 1,086 job advertisements related to Social CRM. The analysis provided a detailed overview of the required skills, technologies, and knowledge. This information offers good opportunities for employers and candidates in the hiring and recruitment process, allowing both to identify more precisely the essential skills needed to excel in the Social CRM sector. In addition, the results of this research can encourage the further development of studies related to this expanding area, contributing to the enrichment and improvement of knowledge in Social CRM.**

***Index Terms*—Data Analysis, Social CRM, Job Analysis, Text Mining, Market Analysis**

I. Introduction

Companies are looking to strengthen customer relationships with the continuous expansion of digital platforms and social media usage [1], [2]. The Social Customer Relationship Management (Social CRM) concept has been significantly boosted, making it an area of extreme relevance in the current scenario [3]. The increased use of the Internet, particularly social networking platforms such as Facebook, Twitter, Instagram, and Snapchat, has brought several opportunities and challenges for marketers looking to manage customer relationships [4]. Therefore, professionals in this sector need to follow trends and adapt to new demands. In this scenario, Customer Relationship Management (CRM) has gradually evolved to take advantage of new technologies and emerging channels, moving from database-driven CRM to electronic CRM (eCRM), mobile CRM (mCRM), and later Social CRM, the last integrates social media tools and customer management to monitor market analysis.

Social CRM emphasizes active participation in online conversations and communities, enabling companies to monitor and respond to social interactions such as comments, mentions, ratings, and shares [3]. In this way, it is possible to build stronger customer relationships and gain valuable insights into their preferences and opinions [1], [5]. Faced with this phenomenon, a growing number of companies have adopted online recruitment as a way to attract candidates [6]. So, using the internet to publicize job postings offers an excellent opportunity to monitor the job market in real-time [7]. A job vacancy consists of two main elements: a title summarizing the position and a detailed description that includes the job requirements and skills needed by the candidate [7].

Text mining in textual databases has become essential to support knowledge discovery in an agile and cost-effective way, surpassing traditional sources [8]. Applying text mining methods to analyze job posts allows for understanding employers' needs, and identifies skills, occupational changes, and working conditions [9]. Based on these facts, this study aimed to perform a textual content analysis on online job advertisements related to Social CRM to identify the knowledge domains, job requirements, skill sets, and tools needed in this segment.

The remainder of the article is organized as follows. In Section II related works are presented. Section III describes the research process. Section IV presents the results resulting from the analyses. Finally, Section V presents the final considerations.

II. Related Work

[10] focused on proposing a semi-autonomous methodology to analyze job advertisements in software engineering and big data. The authors used the Indeed platform as the data source. Furthermore, Latent Dirichlet Allocation (LDA) was chosen as the most appropriate topic modeling method

979-8-3503-0918-8/23/$31.00 ©2023 IEEE
DOI 10.1109/WI-IAT59888.2023.00092

Authorized licensed use limited to: UNIVERSIDADE DE SAO PAULO. Downloaded on April 06,2024 at 14:28:07 UTC from IEEE Xplore. Restrictions apply.

since this method considers the semantics of advertisements, revealing the knowledge domain and critical skills needed for the job market. In addition, the method used can be improved to automatically analyze ads' text and routinely update the skills list.

[11] aimed to understand the most requested skills in big data job advertisements, considering aspects related to the location (city), salary, educational background, and experience. The data were extracted from a Chinese website called Zhaopin, and the texts were grouped using the k-means algorithm. The work highlights the need to integrate more text mining technologies to explore further the information and conditions of the big data job market.

[12] focused on comparing keyword dictionary methods and the LDA model for content analysis in job advertisements. The authors found that the topic identification algorithm performed better when compared to the approach based on the European Dictionary of Skills and Competencies (DISCO). However, there are disadvantages to adopting the LDA model, such as identifying topics that are difficult to interpret. The authors emphasize the preciousness of these discoveries as a way to support communication between the educational network and the labor market.

[13] sought a broader understanding of the requirements related to the spheres of Information Technology in the Commonwealth of Independent States region by mapping the most accurate combinations of soft and hard skills. The methodology proposed in the work is based on natural language processing, hierarchical grouping, and association mining, showing that the methods used can be put in a theoretical and practical way.

The related works have shown how techniques such as N-gram, topic modeling, clustering, and natural language processing are pertinent to reveal important information in many fields. In addition, the works show the importance of the information obtained by analyzing the data in the alignment of education and the labor market. It is essential to point out that a limitation is noted regarding the analysis of online job openings in Social CRM about the needs of the labor market. Thus, the differential of this work consists of a pioneering and comprehensive approach in the analysis of job openings, highlighting the job market's main requirements in the Social CRM field.

III. Materials and Methods

The research method adopted in this work is the Cross-Industry Standard Process for Data Mining (CRISP-DM) used in [14]. This process model is well established for conducting data mining projects [15]. The application of this model is divided into six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Implementation, as shown in Figure 1.

A. Business understanding

In the first stage, the business or application domain was understood. For this, it is necessary to understand the project's goals and identify the research problems based on data mining. It is noteworthy that the definition of the research goals in this study was carried out together with a literature review to obtain information on the role of professionals, techniques, and procedures for mapping the knowledge domains and skills required in the market in Social CRM, according to the related works.

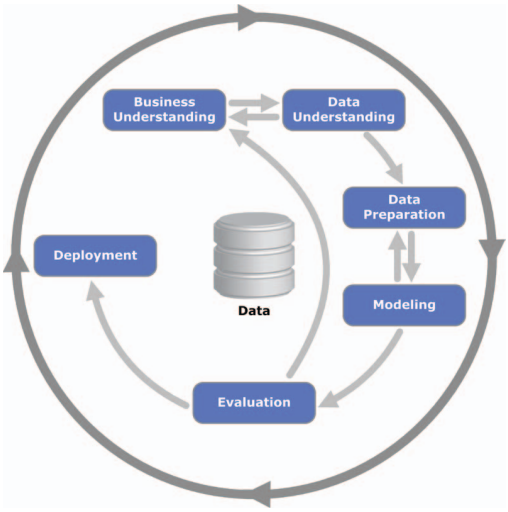


Fig. 1. Phases of the CRISP-DM method. Adapted from [16].

At this stage, an initial data collection was carried out to familiarize them with their characteristics, identifying problems and insights about the data. Poorly designed data is challenging to extract and use on the web in an automated process [17]. In this sense, we use Diffbot, a web scraping tool that can easily collect data from web pages, convert it into some desired format, and store it in some local repository. Through it, it is possible to collect information from various sources on the internet, including news, articles, products, social media, and job vacancies. The platform automatically identifies patterns in the data, such as titles, descriptions, and other relevant elements.

B. Data understanding

We collected 1,296 job advertisements from various websites related to job opportunities using the keywords *"Social Customer Relationship Management"*, *"Social CRM"*, *"eCRM"*, *"mCRM"* and *"iCRM"*. Data were downloaded in Comma-Separated Values (CSV) file format. To understand the composition of the database, several aspects were evaluated, including the data volume, and missing data, to obtain a precise notion of its dimensionality. Table I displays the fields present in the database, including *id*, *title*, *text*, *humanlanguage*, *Uniform Resource Locator (URL)*(URL), *requirements* and *tasks*. It is noteworthy that of the 1,296 advertisements collected, some did not contain information in the fields *title*, *requirements*, and *tasks*. Based on the field *text*, it was possible to determine that the average text length is 758 words.

Authorized licensed use limited to: UNIVERSIDADE DE SAO PAULO. Downloaded on April 06,2024 at 14:28:07 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COLLECTED DATASET FIELDS AND THEIR INFORMATION.

Fields	Description	Amount
Id	Identifier number	1,296
Title	Job title from the corresponding job advertisement	1,191
Text	Job description body text	1,296
Humanlanguage	Text Language	1,296
PageUrl	Job posting source URL	1,296
Requirements	Information related to requirements	731
Tasks	Information about job duties	977

Incorrect data, such as duplicates and *"no description available"* have been removed. Then, based on the average text length of the job descriptions, those containing more than 1,500 words were removed. It was performed because *outliers* were identified, such as content not related to the jobs description (*e.g.,* companies history, website information *etc*). The final dataset included 1,086 documents.

C. Data preparation

From the data understanding, the advertisement titles and text attributes were selected as the main features to be used in the modeling phase. The preprocessing and manipulation steps of the textual content applied to the data were performed using the conventional text mining pipeline [18], [19]. The textual content went through the following preprocessing steps: *a) Removal of duplicates:* The PageUrl field was secondarily manipulated and consisted of removing the prefix and suffixes from the URLs, so, removing duplicate ads were based on checking the identifier (id) and page address (PageUrl) repeated to ensure that each ad appeared only once in the dataset; *b) Text standardization:* The text was converted to lowercase letters to avoid discrepancies due to differences in capital letters, non-English language ads have also been removed; *c) Removal of accents, punctuation, irrelevant characters, URLs, and excessive spaces:* These steps were applied to eliminate elements that do not contribute to the analysis; d) Removal of stopwords: The stopwords from the Natural Language Toolkit (NLTK) library were used - this allowed access to a predefined list of English stopwords removed from the text. In addition to using the NLTK set of stopwords, other words were manually added to the stopwords list. Those added words included: *"view"*, *"e g"*, *"please"*, *"send"*, *"end"*, *"de"*, *"compliance"*, *"risk"*, *"risks"*, *"use"*, *"sanctions"*, *"france"*, *"united"*, *"kingdom"*, *"york"*, *"philippinesofficer"*, *"america"*, *"abbott"*, *"nbsp"*, *"ireland"*, *"northern"*, *"belfast"*, *"citi"*, *"chalhoub"*, *"bcg"*, *"per"*, *"hq"*, *"rohq"*, *"fiserv"*, *"aml"*, *"surveillance"*, *"surveillances"*, *"ct"*, *"legal"*, *"regulatory"*, *"regulations"*, *"rules"*, *"law"*, *"hybridbelfast"*, *"citiover"*, *"tts"*, *"grct"*, *"gft"*, *"email"*, *"apply"*, *"career"*; Finally, *e) Removal of numerical characters:* This step was applied only after evaluating the years of experience required to obtain more accurate results for the topic modelling.

D. Modeling

Several data mining techniques were explored to identify patterns in job advertisements. One approach used to analyze continuous sequences of words was the generation of N-grams, which are neighboring sequences of items in a document [20]. Topic modeling was another technique employed to extract knowledge from unstructured texts in job descriptions. This method aims to group documents and words with similar meanings and is widely used in NLP [21]. Among the various topic modeling options available, for example, Latent Dirichlet Allocation, Non-Negative Matrix Factorization (NMF), BERTopic, Word2Vec [22]–[24]. BERTopic was chosen since it considers semantic information. This approach generates document representations through embeddings of language models based on pre-trained transformers, and a class-based TF-IDF to create dense clusters, allowing the creation of easily interpretable topics and highlighting essential words in topic descriptions [24].

Unlike other techniques such as LDA, BERTopic does not require prior knowledge of the number of clusters [24]. However, it can generate a large number of clusters, making semantic interpretation difficult. To deal with this, it is possible to reduce the number of topics after training, aiming to create a realistic and easily annotated amount by human experts, following the human-in-the-loop approach [25]. In addition, BERTopic offers a *reduce_topics* method that uses the original template information to do a topic reduction, regrouping, and combining existing topics to form broader topics [24]. This can be useful to simplify the interpretation of the results, thus generating meaningful insights.

It should be noted that the original model trained by BERTopic contains Topic -1, which always represents outliers and should not be considered in the analyses [24], [26]. Preprocessing is generally unnecessary for BERTopic as it relies on an embedding approach, and keeping the original structure of the text is vital for transformer models [26]. However, in cases of noisy data, as in the context of this work, conducting preprocessing steps can become beneficial. Refining and cleaning the data before applying BERTopic substantially improved the obtained results.

The third approach was related to building the lexicon related to job posts. The development of this lexicon involved a semi-automatic approach, as described in Figure 2. Initially, the axes were defined, considering the existing elements in the advertisement structure, which resulted in job posts, technologies, skills (technical and behavioral), and knowledge. Next, the authors read a representative set of job postings (around 30% of the dataset was carefully examined).

This step was combined with a script for searching terms related to keywords observed during the reading, which are: *"platforms"*, *"platform"*, *"tools"*, *"software"*, *"knowledge"*, *"certification"*, *"plus"*, *"ability"*, *"required"*, *"programs"* e *"experience"*. This process culminated in building a dictionary that facilitated identifying and visualizing the topics addressed in the job posts. Finally, the terms added to the dictionary were

Authorized licensed use limited to: UNIVERSIDADE DE SAO PAULO. Downloaded on April 06,2024 at 14:28:07 UTC from IEEE Xplore. Restrictions apply.

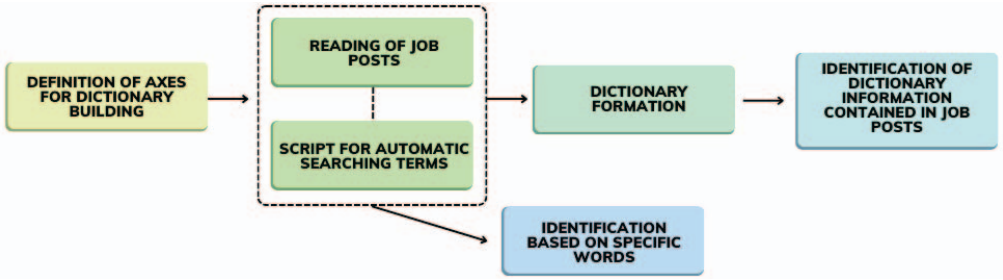


Fig. 2. Stages for building the job posting lexicon.

used to determine the most recurrent ones in the job posts. It is worth mentioning that the evaluation of the results was carried out later, which is described in the following subsection.

E. Evaluation and implementation

The evaluation stage was based on the grounded theory approach, which consists of a qualitative research method that allows researchers to discern explicit and implicit processes in their data [27]. In this study, the approach was used so that the authors could read a representative data sample and cross them with the results obtained. It is also worth mentioning that a sample calculation was made, considering the population's size, a confidence level of 95%, and a margin of error of 5.

The sample size was therefore set at 297 job advertisements, chosen at random, which were read in full as a complementary way of understanding job post characteristics. The authors analyzed the job advertisements to identify domain terms related to Social CRM in order to build the lexicon and increase the study reliability. The dataset with 1,086 job advertisements was used for N-gram analysis, topic modeling, and term counting from the lexicon built.

F. Technologies

The Python (3.9.13) was used to develop the scripts. The experiments were conducted in Jupyter Notebook (6.4.12). In the data preparation and modeling step, several libraries were used to perform different tasks. For example, Pandas (1.4.4) was used for data manipulation; the Natural Language Toolkit (MLTK)(3.7) was used for text preprocessing; Regular expressions were implemented using library (RE) (2022.7.9). The textual representation was performed using the Bag-of-Words method, implemented through the Scikit-learn library (1.0.2). For topic modeling, BerTopic[1] (0.15.0) was used. For data visualization, matplotlib/pyplot (3.5.2) and Seaborn(0.11.2) were adopted for generating graphics and WordCloud[2] (1.8.2.2).

IV. Results

After preprocessing the data, a total set of 1,086 job advertisements was analyzed. Figure 3 visually represents job descriptions in a . This representation provides insight into the skills, knowledge, and requirements referenced in available opportunities. Analyzing these terms highlights the importance of identifying similar words for a deeper understanding of job vacancies. By grouping related terms, it is possible to obtain a comprehensive view of the requirements demanded by employers in Social CRM.

By deepening our analysis of Figure 3, we identified aspects not previously explored in the research questions, highlighting the community's concern with inclusion and equity. This analysis reveals the presence of expressions such as *"equal opportunity"*, *"sexual orientation"*, *"gender identity"*, *"race color"* and *"religion"*. Such terms indicate the value employers give to diversity and guarantee equal opportunities for all candidates. Including these terms call attention to the importance of understanding the technical skills required for positions and the relevance of policies and values that guide the work environment. Furthermore, incorporating the words *"support"* and *"employer qualified"* suggests that companies are focused on creating a professional background in which all employees feel supported and recognized.

Analyzing job advertisements made it possible to examine the distribution of years of experience requested by companies, as shown in Figure 4. Figure 4 depicts an overview of minimum years of experience. Notably, it can be seen that the years of experience distribution vary. Most postings, corresponding to about 50, require candidates with more than 10 years of experience, indicating the search for highly specialized professionals with a broad professional background. The analysis also reveals a substantial group of jobs between 5 and 6 years old, meaning that many opportunities are destined for professionals with intermediate experience. It is important to note that the number of people with 0 to 2 years of experience is relatively low.

To better understand the skills in Social CRM. Initially, an N-gram analysis was carried out, in which the main opportunities and requirements were identified. Subsequently, these skill sets were mapped based on topic modeling. In addition, the most sought-after tools in Marketing, Design, Data Analysis, Software Development, and Programming Languages were analyzed. The results are presented and discussed below.

The preliminary results related to the N-grams in the titles are presented in Figure 5, in which it is possible to observe the 15 most frequent N-grams. The terms indicated the ever-

[1]https://maartengr.github.io/BERTopic/index.html
[2]https://pypi.org/project/wordcloud/

Authorized licensed use limited to: UNIVERSIDADE DE SAO PAULO. Downloaded on April 06,2024 at 14:28:07 UTC from IEEE Xplore. Restrictions apply.

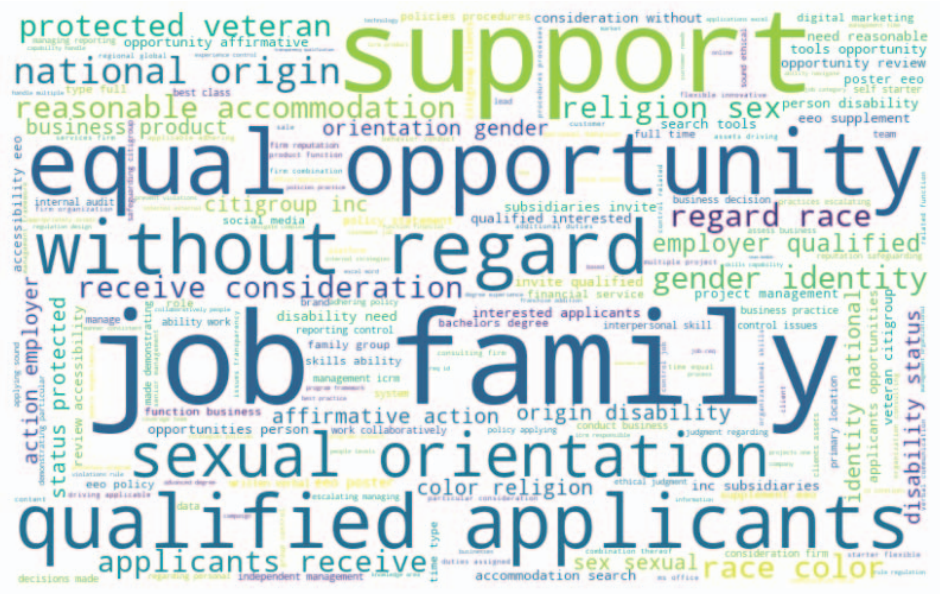


Fig. 3. Word cloud representation of job posts descriptions.

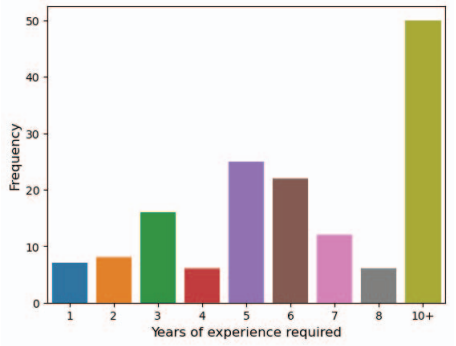


Fig. 4. Years of experience required.

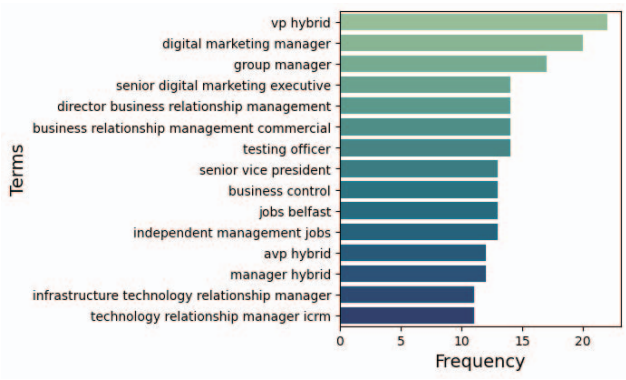


Fig. 5. Representation of N-grams constructed based on job titles.

evolving opportunities and demands related to Social CRM, reflecting the importance of technical skills, advanced research, and interdisciplinary collaboration. Through this analysis, it was possible to verify which titles appear in the job posts, fundamental to understanding which positions are being demanded.

Analyzing Figure 5, it is possible to identify some terms that are directly related to Social CRM, such as *"digital marketing manager", "business relationship management commercial", "senior digital marketing executive", and "director business relationship management"*. Based on a cross-analysis with Figure 4, there is a notable congruence since the high presence of senior and managerial positions corroborates the job vacancies that require professionals with extensive experience since these positions require a broader business view. In addition, these terms have keywords associated with customer relationship management and digital marketing, which are Social CRM essential elements.

On the other hand, some N-grams may not be directly related to Social CRM, such as *"jobs belfast"*, *"independent management jobs"* and *"business control"*. These terms can be considered noise and should be filtered out to ensure the search is more accurate and focused on relevant topics.

The BERTopic-based topic modeling analysis was performed on the job advertisements using different parameters, and, as expected, the initial model identified a large number of topics (31 topics), some of which were redundant or too specific. In this way, reducing topics can lead to a more concise and representative set, facilitating the analysis and understanding of the themes present in the documents. These topics are presented in Table II, along with the BERTopic descriptive

Authorized licensed use limited to: UNIVERSIDADE DE SAO PAULO. Downloaded on April 06,2024 at 14:28:07 UTC from IEEE Xplore. Restrictions apply.

keywords. Topics are listed in the table in descending order of frequency. Discovered topic names were assigned manually, considering descriptive keywords and their frequencies.

In Table II, the first word was the most seen, and the last was the least seen in a topic. In this context, topic names were commonly assigned, taking into account the first four keywords. On the other hand, for some topics, the keywords together defined a specific domain, as detailed below:

Topic 1: Business Management and Control - This topic indicates that companies may be looking for professionals with data management and analysis skills. The social CRM area might be related to managing customer data, analyzing marketing metrics, and implementing efficient processes to improve customer relationships. In addition, companies seek to comply with Equal Employment Opportunity (EEO) standards.

Topic 2: Social CRM and Marketing - This topic is directly relevant to Social CRM. The job posts indicate that companies are looking for professionals who can lead and execute digital marketing strategies and manage customer relationships through social media platforms and CRM systems. This includes campaign management skills, customer data analysis, and utilizing social media to drive customer engagement and loyalty.

Topic 3: Career Analyst and Diversity Officer - This topic addresses career management and promoting diversity in organizations. While not directly tied to Social CRM, there may be some linkage implicitly, as a diverse Social CRM team can better understand and address specific customers' needs. In addition, career analysis can help companies identify suitable talent to work with Social CRM and improve the teams' efficiency.

Topic 4: Business and Portfolio Strategy - This topic addresses business strategies and portfolio management issues, focusing on business expansion and growth. Although business strategy can influence lead capturing and collaborative development of products/services, this topic is not directly related to the specific practices and Social CRM tools.

Topic 5: Growth Ventures and Digital Marketing - This topic is related to the usage of digital marketing strategies to drive business growth and expansion, as well as looking for business opportunities that can benefit from using Social CRM to improve relationships with customers, and boost sales.

Topic 6: Partnership and Microsoft Campaign - This topic suggests the importance of strategic partnerships and campaigns with companies like Microsoft. Under Social CRM, this can involve integrating CRM solutions with Microsoft platforms/tools to improve efficiency, share data, and enhance the customer experience.

In general, topics 1, 2, 5, and 6 seem more in line with the field of Social CRM, as they address aspects related to business management, digital marketing, business growth, strategic partnerships, and campaigns. They can directly influence the Social CRM implementation. While topics 3 and 4 are relevant in other business areas, they do not directly relate to Social CRM.

Through N-gram analysis and topic modeling, it was possible to glimpse the diversity and nuances of professional opportunities in this field. However, the above reviews provide an overview of requirements and areas of focus, ranging from business management and digital marketing to data analysis. In this way, a lexicon was built to categorize and organize this complex information to obtain a more precise and structured view of the employment scenario in Social CRM.

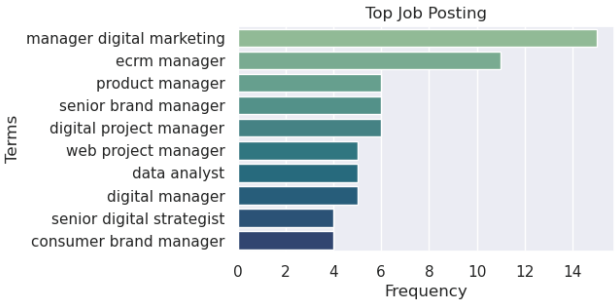


Fig. 6. Job Posting

Figure 6 shows the term frequency in the lexicon built from job posts related to Social CRM. The predominance of positions such as *"digital marketing manager"*, *"ecrm manager"*, and *"product manager"* suggests a growing need for professionals capable of leading digital and CRM strategies. This indicates a strong emphasis on digital marketing management and utilizing customer relationship systems to drive business results. The priority given to *"senior brand manager"*, *"consumer brand manager"*, *"digital project manager"* and *"web project manager"* reinforces CRM strategies relevance to build and maintain solid brands in the digital environment. The growing importance of data analysis is represented by the *"Data Analyst"*, suggesting the need for professionals capable of getting valuable insights from customer data.

Figure 7 reveals insights into key technologies present in job posts. Design tools such as *"photoshop"* and *"InDesign"* are prevalent, evidencing the market need for professionals capable of creating attractive and engaging visual content. Technologies related to data analysis, such as *"google analytics"*, *"tableau"*, and *"power bi"* indicate an increasing focus on data-driven decision-making.

Regarding the skills mentioned in the job posts through Figure 8, we can observe that the frequency of data analysis and programming skills calls attention to the need for professionals who can interpret valuable insights from the collected data and implement solutions techniques. The emphasis on *"communication"*, *"interpersonal skills"*, and *"flexible"* reflects the significance of interpersonal skills in a scenario where collaboration and adaptation are crucial to success. The prevalence of *"proactive"* suggests that this ability is highly valued in dealing with the ever-evolving complexities of the digital environment.

Figure 9 shows an overview of the main knowledge required by companies. The inclusion of *"digital marketing"*,

Authorized licensed use limited to: UNIVERSIDADE DE SAO PAULO. Downloaded on April 06,2024 at 14:28:07 UTC from IEEE Xplore. Restrictions apply.

TABLE II
REPRESENTATION OF SELECTED TOPICS WITH TERMS.

Topic	Labels	Key words
1	Business Management and Control	business, management, control, experience, data, related, internal, eeo, work, job
2	Social CRM and Marketing	marketing, digital, customer, crm, social, sales, media, brand, campaigns, management
3	Career Analyst and Diversity Officer	learn, analyst, career, diversity, jobs, officer, hybridmultiple, search, cookie, preferences
4	Business and Portfolio Strategy	business, commercial, portfolio, nutrition, strategy, management, global, divisional, brm, life
5	Growth Ventures and Digital Marketing	growth, ventures, architects, execute, customer, validate, digital, worlds, adoption, marketing
6	Partnership and Microsoft Campaign	marketing, partnership, drive, campaign, microsoft, windows, partners, consumer, team, across

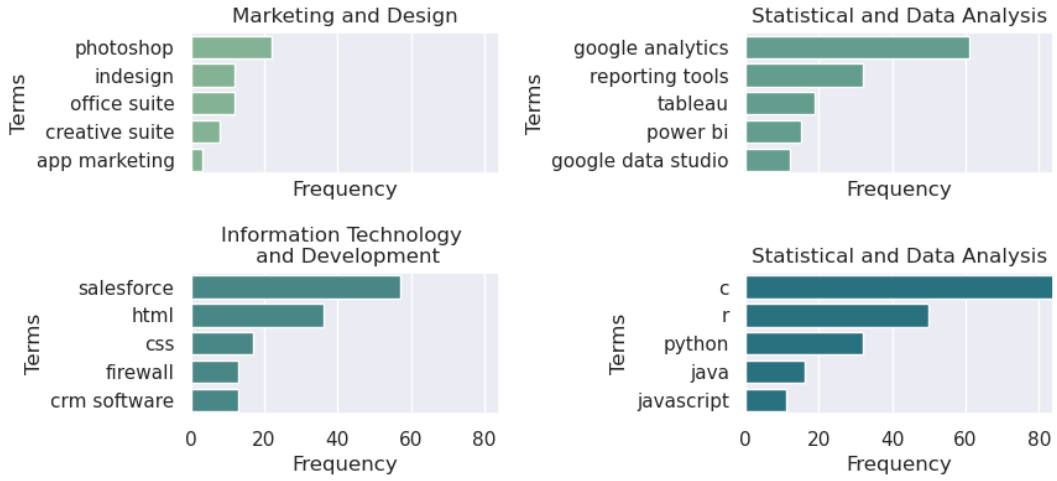


Fig. 7. Top Technologies

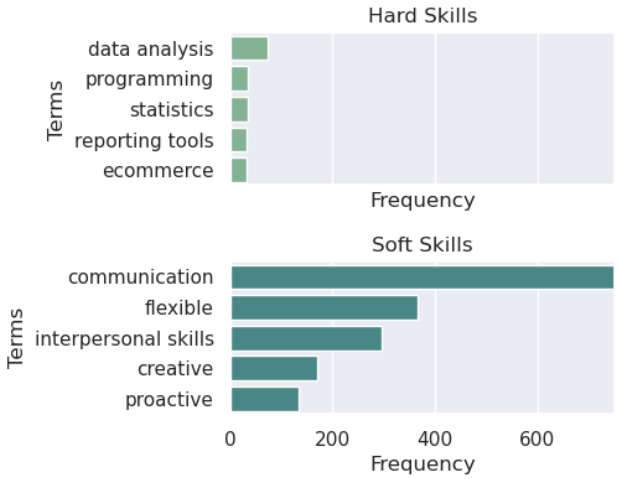


Fig. 8. Top Skills

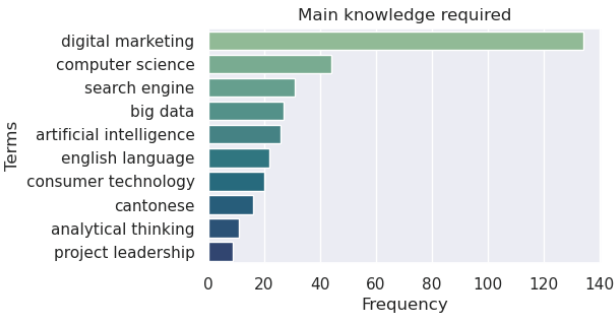


Fig. 9. Main knowledge required

"computer science", and *"big data"* as frequent terms show up the need for professionals with a solid understanding of modern digital marketing concepts/technologies. The emphasis on *"search engine"* and *"artificial intelligence"* suggests that optimizing for search engines and adopting artificial intelligence solutions are considered vital skills. The presence of the *"English language"* as essential knowledge reflects the globalization of the Social CRM field, thus, the importance of communication in an international context.

The results obtained from the lexicon indicate that Social CRM is rapidly evolving towards more sophisticated digital strategies and data-driven analytics. Professionals with a balance of technical and soft skills, whith a deep understanding of digital and technology trends, have a competitive advantage.

Authorized licensed use limited to: UNIVERSIDADE DE SAO PAULO. Downloaded on April 06,2024 at 14:28:07 UTC from IEEE Xplore. Restrictions apply.

V. Final Remarks

The research novelty relies on a comprehensive approach to analyzing job advertisements related to Social CRM by combining text mining and human-in-the-loop methods/strategies. To carry out the study, 1,086 job posts were collected using Diffbot, which facilitates data scraping. The analysis focused on identifying and distinguishing skills, knowledge, positions, and technologies. Good data science practices were used, which allowed the data analysis centered on the N-gram approach, topic modeling, and the building of a lexicon oriented to the Social CRM domain. In addition, the concepts related to grounded theory are highlighted to validate the results. These steps were crucial to understanding the Social CRM market's dynamics, patterns, and trends.

In short, the results converge to identify a growing demand for highly specialized professionals in Social CRM, who would be capable of leading digital and CRM strategies. However, we also identified non-relevant N-grams, suggesting the need to filter out noise to target the search more precisely. Topic modeling also suggests a growing demand for professionals with skills in digital strategies, data analysis, and customer relationship management.

This information has the potential to guide people looking to enter this field. It also makes promoting public policies aimed at this job market easier. Text mining and data analysis have become powerful tools for understanding the job market dynamics and making important decisions about hiring and qualifying professionals in this ever-changing field. In future work, we intend to expand our data collection, including other search terms. We also would like to refine our lexicon by inviting Human Resources specialists to validate the dictionary. Finally, we intend to apply more text mining methods, such as k-means for grouping the data, and large language models (e.g., GPT) for testing a question-and-answering approach.

Acknowledgements

This work was supported by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) - DT-308334/2020; and by *Fundação Amazônia de Amparo a Estudos e Pesquisas* (FAPESPA) PRONEM-FAPESPA/CNPq nº 045/202.

References

[1] M. M. Mariani, R. Perez-Vega, and J. Wirtz, "Ai in marketing, consumer research and psychology: A systematic literature review and research agenda," *Psychology & Marketing*, vol. 39, no. 4, pp. 755–776, 2022.

[2] F. Lobato, M. Pinheiro, A. Jacob, O. Reinhold, and Á. Santana, "Social crm: Biggest challenges to make it work in the real world," in *Business Information Systems Workshops: BIS 2016 International Workshops, Leipzig, Germany, July 6-8, 2016, Revised Papers 19*. Springer, 2017.

[3] R. Alt and O. Reinhold, *Social customer relationship management: Fundamentals, applications, technologies*. Springer Nature, 2019.

[4] W. Silva, Á. Santana, F. Lobato, and M. Pinheiro, "A methodology for community detection in twitter," in *Proceedings of the International Conference on Web Intelligence*, 2017, pp. 1006–1009.

[5] B. N. R. Chagas, J. A. N. Viana, O. Reinhold, F. Lobato, A. F. Jacob, and R. Alt, "Current applications of machine learning techniques in crm: a literature review and practical implications," in *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 2018.

[6] A. B. Holm, "The effect of e-recruitment on the recruitment process: Evidence from case studies of three danish mncs," in *Proceedings of the 3rd European academic workshop on electronic human resource management*, 2010, pp. 91–111.

[7] R. Boselli, M. Cesarini, S. Marrara, F. Mercorio, M. Mezzanzanica, G. Pasi, and M. Viviani, "Wolmis: A labor market intelligence system for classifying web job vacancies," *Journal of intelligent information systems*, vol. 51, pp. 477–502, 2018.

[8] K. Ghazzawi and A. Accoumeh, "Critical success factors of the e-recruitment system," *Journal of Human Resources Management and Labor Studies*, vol. 2, no. 2, pp. 159–170, 2014.

[9] L. M. Kureková, M. Beblavỳ, and A. Thum-Thysen, "Using online vacancies and web surveys to analyse the labour market: A methodological inquiry," *IZA Journal of Labor Economics*, vol. 4, pp. 1–20, 2015.

[10] F. Gurcan and N. E. Cagiltay, "Big data software engineering: Analysis of knowledge domains and skill sets using lda-based topic modeling," *IEEE access*, vol. 7, pp. 82 541–82 552, 2019.

[11] D. Debao, M. Yinxia, and Z. Min, "Analysis of big data job requirements based on k-means text clustering in china," *PloS one*, 2021.

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

[13] A. Ternikov, "Soft and hard skills identification: insights from it job advertisements in the cis region," *PeerJ Computer Science*, vol. 8, 2022.

[14] S. Moro, R. Laureano, and P. Cortez, "Using data mining for bank direct marketing: An application of the crisp-dm methodology," in *Proceedings of European Simulation and Modelling Conference - ESM'2011*, 2011.

[15] R. Wirth and J. Hipp, "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*, vol. 1. Manchester, 2000, pp. 29–39.

[16] K. Jensen, "A diagram showing the relationship between the different phases of crisp-dm and illustrates the recursive nature of a data mining project," *Wikimedia Commons*, vol. 26, 2012.

[17] H. Nigam and P. Biswas, "Web scraping: from tools to related legislation and implementation using python," in *Innovative Data Communication Technologies and Application: ICIDCA 2020*. Springer, 2021.

[18] D. Cirqueira, M. F. Pinheiro, A. Jacob, F. Lobato, and Á. Santana, "A literature review in preprocessing for sentiment analysis for brazilian portuguese social media," in *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 2018, pp. 746–749.

[19] L. C. Fernandes, J. Silva, A. Jacob, and F. Lobato, "An extensive analysis of online restaurant reviews: a case study of the amazonian culinary tourism," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2020, pp. 81–84.

[20] A. Balfagih, V. Keselj, and S. Taylor, "N-gram and word2vec feature engineering approaches for spam recognition on some influential twitter topics in saudi arabia," in *Proceedings of the 6th International Conference on Information System and Data Mining*, 2022, pp. 101–107.

[21] J. Garcia, G. Villavicencio, F. Altimiras, B. Crawford, R. Soto, V. Minatogawa, M. Franco, D. Martínez-Muñoz, and V. Yepes, "Machine learning techniques applied to construction: A hybrid bibliometric analysis of advances and future directions," *Automation in Construction*, vol. 142, p. 104532, 2022.

[22] D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler *et al.*, "Applying lda topic modeling in communication research: Toward a valid and reliable methodology," in *Computational methods for communication science*. Routledge, 2021.

[23] Y. Chen, H. Zhang, R. Liu, Z. Ye, and J. Lin, "Experimental explorations on short text topic mining between lda and nmf based schemes," *Knowledge-Based Systems*, vol. 163, pp. 1–13, 2019.

[24] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[25] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," *Future Generation Computer Systems*, vol. 135, pp. 364–381, 2022.

[26] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Frontiers in sociology*, vol. 7, p. 886498, 2022.

[27] K. Charmaz and R. Thornberg, "The pursuit of quality in grounded theory," *Qualitative research in psychology*, vol. 18, no. 3, 2021.

Authorized licensed use limited to: UNIVERSIDADE DE SAO PAULO. Downloaded on April 06,2024 at 14:28:07 UTC from IEEE Xplore. Restrictions apply.

APÊNDICE C – ARTIGO PUBLICADO NO *ENIAC* 2023

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

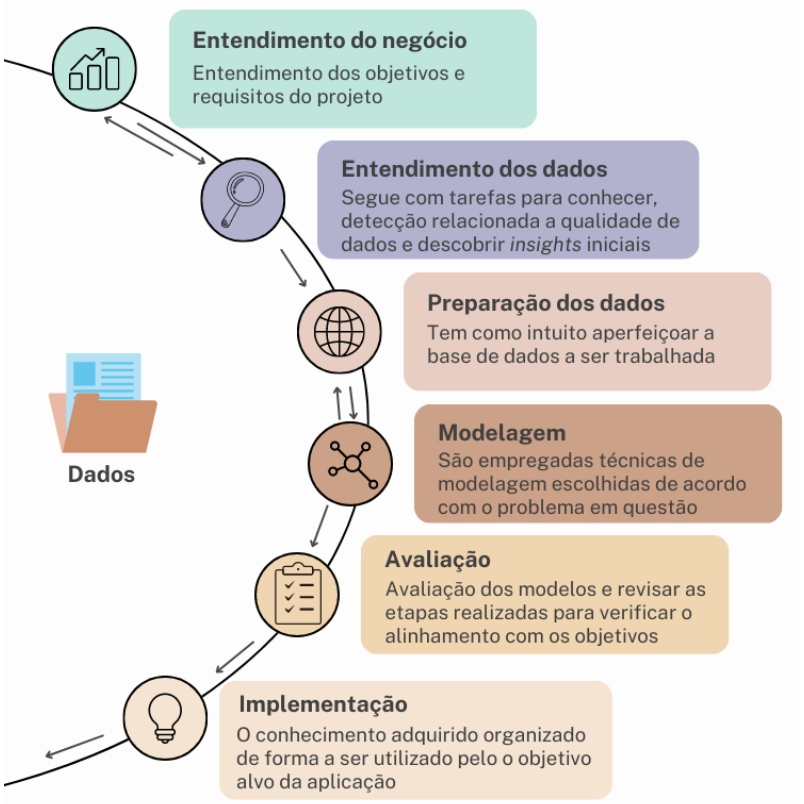


Figura 1. Fases do método CRISP-DM.

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-

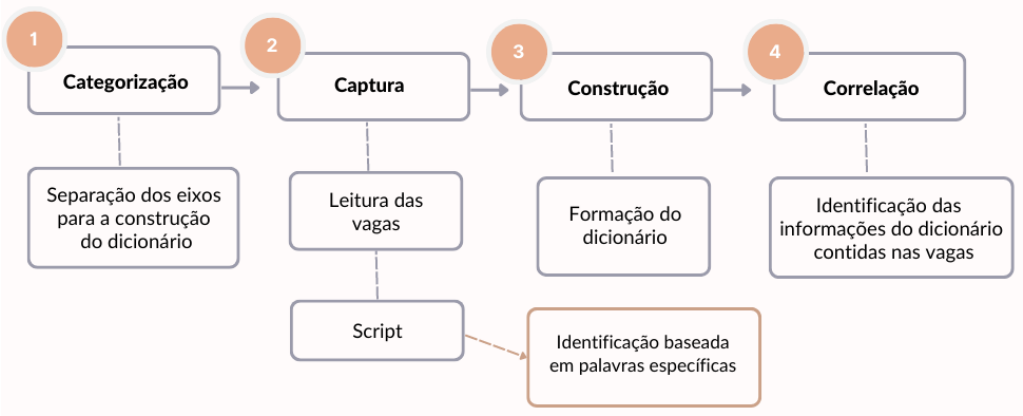


Figura 2. Etapas da construção do léxico.

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.

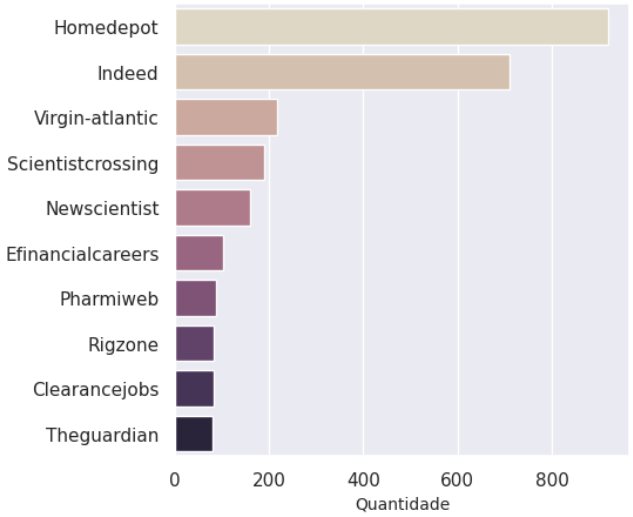


Figura 4. Plataformas com a maior quantidade de anúncios de vagas de ciência de dados.

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.

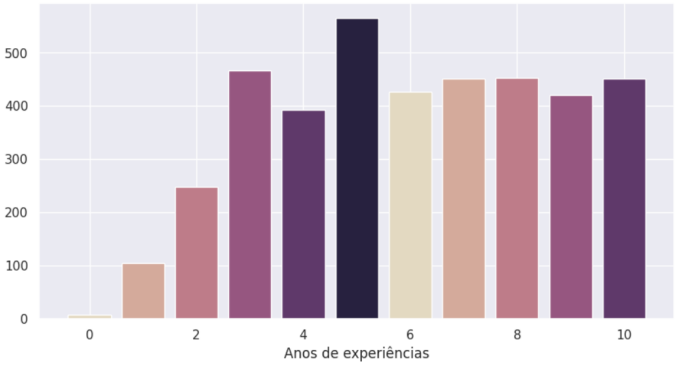


Figura 5. Ano mínimo de experiência indicado nas vagas.

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

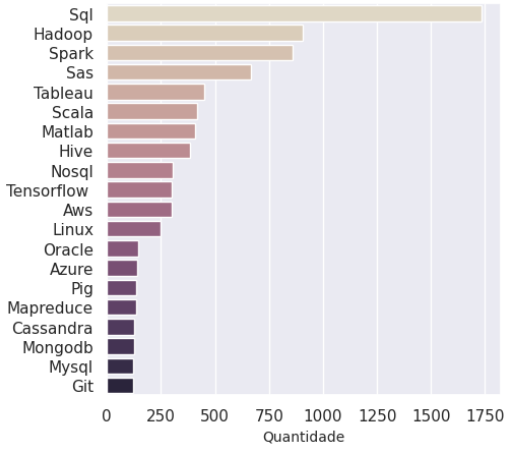


Figura 6. As tecnologias mais frequentes encontradas nas descrições das vagas.

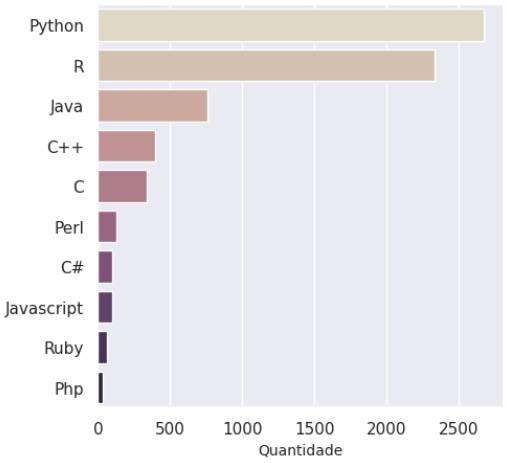


Figura 7. Linguagens de programação mais mencionadas nos anúncios.

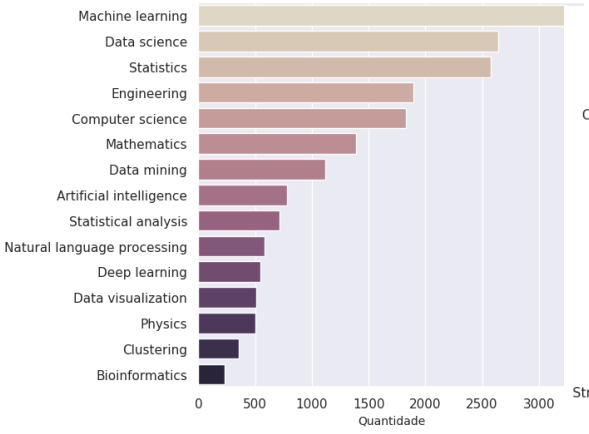


Figura 8. Conhecimentos mais recorrentes nas descrições das vagas.

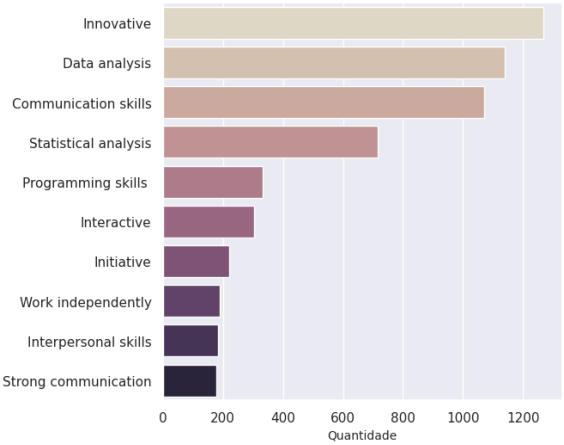


Figura 9. Habilidades mais recorrentes.

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

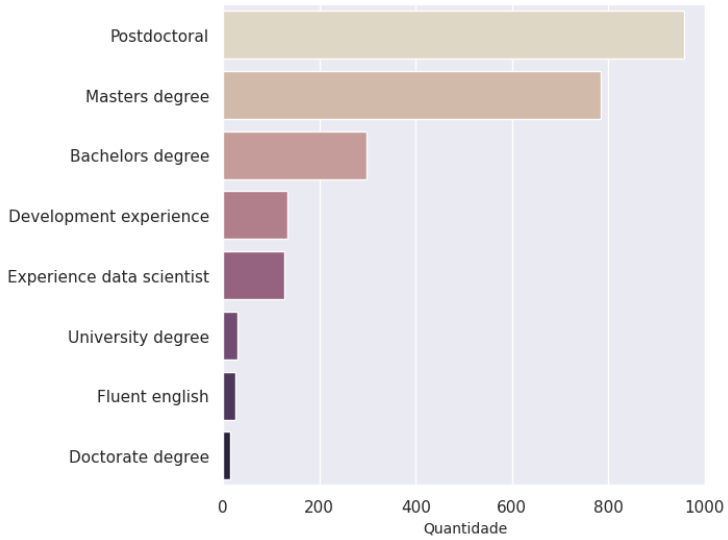


Figura 10. Indicação dos requisitos predominantes nas descrições.

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.

APÊNDICE D – INFOGRÁFICO ANUAL *DATA NEVER SLEEPS 11.0* - IMAGEM ORIGINAL

Figura 24 – Infográfico anual *Data Never Sleeps 11.0*

Fonte: (Domo, 2023)

APÊNDICE E – MANEIRAS PELAS QUAIS AS HABILIDADES INTERPESSOAIS PODEM CONTRIBUIR PARA O SUCESSO DO NEGÓCIO - IMAGEM ORIGINAL

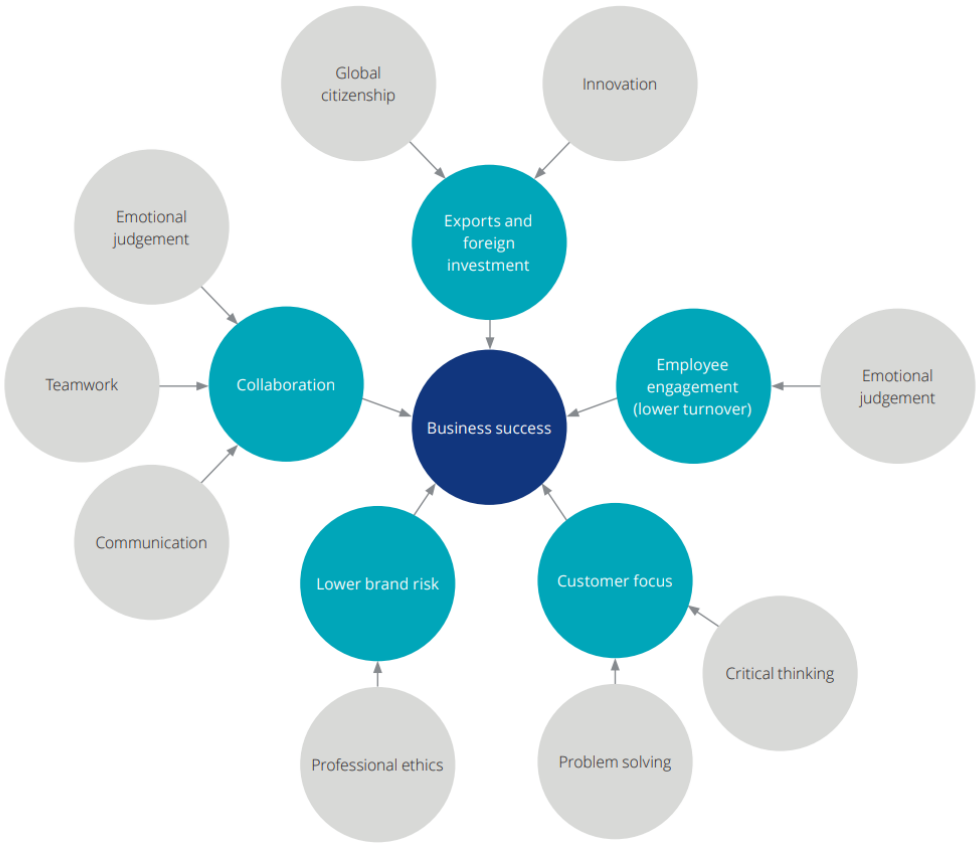


Figura 25 – Maneiras pelas quais as habilidades interpessoais podem contribuir para o sucesso do negócio.

Fonte: (ECONOMICS, 2017)