UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
IEG-INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

FERNANDO ALMEIDA DO CARMO

VISÃO GERAL DO MERCADO EM CRM: UM ESTUDO DE CASO DE VAGAS DE EMPREGO

SANTARÉM-PA
2024

FERNANDO ALMEIDA DO CARMO

VISÃO GERAL DO MERCADO EM CRM: UM ESTUDO DE CASO DE VAGAS DE EMPREGO

Trabalho de Conclusão de Curso apresentado ao Programa de Computação, para obtenção do grau de Bacharel em Ciência da Computação; Universidade Federal do Oeste do Pará, Instituto de Engenharia e Geociências.

Orientador: Fábio Manoel França Lobato

SANTARÉM-PA

2024

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema Integrado de Bibliotecas – SIBI/UFOPA

C287v Carmo, Fernando Almeida do

Visão geral do mercado em CRM: um estudo de caso de vagas de emprego./ Fernando Almeida do Carmo. - Santarém, 2024.

20 p. : il.

Inclui bibliografias.

Orientador: Fábio Manoel França Lobato.

Trabalho de Conclusão de Curso (Graduação) – Universidade Federal do Oeste do Pará, Instituto de Engenharia e Geociências, Bacharelado em Ciência da Computação.

1. CRM. 2. Mineração de texto. 3. Análise de dados. 4. Vagas de emprego. 5. Análise de Mercado. I. Lobato, Fábio Manoel França, *orient*. II. Título.

CDD: 23 ed. 005.7

Bibliotecária - Documentalista: Cátia Alvarez – CRB/2 843

UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS
PROGRAMA DE COMPUTAÇÃO

Para gerar os formulários de avaliação basta:

(i) Preencher todos os dados na tabela abaixo;

(ii) Selecionar todo o texto do documento (atalho: Ctrl+T);

(iii) Clique com o lado direito em qualquer parte do texto e clique na opção "Atualizar Campo" (atalho: F9).

Título	*CRM Market Overview: A Case Study of Job Vacancies*
Discente	Fernando Almeida do Carmo
Examinador 1/Orientador	Fábio Manoel França Lobato
Examinador 2	Deam James Azevedo da Silva
Examinador 3	Paula Myrian Lima Pedroso
Data da Apresentação	26 de setembro de 2024

FORMULÁRIO DE AVALIAÇÃO DE TCC

Identificação:

Título do Trabalho: *CRM Market Overview: A Case Study of Job Vacancies*
Aluno (a): Fernando Almeida do Carmo
Orientador (a): Fábio Manoel França Lobato

Avaliação:

Examinador (a) 1: Fábio Manoel França Lobato	**Nota:** 10,0
Assinatura:	

Examinador (a) 2: Deam James Azevedo da Silva	**Nota:** 10,0
Assinatura:	

Examinador (a) 3: Paula Myrian Lima Pedroso	**Nota:** 10,0
Assinatura:	

Parecer:

O trabalho deve ser adequado para depósito na biblioteca, devendo ser incluído os elementos pré e pós-textuais.

Resumo da Avaliação:

	Aceitação incondicional
X	Aceitação condicionada a modificações (especificar no verso)
	Recusado

Nota Final:

Santarém-PA, 26 de setembro de 2024.

gov.br
Documento assinado digitalmente
FABIO MANOEL FRANCA LOBATO
Data: 27/09/2024 10:44:25-0300
Verifique em https://validar.iti.gov.br

_____

Presidente da Banca Examinadora

FORMULÁRIO DE AVALIAÇÃO INDIVIDUAL DE TCC

Identificação:

Título do Trabalho:	*CRM Market Overview: A Case Study of Job Vacancies*
Nome da Aluno (a): Fernando Almeida do Carmo	
Orientador (a): Fábio Manoel França Lobato	
Avaliador (a): Deam James Azevedo da Silva	

Critérios de Avaliação:

REDAÇÃO	0,0 0,5 1,0
Clareza, concisão e precisão do texto	☐☐☐☐☐☐☐☐☐☐
Organização do trabalho	☐☐☐☐☐☐☐☐☐☐
Adequação e qualidade do embasamento teórico	☐☐☐☐☐☐☐☐☐☐
Adequação das referências bibliográficas	☐☐☐☐☐☐☐☐☐☐
Adequação e qualidade do modelo proposto	☐☐☐☐☐☐☐☐☐☐
Adequação da metodologia e avaliações	☐☐☐☐☐☐☐☐☐☐
Adequação e qualidade da contribuição	☐☐☐☐☐☐☐☐☐☐
APRESENTAÇÃO	
Respeito ao tempo e qualidade do material	☐☐☐☐☐☐☐☐☐X
Clareza na exposição das ideias	☐☐☐☐☐☐☐☐☐X
Domínio do tema	☐☐☐☐☐☐☐☐☐X

Observações:

1. Quesitos de avaliação que puderem ser corrigidos ou melhorados não deverão ter conceito inferior a seis;

2. As sugestões para correções e melhoramentos deverão ser descritas de forma clara, concisa e precisa, no próprio trabalho do aluno, e com resumo no verso desta folha;

3. Conforme o Art. 23, § 4º, no caso do produto de TCC ser um artigo publicado, será avaliada apenas a apresentação oral

Notas para a redação no caso de trabalhos publicados	
Qualis	**Notas**
A1	10,0
A2	9,5
B1	9,0
B2	8,5
B3	8,0
B4	7,5
B5	7,0

Santarém-PA, 26 de setembro de 2024.

gov.br Documento assinado digitalmente
DEAM JAMES AZEVEDO DA SILVA
Data: 27/09/2024 14:11:36-0300
Verifique em https://validar.iti.gov.br

_____

Assinatura do Avaliador

FORMULÁRIO DE AVALIAÇÃO INDIVIDUAL DE TCC

Identificação:

Título do Trabalho:	*CRM Market Overview: A Case Study of Job Vacancies*
Nome da Aluno (a):	Fernando Almeida do Carmo
Orientador (a):	Fábio Manoel França Lobato
Avaliador (a):	Paula Myrian Lima Pedroso

Critérios de Avaliação:

REDAÇÃO	0,0 0,5 1,0
Clareza, concisão e precisão do texto	☐☐☐☐☐☐☐☐☐☐
Organização do trabalho	☐☐☐☐☐☐☐☐☐☐
Adequação e qualidade do embasamento teórico	☐☐☐☐☐☐☐☐☐☐
Adequação das referências bibliográficas	☐☐☐☐☐☐☐☐☐☐
Adequação e qualidade do modelo proposto	☐☐☐☐☐☐☐☐☐☐
Adequação da metodologia e avaliações	☐☐☐☐☐☐☐☐☐☐
Adequação e qualidade da contribuição	☐☐☐☐☐☐☐☐☐☐
APRESENTAÇÃO	
Respeito ao tempo e qualidade do material	☐☐☐☐☐☐☐☐☐X
Clareza na exposição das ideias	☐☐☐☐☐☐☐☐☐X
Domínio do tema	☐☐☐☐☐☐☐☐☐X

Observações:

1. Quesitos de avaliação que puderem ser corrigidos ou melhorados não deverão ter conceito inferior a seis;

2. As sugestões para correções e melhoramentos deverão ser descritas de forma clara, concisa e precisa, no próprio trabalho do aluno, e com resumo no verso desta folha;

3. Conforme o Art. 23, § 4º, no caso do produto de TCC ser um artigo publicado, será avaliada apenas a apresentação oral

Notas para a redação no caso de trabalhos publicados	
Qualis	**Notas**
A1	10,0
A2	9,5
B1	9,0
B2	8,5
B3	8,0
B4	7,5
B5	7,0

Santarém-PA, 26 de setembro de 2024.

Documento assinado digitalmente
gov.br
PAULA MYRIAN LIMA PEDROSO
Data: 27/09/2024 11:20:50-0300
Verifique em https://validar.iti.gov.br

_____

Assinatura do Avaliador

FORMULÁRIO DE AVALIAÇÃO INDIVIDUAL DE TCC

Identificação:

Título do Trabalho:	*CRM Market Overview: A Case Study of Job Vacancies*
Nome da Aluno (a):	Fernando Almeida do Carmo
Orientador (a):	Fábio Manoel França Lobato
Avaliador (a):	Fábio Manoel França Lobato

Critérios de Avaliação:

REDAÇÃO	0,0 0,5 1,0
Clareza, concisão e precisão do texto	☐☐☐☐☐☐☐☐☐☐☐
Organização do trabalho	☐☐☐☐☐☐☐☐☐☐☐
Adequação e qualidade do embasamento teórico	☐☐☐☐☐☐☐☐☐☐☐
Adequação das referências bibliográficas	☐☐☐☐☐☐☐☐☐☐☐
Adequação e qualidade do modelo proposto	☐☐☐☐☐☐☐☐☐☐☐
Adequação da metodologia e avaliações	☐☐☐☐☐☐☐☐☐☐☐
Adequação e qualidade da contribuição	☐☐☐☐☐☐☐☐☐☐☐
APRESENTAÇÃO	
Respeito ao tempo e qualidade do material	☐☐☐☐☐☐☐☐☐☐X
Clareza na exposição das ideias	☐☐☐☐☐☐☐☐☐☐X
Domínio do tema	☐☐☐☐☐☐☐☐☐☐X

Observações:

1. Quesitos de avaliação que puderem ser corrigidos ou melhorados não deverão ter conceito inferior a seis;

2. As sugestões para correções e melhoramentos deverão ser descritas de forma clara, concisa e precisa, no próprio trabalho do aluno, e com resumo no verso desta folha;

3. Conforme o Art. 23, § 4º, no caso do produto de TCC ser um artigo publicado, será avaliada apenas a apresentação oral

Notas para a redação no caso de trabalhos publicados	
Qualis	**Notas**
A1	10,0
A2	9,5
B1	9,0
B2	8,5
B3	8,0
B4	7,5
B5	7,0

Santarém-PA, 26 de setembro de 2024.

_____

Assinatura do Avaliador

RESUMO

As mídias sociais revolucionaram a forma como as empresas se relacionam com seus clientes, impactando diretamente os sistemas de Gestão de Relacionamento com o Cliente (CRM). Além disso, essas mídias mudaram drasticamente a forma como as empresas anunciam vagas de emprego e recrutam funcionários. Considerando a multitude de plataformas de recrutamento e o aumento do volume de dados, a extração de conhecimento para acompanhar a evolução mercadológica representa um desafio de pesquisa importante. Neste trabalho, apresentou-se uma análise textual de anúncios de emprego relacionados a CRM usando métodos de mineração de texto e uma construção lexical. Identificamos as vagas mais comuns, bem como conhecimentos, habilidades técnicas (hard skills) e habilidades interpessoais (soft skills). Este trabalho foi concebido sob a perspectiva da Teoria do Design, principalmente no que diz respeito à construção de artefatos que permitam a análise textual. As análises foram realizadas usando uma metodologia baseada na *Design Science Research*, onde o problema foi identificado, os objetivos foram delineados e passou-se para a etapa de design, onde aplicamos técnicas de mineração de texto incluindo análise de N-gramas de descrições de cargos, modelagem de tópicos e construção de léxico orientada para a área de CRM. Os achados da pesquisa revelam as principais áreas de conhecimento, requisitos de experiência, habilidades, treinamentos e tecnologias que se destacam neste setor, bem como características como preferência por equipes diversas, preocupação com equidade de gênero, inclusão da comunidade LGBQIAP+ e diversidade de raça/cor. O léxico construído pode ser usado para uma visão mais precisa e estruturada do cenário de emprego de CRM. Essas informações fornecem boas oportunidades para candidatos e recrutadores no processo de contratação, permitindo que ambos identifiquem tais aspectos com mais precisão e apoiem a tomada de decisão.

Palavras-Chave: CRM, Mineração de Texto, Análise de dados, Vagas de Emprego, Análise de Mercado

ABSTRACT

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. 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. 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). This work was conceived from the perspective of Design Theory, mainly concerning the construction of artifacts that allow textual analysis. 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. 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. 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.

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

SUMÁRIO

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**.

SBSI '24, May 20–23, 2024, Juiz de Fora, Brazil

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.

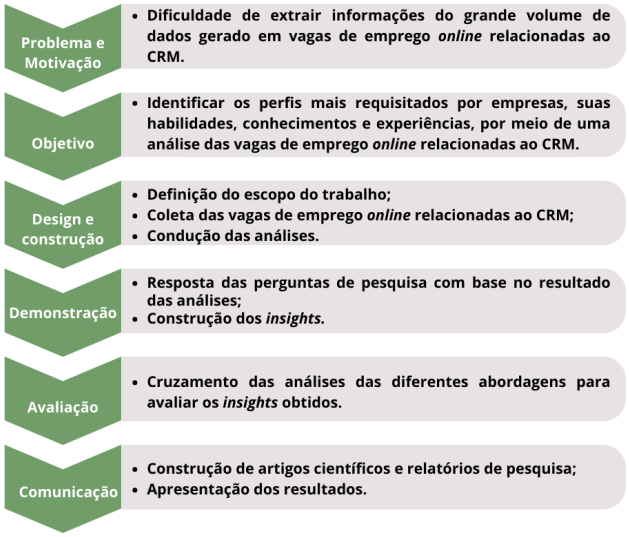

Figura 1: Etapas do *Design Science Research*.

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/

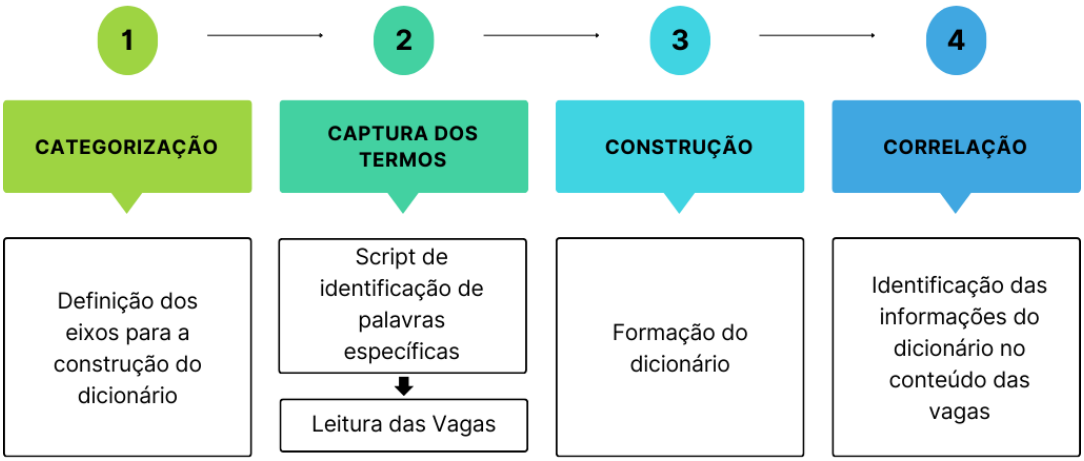

Figura 2: Etapas para a construção do léxico.

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.

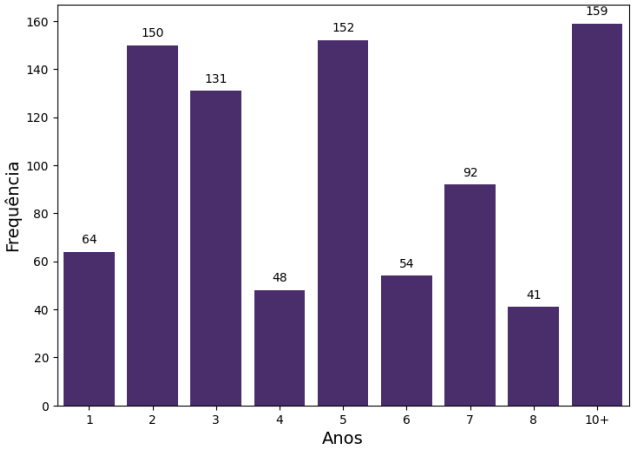

Figura 4: Tempo de experiência exigido nas vagas envolvendo CRM.

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

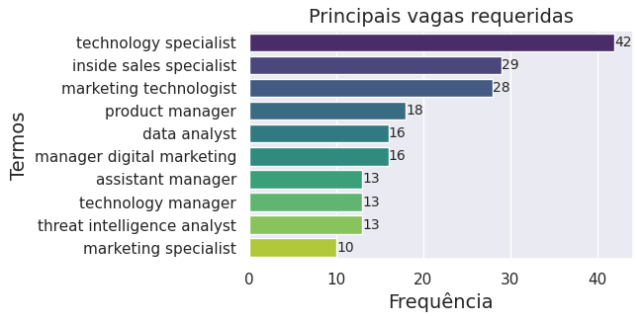

Figura 6: Principais vagas relacionadas ao CRM.

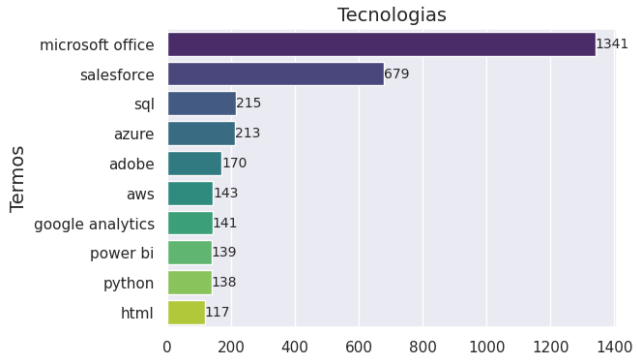

Figura 7: Principais tecnologias requeridas em vagas relacionadas ao CRM.

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;

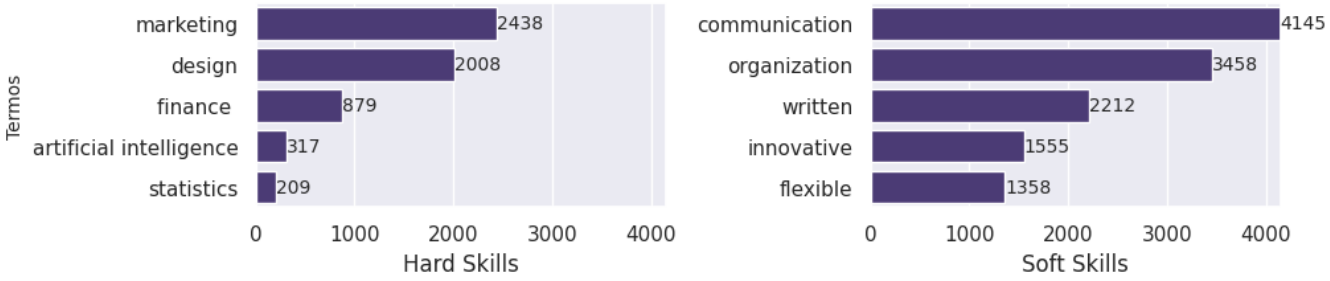

Figura 8: *Hard* e *Soft skills* identificadas nos anúncios de vagas de emprego.

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