Análise Comparativa das Demandas do Mercado de Trabalho em Ciência de Dados em 2024 com a Classificação Brasileira de Ocupações Utilizando Processamento de Linguagem Natural

Gustavo Coxev Wolski
Mackenzie
SP, Brasil
0000-0002-5408-1139

Dirceu Matheus Junior
Mackenzie
SP, Brasil
0000-0001-7823-7284

***Abstract*—This study aimed to compare the competencies demanded by the job market in 2024 in the field of Data Science with the parameters established by the Brazilian Classification of Occupations (CBO). The analysis highlighted the collaborative nature and versatility required to work in various contexts, emphasizing the importance of understanding the transformations in contemporary occupational structures and the need to revise the CBO to reflect these dynamic changes. The essential steps of data preprocessing were addressed, highlighting their relevance in ensuring robust analysis. The study also explored word frequency and similarity analyses between occupation descriptions and job market descriptions. Preliminary results revealed a broader understanding of the professional landscape, contributing to a detailed analysis of the specific demands in the field of Data Science, and highlighting the importance of continuous adaptation of classifications to better reflect the evolving dynamics of the job market.**

***Keywords*—Changes in occupational structures; Natural language processing; Data science and analytics.**

***Resumo*— Este estudo teve como objetivo comparar as competências demandadas pelo mercado de trabalho em 2024 na área de Ciência de Dados com os parâmetros estabelecidos pela Classificação Brasileira de Ocupações (CBO). A análise destacou a natureza colaborativa e a versatilidade exigidas para atuar em diversos contextos, sublinhando a importância de compreender as transformações nas estruturas ocupacionais contemporâneas e a necessidade de revisar a CBO para refletir essas mudanças dinâmicas. Foram abordadas as etapas essenciais de pré-processamento de dados, destacando sua relevância para garantir uma análise robusta. O estudo também explorou a análise de frequência de palavras e de similaridades entre a descrição. Os resultados preliminares revelaram uma compreensão mais abrangente do panorama profissional, contribuindo para uma análise detalhada das demandas específicas no campo de Ciência de Dados e evidenciando a importância da adaptação contínua das classificações para melhor refletir a dinâmica evolutiva do mercado de trabalho.**

***Palavras-chave*—Ciência e análise de dados; Processamento de linguagem Natural; Mudanças nas estruturas ocupacionais.**

I. Introdução

Com a situação atual dos trabalhadores em pauta, é crucial entender como as mudanças nas estruturas ocupacionais podem afetar diretamente suas vidas e carreiras, além de fornecer um direcionamento que reflete a atualidade do mercado. Em resposta a essa necessidade, o Ministério do Trabalho e Emprego (MTE) tomou uma medida significativa em 2002. Nesse ano, foi disponibilizada à população uma reformulação da CBO, com o objetivo de substituir a versão anterior, datada de 1994. Este documento, tem por finalidade a identificação das ocupações no mercado de trabalho, para fins classificatórios junto aos registros administrativos e domiciliares [1]. Segundo o MTE a nova CBO tem uma dimensão estratégica importante, na medida em que, com a padronização de códigos e descrições, poderá ser utilizada pelos mais diversos atores sociais do mercado de trabalho. Terá relevância também para a integração das políticas públicas do Ministério do Trabalho e Emprego, sobretudo no que concerne aos programas de qualificação profissional e intermediação da mão-de-obra, bem como no controle de sua implementação.

Ao visualizar a CBO, é evidente que não se limita apenas a listar ocupações de forma estática. Pelo contrário, ela representa um esforço para reconhecer, nomear, codificar e descrever as características das ocupações de maneira abrangente [2]. Uma das principais mudanças introduzidas foi a organização das ocupações em famílias, agrupando-as de acordo com suas similaridades e correspondências a domínios de trabalho mais amplos. Este método, além de proporcionar uma visão mais holística do mercado de trabalho, se preocupou em integrar a expertise prática dos profissionais que atuam nessas famílias ocupacionais [3].

Essa abordagem é evidenciada pelo MTE que em sua descrição da CBO diz que a melhor descrição de uma ocupação é aquela feita por quem a exerce diariamente, por quem compreende suas nuances e demandas intricadas. Assim, ao analisarmos as descrições presentes nesta nova versão da CBO em paralelo com as habilidades exigidas pelo mercado, podemos desvendar discrepâncias e convergências

que não apenas ilustram o cenário atual, mas também delineiam um caminho claro para as próximas etapas do desenvolvimento profissional e educacional.

A escolha do campo de comparação em ciência de dados é fundamentada no fato de que, segundo [4] dada a constante evolução da área, a pesquisa sobre os conhecimentos e habilidades necessárias nessas profissões ainda está em estágio de desenvolvimento e aprimoramento. Somada A falta de clareza na definição de cargo em ciência de dados representa um desafio tanto para o meio acadêmico, que necessita de diretrizes claras para estabelecer currículos que preparem adequadamente os futuros profissionais, quanto para a indústria, que precisa de informações precisas para definir os requisitos de cargos e recrutar profissionais qualificados [5].

No decorrer deste estudo, será explorado as competências descritas na CBO, discernindo as lacunas e os pontos de convergência com as demandas do mercado de trabalho atual em Ciência de Dados. A escolha desta área foi realizada por ser uma profissão cada vez mais recorrente nas postagens de posições e oportunidades de emprego. Uma busca em outubro de 2023 na plataforma de empregos Linkedin por vagas com a designação “ciência de dados” resultou em cerca de 1.800 ofertas vigentes de emprego no Brasil. Dessa maneira, pretende-se a partir da API de mineração de dados; Linkedin jobs scraper API, coletar estes dados da plataforma RapidAPI, não apenas, para analisar as diferenças entre as competências reconhecidas oficialmente e as necessárias para a Ciência de Dados, mas também para oferecer uma visão crítica e esclarecedora sobre como a evolução das estruturas ocupacionais pode moldar o futuro do emprego e da educação no Brasil nesta área

II. Revisão da Literatura

O uso do Processamento de Linguagem Natural (PLN) tem se expandido para além das áreas tradicionais de tecnologia, alcançando também o campo de recursos humanos e análises de mercado de trabalho. Diversos estudos demonstram como ferramentas de PLN podem ser utilizadas para analisar grandes volumes de dados de vagas de emprego, facilitando a extração de competências e habilidades demandadas pelas empresas. Por exemplo, [6] exploraram as técnicas de PLN para avaliar requisitos em vagas de ciência de dados e big data, destacando que a tecnologia permite não apenas a análise das qualificações explícitas, mas também a interpretação de competências implícitas valorizadas no mercado. Essas análises fornecem uma visão panorâmica das tendências de contratação e auxiliam na identificação de competências que surgem com frequência nas descrições de vagas.

Ferramentas de PLN também têm sido aplicadas para rastrear tendências e identificar lacunas entre as demandas do mercado e a formação profissional oferecida por instituições acadêmicas. [7] mostraram que a análise de grandes quantidades de descrições de vagas com PLN é eficaz para captar as mudanças rápidas nas habilidades técnicas, como o uso de linguagens de programação e ferramentas analíticas, especialmente em áreas como Ciência de Dados e Inteligência. No entanto, apesar do avanço dessas técnicas, desafios permanecem, especialmente na comparação de vagas com descrições ocupacionais padronizadas, como as da CBO, que nem sempre estão atualizadas com as demandas tecnológicas emergentes.

Classificações ocupacionais, como a CBO no Brasil, foram inicialmente estruturadas para proporcionar uma visão consolidada das ocupações formais, promovendo padronização na descrição de atividades e competências associadas a cada profissão [3]. No entanto, não há uma crescente discussão sobre a necessidade de atualizações frequentes dessas classificações, especialmente em áreas de alta inovação. A última atualização da CBO, realizada em 2002, abordou as transformações nas estruturas ocupacionais até então, mas as rápidas mudanças tecnológicas e a emergência de novas ocupações, como as relacionadas a dados e tecnologia, têm evidenciado uma defasagem na adequação dessas descrições às demandas reais do mercado.

Estudos recentes têm apontado a Ciência de Dados como uma área com alta demanda e com um perfil de competências que muda constantemente. [8] e [9] indicam que, além de habilidades técnicas, como domínio de ferramentas específicas e linguagens de programação, os profissionais de Ciência de Dados devem apresentar competências analíticas e uma forte capacidade de adaptação a novos métodos e tecnologias. Tais competências emergentes desafiam a estrutura de classificações como a CBO, que precisam equilibrar descrições padronizadas com flexibilidade suficiente para acompanhar as mudanças nas demandas.

III. Materiais e Métodos

A metodologia adotada neste estudo é centrada no uso de técnicas de Processamento de Linguagem Natural (PLN) para comparar descrições de vagas de trabalho de Ciência de Dados com as descrições ocupacionais formais da Classificação Brasileira de Ocupações (CBO). A escolha da análise de similaridade por meio de vetorização e medidas de similaridade cosseno baseia-se na capacidade dessas ferramentas de medir a proximidade entre textos, fornecendo uma comparação objetiva das descrições. No entanto, como essa abordagem não substitui a experiencia de campo humana, é visada a progressão posterior deste trabalho com a realização de uma revisão crítica dos resultados por especialistas da área de Ciência de Dados, o que ajudaria a validar se os termos identificados representam, de fato, as competências essenciais demandadas no mercado.

A escolha de limitar a amostra de vagas à região de São Paulo também influencia a generalização dos resultados. Embora São Paulo seja um centro importante para empregos em tecnologia no Brasil, as demandas de mercado podem variar entre regiões. O presente estudo reconhece que essa restrição geográfica pode reduzir a representatividade dos dados em nível nacional.

O fluxo de dados proposto para este trabalho é ilustrado na Figura 1, delineando os diversos módulos empregados para avaliar a correspondência das competências descritas na CBO

com a realidade do mercado. A metodologia é composta por múltiplas etapas, sendo elas: Primeiramente, a coleta de dados é feita através de uma API no caso da base de dados de vagas de empresa, e no caso do descritivo da CBO, é realizado o download do arquivo PDF, estes dados então são Pré-processados com filtros e técnicas padrão de PLN. Em segundo lugar, é realizada a extração de características de cada texto, seguido pela análise de similaridade entre as duas bases de dados. Finalmente, as métricas de desempenho são calculadas e os resultados são relatados. Cada etapa é explicada nas seguintes subseções

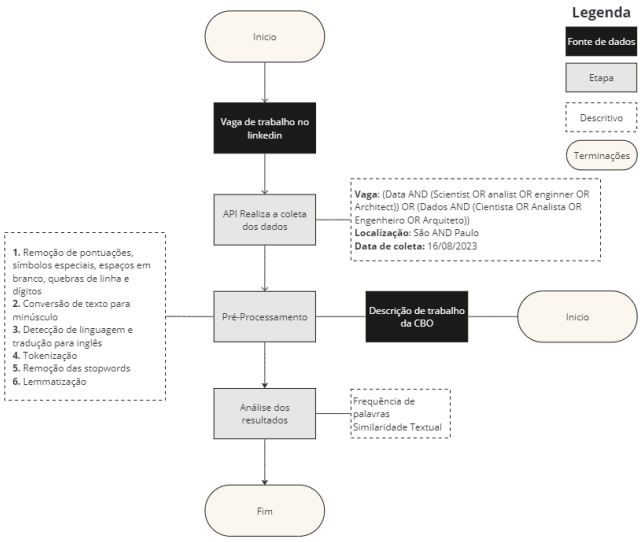

Fig. 1. Autoria própria - Metodologia de Processamento e comparação dos dados do mercado de trabalho com o descritivo CBO.

A extração relativa as ocupações de um cientista de dados são acessíveis através da seção dedicada à Classificação Brasileira de Ocupações no portal do Ministério do Trabalho do site gov.br. A extração desses arquivos específicos ocorreu em 12 de novembro de 2023. E abrangem informações que incluem os descritivos de áreas de atividade, competências pessoais, condições gerais de exercício, formação e experiência relacionadas às diversas ocupações. Esses dados fornecerão a base essencial para a comparação com as descrições de vagas extraídas do LinkedIn, enriquecendo a análise das competências exigidas no mercado de trabalho.

Na Figura 2 temos as condições gerais de exercício, extraídas do site da CBO a partir dela é proporcionado uma visão sobre a expectativa do modus operandi diário dos profissionais em questão. A descrição delineia com o cenário laboral, esboçando um retrato detalhado do ambiente em que esses profissionais desempenham suas funções.

Condições gerais de exercício

Os profissionais dessa Família ocupacional atuam em organizações públicas e privadas das mais variadas atividades econômicas. Trabalham de forma individual e/ou equipe multidisciplinar com supervisão ocasional em alguns casos. Trabalham em ambientes fechados, incluindo a modalidade à distância, em horários irregulares. O vínculo predominante é o de assalariado com carteira assinada, mas também tem profissionais que exercem suas atividades de forma autônoma.

Fig. 2. Brasil - Condições gerais de exercício de profissionais da família ciência de dados extraído do site da CBO [3].

Explorando este descritivo, é almejado empreender uma análise das condições de trabalho intrínsecas às vagas de emprego coletadas, e assimilar a realidade dos ambientes organizacionais, compreendendo as dinâmicas de trabalho e, por conseguinte, avaliando a consonância entre as expectativas apresentadas e o contexto efetivo das oportunidades de emprego. Na mesma página também é possível encontrar os requisitos de formação e experiência, A Figura 3 abaixo delineia os critérios que norteiam a preparação e trajetória profissional para o cargo de cientista de dados.

Formação e experiência

Para o exercício da ocupação de estatístico, requer-se curso superior completo, sendo desejável curso de tecnologia, cursos de especialização ou de pós-graduação. Para os cientistas de dados é desejável nível superior sem especificação de área, embora haja registro de empresas que contratam profissionais com nível médio acrescentado de cursos livres na área. O exercício pleno das atividades para o estatístico, em média, ocorre no período de três a quatro anos de experiência profissional e de um a dois anos para o cientista de dados.

Fig. 3. Brasil - Formação e experiência de profissionais da família ciência de dados [3].

O objetivo é realizar uma análise detalhada dos requisitos de formação e experiência para a ocupação de estatístico e cientista de dados. Somado a isso examinar as exigências educacionais, como cursos superiores e adicionais, bem como o tempo médio de experiência profissional necessário, proporcionando insights cruciais. Com esse entendimento, é possível avaliar a coerência entre esses requisitos e as práticas efetivas do mercado de trabalho.

Considerando as competências pessoais e técnicas esperadas, é possível extrair do descritivo da CBO uma abrangente lista de atividades que um cientista de dados é chamado a desempenhar em sua rotina diária. A Figura 4, extraída do site da CBO, exemplifica essas competências e detalham habilidades que refletem a realidade da profissão.

Competências Pessoais	
1	Demonstrar raciocínio matemático
2	Demonstrar raciocínio lógico
3	Demonstrar capacidade de síntese
4	Demonstrar organização
5	Demonstrar objetividade
6	Desenvolver capacidade analítica
7	Desenvolver perspicácia

Fig. 4. Brasil - Exemplo de Competências adotadas para os profissionais da família cientista de dados no site da CBO [3].

Esta descrição pormenorizada das competências delineadas pela CBO nos permite realizar uma comparação com as exigências técnicas e pessoais encontradas nas vagas de emprego coletadas. O intuito é discernir a congruência entre as expectativas delineadas e a dinâmica efetiva do mercado de trabalho, proporcionando uma análise das demandas e requisitos para esses profissionais especializados.

A. Captura dos dados

Na fase de coleta de dados, a base de informações relativa às vagas provém diretamente da plataforma LinkedIn, sendo obtida por meio de uma conexão da API (Interface de Programação de Aplicações) Linkedin Jobs API da plataforma RapidAPI, está por sua vez refere-se a um conjunto de regras e definições que permitem a interação entre diferentes softwares. Ela possibilita que um programa (ou parte dele) se comunique com outro, permitindo a troca de dados e funcionalidades [10].

Para a efetivação dessa captação, foi empregada uma string de pesquisa específica para os títulos das vagas nesta API, sendo ela: "(Data AND (Scientist OR Analyst)) OR (Dados AND (Cientista OR Analista))".A escolha da string de pesquisa baseou-se na análise do título e descrição da CBO, onde "Cientista de dados" e "Analista de dados" emergiram como palavras-chaves principais, tal descrição trata de atividades como desenhar amostras, analisar e processar dados, planejar atividades de pesquisa, desenvolver metodologias de análise, criar bancos de dados e comunicar resultados de análises de dados, justificando a escolha específica da string de pesquisa. Quanto à localização das vagas, a pesquisa foi restrita a São Paulo.

Os dados coletados refletem as vagas ativas na plataforma na data de 12 de novembro de 2023. Dentre as variáveis obtidas, incluem-se: quantidade de candidatos, ID da empresa, nome da empresa, URL da empresa, tipo de contrato, descrição da vaga, nível de experiência requisitado, URL da vaga, localização, tempo de postagem, nome do entrevistador, URL do perfil do entrevistador, data da publicação, salário, setor, título da vaga e tipo de trabalho. Importante ressaltar que, para o escopo deste trabalho, será utilizado apenas o descritivo da vaga para a análise em questão.

B. Pré-Processamento:

Para o pré-processamento de dados, usamos a linguagem Python e o Jupyter Notebook, uma ferramenta que permite criar notas e realizar análises em um único ambiente. Essa combinação foi escolhida pela facilidade de uso e leitura, que tornam Python uma linguagem acessível para humanos e máquinas. O Jupyter Notebook também é muito útil para documentar todo o processo de pesquisa, facilitando o acompanhamento dos fluxos de trabalho, dados e visualizações [11].

Para o carregamento e armazenamento dos dados, a biblioteca “pandas” foi utilizada, criando um dataframe para as vagas de emprego do LinkedIn e outro para as descrições de trabalho da CBO. Após o carregamento dos dados, a primeira ação consistiu na remoção de pontuações, acentos, emojis, caracteres especiais, dígitos, quebras de linhas e espaços em branco subsequentes. Além disso, todas as palavras foram transformadas para minúsculas. Essa etapa foi realizada com o auxílio da biblioteca "re", que é usada para trabalhar com expressões regulares. Expressões regulares são padrões de busca de texto que permitem realizar correspondências complexas em strings [12]. Também foi empregado o módulo "string", que fornece um conjunto de constantes e funções relacionadas a operações de strings. Enquanto a biblioteca "re" é mais focada em operações avançadas de manipulação de strings usando expressões regulares, o módulo "string" oferece constantes úteis e funções básicas para lidar com caracteres e strings.

A padronização de todas as palavras foi adotada para eliminar inconsistências devido à variação do uso de entre cada descrição de vaga, garantindo que as análises nas etapas posteriores sejam afetadas por diferenças de formatação. Somado a isso a remoção de elementos específicos do texto foi fundamentada em garantir a uniformidade e consistência nos dados, aprimorando a qualidade, além disso servem como uma medida essencial para facilitar a legibilidade do texto e evitar possíveis distorções na comparação entre as descrições a frente.

Ao examinar o dataframe das vagas do LinkedIn, após a padronização, ainda é evidente a influência das culturas e regionalidades de cada empresa. Uma característica proeminente é a variação nas linguagens utilizadas nas descrições de vagas, com destaque para a presença frequente de idiomas além do português, sendo o mais notável o inglês. Essa diversidade linguística reflete as exigências específicas de cada empresa, para subverter essa diferenciação e realizar a uniformização dos textos é necessário que todos estejam na mesma linguagem.

Para contornar essa variabilidade, torna-se imperativo identificar as diferentes linguagens presentes nas descrições e proceder com a tradução quando necessário. Para realizar essa identificação, adotou-se a biblioteca langid, uma ferramenta autônoma de identificação de idioma. A Tabela 1 a seguir apresenta a quantidade de idiomas encontrados em cada

linguagem, oferecendo insights sobre a diversidade linguística nas vagas analisadas. Este processo de identificação, contribuirá para a velocidade do processamento das traduções, pois não será necessário traduzir as descrições já presentes na linguagem final.

TABELA I
QUANTIDADE DE IDIOMAS IDENTIFICADOS PARA TRADUÇÃO

Linguagem	Quantidade de identificações
Português	
Inglês	32
Espanhol	1
Total	478

A etapa de classificação da linguagem não será aplicada ao descritivo da CBO, uma vez que este consiste em uma única string e já se estabeleceu que o texto está redigido na língua portuguesa. Adicionalmente, a consideração de tal classificação para este contexto revela-se dispendiosa em termos de processamento, dado que envolveria a identificação da língua seguida pela tradução, procedimento este mais custoso do que a tradução direta.

Para realizar as traduções, foi empregada a biblioteca deep_translator, a qual se integra por meio de API com as principais ferramentas de tradução disponíveis, viabilizando uma conversão dos textos. A linguagem escolhida para traduzir o texto foi a língua portuguesa, essa escolha se deu por conta da quantidade e disponibilidade de conteúdos já encontrados em português e que já oferecem suporte a esta língua. Uma outra vantagem significativa desse método reside na correção automática dos principais erros gramaticais durante o processo de tradução. Como parte desse procedimento, foi essencial incorporar tratamentos para exceções de conexão, implementando tentativas adicionais após um intervalo de um segundo. Adicionalmente, a subdivisão das descrições de vagas em blocos de, no máximo, 5000 tokens se fez necessária devido às limitações da biblioteca utilizada.

A seguir, procedemos com o processo de tokenização do texto, essa etapa visa a fragmentação das sequências de caracteres da entrada (nesse caso as descrições em inglês), percebida como uma longa sucessão de caracteres para um computador, em subunidades chamadas tokens, está por sua vez, é uma prática essencial, uma vez que representa a base para análises mais granulares e profundas do texto [13]. Para efetuar esse processo, recorremos ao pacote RegexpTokenizer da biblioteca NLTK. Essa abordagem nos permite explorar as nuances e características específicas de cada unidade textual, contribuindo para uma análise mais refinada e precisa.

Em seguida, utilizamos o pacote stopwords, proveniente da biblioteca NLTK, com o intuito de eliminar as stopwords contidas no dataframe. O termo "stopwords" refere-se a partes do texto em um documento que carregam poucas informações sobre a parte específica do texto a que pertencem e, portanto, necessitam ser removidas durante a etapa de préprocessamento [14]. Essas palavras comuns, como artigos, preposições e conjunções, não contribuem significativamente para a compreensão do conteúdo e podem introduzir ruídos nos resultados da análise.

Para realizar essa remoção, foi feito o download das stopwords para a língua inglesa e excluímos todas as ocorrências dessas palavras insignificantes que aparecem nas descrições de vagas. Isso aprimora a qualidade e relevância dos dados analisados. Importante mencionar que, nesse estágio, é de caráter essencial uma função para a inclusão de palavras a serem excluídas. Essa abordagem se tornou crucial, pois certas palavras, como "São Paulo", "Localização" e "Remuneração", ocorrem frequentemente, mas contribuem pouco para a comparação de um descritivo genérico da CBO com outro regionalizado para o estado de São Paulo. A inclusão dessa função proporciona flexibilidade e adaptabilidade ao processo, permitindo ajustes específicos dessas palavras conforme necessário.

A etapa subsequente à remoção de stopwords consiste na exclusão de palavras infrequentes. Esta medida é adotada com base na imperatividade de eliminar termos como nomes de empresas, marcas, nome de entrevistadores, erros de gramática graves e que, embora comuns em determinados contextos, podem introduzir ruído indesejado na análise. A justificativa para tal procedimento reside na necessidade de otimizar a precisão da análise, focando nas palavras que verdadeiramente contribuem para a caracterização das competências em questão, visando a remoção de termos redundantes e pouco informativos, sem comprometer a integridade das palavras-chave cruciais para a análise final.

A determinação do limiar de frequência para exclusão automatizada de palavras foi fundamentada em uma abordagem experimental contínua ao longo do desenvolvimento do trabalho. Optou-se por estabelecer que palavras cuja frequência de ocorrência fosse inferior a 5 repetições fossem automaticamente excluídas do conjunto de dados. Essa decisão foi meticulosamente ponderada, representando uma estratégia para assegurar que apenas termos relevantes e significativos perdurassem no corpus textual, promovendo uma seleção criteriosa de vocabulário que resiste à interferência de elementos menos contributivos para o escopo do estudo em uma busca equilibrada por manter tokens importantes e excluir os desnecessários.

A última etapa do pré-processamento é a lematização, que normaliza as palavras, ou seja, as reduz à sua forma base. Essa etapa é parecida com a stemmatização, mas não reduz a palavra até sua raiz, mantendo-a em uma forma mais natural e legível [15].

Para ilustrar, no dataset de vagas podemos encontrar as palavras “Trabalhando”, “Trabalho” e “Trabalhará” A lematização dessas palavras ajustaria seus sufixos de maneira que convergissem para a forma infinitiva " Trabalhar”. Notavelmente, nesta instância, tanto a forma normalizada quanto a raiz da palavra é idêntica. Cabe ressaltar que, em

alguns casos, a forma normalizada pode divergir da raiz da palavra. Por exemplo, as palavras "computador" "computar" e "computação" seriam derivadas para a raiz “ comput”, mas a forma normalizada seria o infinitivo do verbo, ou seja, "computar” Esta intricada operação de lematização proporciona uma representação unificada e simplificada das palavras, contribuindo, assim, para a consistência e clareza na análise linguística e textual [16].

C. Metodologia de processamento da análise de frequência:

Após o pré-processamento, a análise de frequência de palavras foi conduzida para ambas as fontes de dados. Este método consiste em utilizar a funcionalidade freqdist da biblioteca nltk para avaliar a frequência de palavras em ambas as fontes de dados, permitindo identificar as palavras mais comuns em cada conjunto e proporcionando uma visão abrangente das competências e habilidades frequentemente mencionadas tanto pela CBO quanto pelos empregadores do LinkedIn. A expectativa é que ao examinar as palavras com maior frequência, seja possível realizar uma comparação direta das competências e requisitos exigidos por ambas as fontes.

Além disso, foi construída a identificação de tokens comuns e incomuns entre os conjuntos de dados. Ao verificar tais tokens compartilhados, será possível destacar as competências universalmente reconhecidas em determinada ocupação. E em contraste, a identificação de tokens exclusivos em cada conjunto de dados revelará nuances específicas das descrições de vagas do LinkedIn em comparação com a descrição geral da CBO.

D. Metodologia de processamento da análise de similaridade:

A análise de similaridade utiliza ferramentas como o CountVectorizer para criar representações vetoriais. Este processo converte informações textuais em espaços numéricos multidimensionais, onde palavras, frases ou documentos são representados por vetores numéricos. Assim, a metodologia permite a comparação quantitativa entre as descrições das vagas.

Para isso foi realizada a criação de uma nova coluna no DataFrame de vagas, denominada 'similaridade', destinada a armazenar os resultados das análises de similaridade entre o dataframe a as descrições da CBO. O vetorizador é configurado para considerar n-gramas de 1 a 3 tokens, este por sua vez são unidades de texto compostas por n elementos consecutivos, geralmente palavras ou caracteres. O valor de n denota o número de elementos que compõem cada n-grama. O objetivo é gerar uma representação numérica das palavras presentes nas descrições, levando em conta não apenas palavras isoladas, mas também sequências contíguas de até três palavras.

O loop sobre as linhas do DataFrame permite a aplicação do vetorizador a cada descrição de vaga tratada, transformando o texto em um vetor numérico. A similaridade cosseno entre a descrição atual e uma descrição de referência também transformada é calculada, essa medida do cosseno do ângulo entre esses vetores no espaço vetorial, pode ser entendida porquanto mais próximos estão os vetores, menor é o ângulo e, consequentemente, maior é o valor do cosseno e, portanto, da similaridade proporcionando uma medida de quão próximas ou distantes estão essas duas descrições no espaço vetorial criado pelo vetorizador [17], de acordo com a eq. 1. A similaridade resultante é então armazenada na coluna 'similaridade' do DataFrame, onde são extraídos os principais indicadores estatísticos (Média, desvio padrão, soma, mínimo, máximo e quartis).

$$\text{Eq.1 } Similaridade\ Cosseno(A,B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

onde, A·B denota o produto escalar entre os vetores A e B; ‖A‖ representa a norma Euclidiana do vetor A; ‖B‖ representa a norma Euclidiana do vetor B.

IV. Análise dos Resultados

A. Análise de frequência:

Os resultados obtidos ao analisar a frequência dos dados disponíveis no mercado para a posição de Cientista de Dados fornecem insights valiosos sobre as expectativas dos empregadores em relação aos candidatos. A visualização da nuvem de palavras das descrições de vagas para cientista de dados na Figura 5, revela termos recorrentes que são altamente enfatizados no ambiente de trabalho. Palavras-chave como "experiência", "conhecimento", "desenvolvimento" e "equipe" destacam-se como elementos essenciais, apontando para a relevância de habilidades práticas e experiência profissional na busca por profissionais qualificados.

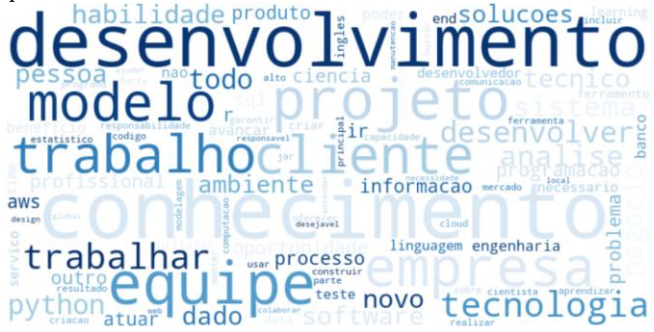

Fig. 5. Autoria própria - Nuvem de palavras da frequência extraída para os dados do mercado de trabalho.

Além disso, a presença de termos como "Python", "AWS", "SQL", "data", "machine learning" e "tecnologia" sinaliza uma demanda vigorosa por habilidades técnicas e conhecimento específico em ferramentas essenciais para o campo da Ciência de Dados. Essa ênfase técnica indica a necessidade de profissionais familiarizados e competentes em plataformas e linguagens de programação relevantes para o cenário contemporâneo. Adicionalmente, a inclusão de termos como "remoto", "flexível" e "benefício" reflete uma tendência moderna de valorização do trabalho remoto e benefícios

flexíveis. Esses elementos não apenas denotam uma adaptação às mudanças nas dinâmicas de trabalho, mas também se alinham as condições gerais de exercício encontradas na Classificação Brasileira de Ocupações (CBO), indicando uma convergência entre as expectativas do mercado e as diretrizes oficiais. Essa contextualização ampla das palavras mais frequentes proporciona uma compreensão mais completa das tendências atuais no recrutamento para profissionais de Ciência de Dados.

Ao emitir a mesma análise de frequência para o descritivo da CBO podemos ampliar a compreensão das competências essenciais associadas a ocupação de cientista de dados destacando as demandas mais genéticas do mercado de trabalho, segundo a Figura 6.

Fig. 6. Autoria própria - Nuvem de palavras da descrição da CBO para família de cientista de dados.

Torna-se evidente a ênfase em termos como "análise", "estatístico", "matemático" e "modelo", ressaltando a demanda premente por habilidades analíticas e quantitativas inerentes a esta ocupação específica. A presença de palavras como "definir", "elaborar", "amostra", "identificar", "dados" e "validar" sugere uma abordagem metodológica e científica, alinhada à área de dados, indicando uma aplicação robusta de métodos estatísticos, conforme esperado para a ocupação.

Além disso, a visualização de termos como "multidisciplinar" e "equipe" destaca a natureza colaborativa da profissão, ressaltando a importância de atuar em diversos contextos e interagir efetivamente em ambientes de equipe.

Comparando as habilidades demandadas pelo mercado com as da CBO, nota-se a presença frequente de termos como ‘modelo’, ‘análise’, ‘processo’ e ‘dados’, o que destaca a importância de competências voltadas ao desenvolvimento e análise de modelos. O ensino superior, mencionado de forma geral pela CBO, aparece como requisito em 77 das vagas analisadas, reforçando a necessidade de formação especializada.

No entanto, ao explorar as divergências, surge uma notável presença de termos técnicos específicos no conjunto de dados das vagas, como "python", "r", "aws", "sql" e "inglês", indicando diferenças nas ênfases e competências percebidas em cada fonte. Essa análise ressalta a importância de uma abordagem dinâmica e adaptável por parte dos profissionais, que devem não apenas atender às expectativas gerais delineadas pela CBO, mas também incorporar habilidades técnicas específicas para se destacarem em um mercado de trabalho em constante evolução.

A Enquanto a CBO utiliza termos mais gerais, o mercado de trabalho enfatiza tecnologias e ferramentas específicas e reflete tendências atuais, mostrando maior demanda por habilidades técnicas modernas. Isso sugere que a CBO pode precisar de atualizações frequentes para acompanhar as rápidas mudanças tecnológicas e as novas exigências do mercado.

B. Análise de similaridade

A aplicação da estatística descritiva na avaliação das similaridades revelou dados significativos, apresentando uma média de similaridade de 14,02% entre o descritivo da CBO e as vagas do mercado de trabalho, com um desvio padrão de 11,08%. No entanto, destaca-se a presença de vagas que exibem similaridades surpreendentes, atingindo até 48,52% em relação ao descritivo padrão da CBO. Essa variação substancial aponta para uma diversidade notável na aderência das descrições de vagas às diretrizes estabelecidas pela CBO.

A representação visual dessas distribuições é apresentada na Figura 7, onde um histograma delineia a contagem de vagas de acordo com seus níveis de similaridade. Esse gráfico oferece uma visão panorâmica da dispersão das similaridades, permitindo uma análise mais aprofundada das características distintas de cada categoria.

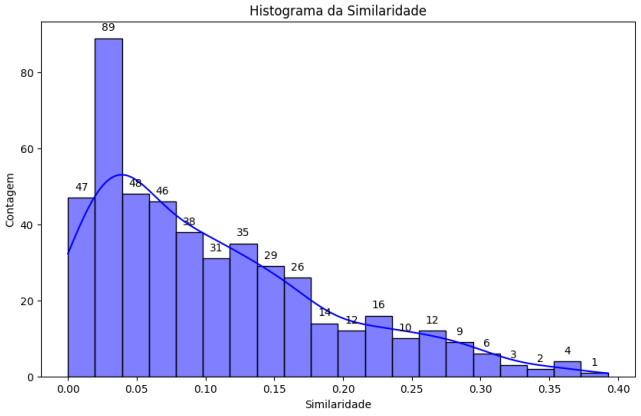

Fig. 7. Autoria propria - Histograma da distribuição de similaridade das vagas do mercado de trabalho para cientista de dados com a descrição da CBO para família de cientista de dados.

Ao aprofundar a análise das vagas que compartilham as semelhanças mais proeminentes, destaca-se uma tendência notável dessas oportunidades em adotar de forma mais abrangente os requisitos delineados pela CBO, concentrando se em aspectos mais amplos. Alguns dos exemplos encontrados como "conhecimentos estatísticos", manipulação de bancos de dados e linguagens de programação, evitam entrar em detalhes significativos sobre as competências técnicas específicas de cada uma dessas áreas. Por exemplo, na área de linguagem de programação, 56,5% das vagas não mencionaram explicitamente as competências técnicas específicas, como Python e SQL, nem suas funcionalidades relacionadas, como modelagem de dados e visualização.

Em contraste, as vagas abaixo de 25% similaridade com a CBO tendem a oferecer detalhamentos mais extensos sobre as competências necessárias, explicando o que desejam de cada área tais como linguagens de programação e bancos de dados além de dedicar espaço para uma explanação mais abrangente sobre a própria empresa e suas particularidades.

O Box Plot apresentado na Figura 8 mostra a similaridade das vagas do mercado de trabalho para cientista de dados em relação à descrição da CBO para a família de cientista de dados. A maior parte das similaridades está concentrada entre 3,41% e 15,34%, com a mediana em torno de 8,40%, indicando que metade das vagas possui uma similaridade abaixo desse valor e a outra metade acima. A presença de outliers acima de 33,23% revela que algumas vagas possuem uma alta similaridade com a descrição da CBO, destacando casos onde a demanda do mercado se alinha fortemente com as descrições oficiais.

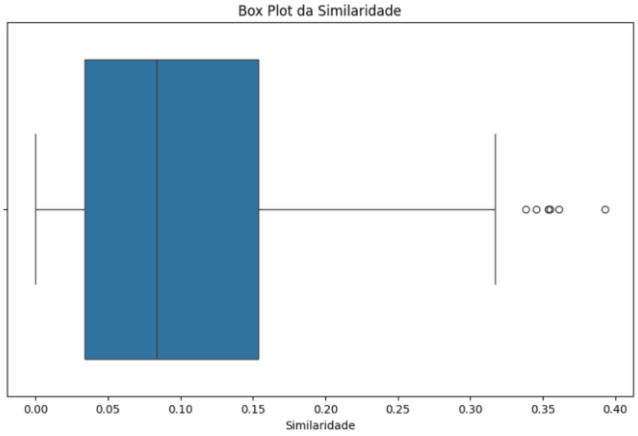

Fig. 8. Autoria propria - Box-plot da similaridade entre o descritivo e as vagas do mercado de trabalho para ciência de dados.

A análise sugere que, embora exista uma correspondência entre as descrições da CBO e as demandas do mercado de trabalho para cientistas de dados, há uma variabilidade significativa. Essa variabilidade reflete as diferentes exigências e especificidades das vagas disponíveis no mercado atual, indicando que as descrições da CBO podem não capturar totalmente as nuances e habilidades específicas procuradas pelos empregadores modernos.

Em última análise, essa estatística ressalta a importância de considerar a diversidade nas abordagens das descrições de vagas, reconhecendo que há uma variedade de estratégias adotadas pelos empregadores na formulação de anúncios de emprego. Além da omissão de informações para atrair uma gama mais ampla de candidatos, outros métodos incluem a utilização de jargões específicos da indústria, a aplicação de requisitos exagerados ou pouco realistas e a descrição de responsabilidades vagas ou genéricas. Essa compreensão mais profunda é fundamental para os profissionais que buscam alinhar suas habilidades e experiências às expectativas do mercado de trabalho de maneira mais eficaz.

V. Conclusão

A execução da análise de frequência revelou-se uma ferramenta essencial para a obtenção de uma visão mais abrangente do cenário profissional, proporcionando uma compreensão detalhada das demandas específicas do mercado de trabalho. É notável que, ao conduzir essa análise, foi possível identificar uma gama de competências técnicas que não foram abordadas nas descrições padronizadas da CBO, mas que aparecem em mais de 50% de todas as vagas, isso inclui habilidades de programação em Python, SQL, machine learning e banco de dados. E que embora a CBO seja eficaz em orientar, de maneira geral, os profissionais que ingressam no mercado, sua abordagem genérica torna-se insuficiente para fornecer, de modo específico, as habilidades atualmente requisitadas pelos empregadores.

Essa lacuna destaca a importância de identificar e analisar as competências técnicas mais relevantes para que os profissionais possam se manter competitivos em um mercado dinâmico e em constante mudança. Estar atento às competências emergentes é essencial para que os profissionais acompanhem as tecnologias, linguagens e ferramentas mais atuais em suas áreas.

Dessa forma, a análise de frequência não apenas expõe lacunas entre as expectativas do mercado e as descrições padrão, mas também destaca a necessidade contínua de os profissionais estarem sintonizados com as tendências e inovações do setor para garantir sua relevância e competitividade. Em última análise, a realização dessa abordagem não apenas informa, mas capacita os profissionais que queiram adentrar no mercado de forma a serem capazes de tomar medidas proativas na construção de suas trajetórias profissionais.

Os resultados da análise indicam uma média de similaridade de apenas 14,02% entre as descrições da CBO e as vagas de trabalho de Ciência de Dados. Esse percentual de baixa correspondência reflete uma discrepância significativa entre as competências oficialmente reconhecidas e as exigidas pelo mercado, o que sugere uma defasagem na CBO para atender às necessidades de uma área tão dinâmica quanto a Ciência de Dados.

As implicações dessa baixa similaridade podem ser vistas especialmente para a formação de novos profissionais e a atualização de currículos acadêmicos. Em instituições que utilizam a CBO como base para a estrutura curricular, há um risco de os estudantes concluírem os cursos sem as habilidades específicas que o mercado demanda, como o domínio de linguagens de programação específicas (Python e SQL) e conhecimentos em machine learning. Essa lacuna pode dificultar a inserção dos graduandos no mercado e gerar uma necessidade de treinamento complementar.

Além das consequências para a formação profissional, as divergências encontradas neste estudo revelam a necessidade de uma atualização contínua das descrições da CBO para alinhá-las com as exigências práticas das empresas. O estudo sugere que políticas públicas na área de educação e mercado

de trabalho poderiam se beneficiar de descrições ocupacionais mais dinâmicas, especialmente em áreas de alta inovação.

Ao verificar trabalhos similares como os de [17] e [18] é possível ver a corroboração com a análise ao destacar as principais competências e a dinâmica do mercado de trabalho, tecnologias e competências exigidas. Os autores enfatizam os impactos da transformação digital e a necessidade de habilidades emergentes, enquanto Schuster aborda o perfil demandado dos profissionais de tecnologia da informação. A integração dessas perspectivas reforça a importância de uma abordagem adaptativa e informada para os profissionais se manterem competitivos no mercado atual.

AGRADECIMENTOS

À vida, cujos detalhes meticulosos moldaram minha jornada. Em meio a desafios, encontrei inspiração em suas sutilezas, permitindo-me ser quem sou hoje. Agradeço à beleza simples e grandiosa que, silenciosamente, guiou meu caminho.

REFERÊNCIAS

[1] Brasil. 2002. Ministério do Trabalho. Classificação Brasileira de Ocupações. Brasília. Disponível em: < http://www.mtecbo.gov.br/cbosite/pages/saibaMais.jsf>. Acesso em: 20 out. 2023.

[2] Ayoobzadeh, M. 2022. "Freelance job search during times of uncertainty: protean career orientation, career competencies and job search. Personnel Review 6: 40–56.

[3] Brasil. 2024. Ministério do Trabalho. Classificação Brasileira de Ocupações. Brasília. Acesso em: 12 nov. 2023.

[4] Radovilsky, Z. et al. 2018. "Skills requirements of business data analytics and data science jobs: A comparative analysis. In: Journal of Supply Chain and Operations Management. 2018, California, USA. p. 82-101.

[5] Halwani, M.A., Amirkiaee, S.Y., Evangelopoulos, N. and Prybutok, V. 2022, "Job qualifications study for data science and big data professions", Information Technology & People, p 510-525.

[6] Cegielski, Casey & Jones-Farmer, L.A.. (2016). Knowledge, Skills, and Abilities for Entry-Level Business Analytics Positions: A Multi-Method Study. Decision Sciences Journal of Innovative Education. 14. 91-118. 10.1111/dsji.12086.

[7] Halwani, Marwah & Amirkiaee, S. & Evangelopoulos, Nicholas & Prybutok, V.R.. (2021). Job qualifications study for data science and big data professions. Information Technology & People. ahead-of-print. 10.1108/ITP-04-2020-0201.

[8] Glez-Peña, D; Lourenço, A; López-Fernández, H.; Reboiro-Jato, M.; Fdez-Riverola, F. 2014. “Web scraping technologies in an API world. “In Briefings in Bioinformatics 15: 788–797.

[9] Wilkinson, M.; Dumontier, M.; Aalbersberg, I. et al. 2016. “The FAIR Guiding Principles for scientific data management and stewardship.” In Scientific Data 3:18-16.

[10] Palleta, Francisco Carlos e Moreiro Gonzalez, José Antonio. 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, 2021Tradução . . Disponível em: https://doi.org/10.47989/irpaper904. Acesso em: 01 nov. 2024.

[11] Python. 2022. re — Operações com expressões regulares. Acesso em: 20 jan. 2024.

[12] Grefenstette, G. 1999. Tokenization. p.117-133. In: van Halteren, H. Syntactic Wordclass Tagging: Text, Speech and Language Technology, 9 ed.Kremmen, Berlim, Alemanha.

[13] Sarica, S.; Luo, J. 2021. Stopwords in technical language processing. PLOS ONE 16: 25 - 49.

[14] Balakrishnan, V.; Lloyd-Yemoh, E. 2014. “Stemming and lemmatization: A comparison of retrieval performances.” in Software Engineering 2: 174-179.

[15] Bergmanis, T.; Goldwater, S. 2018. Context-sensitive neural lemmatization with lematus. In 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018, New Orleans, Louisiana, United States. Anais...p.1391-1400.

[16] Sintia, S.; Defit, S.; Nurcahyo, G. 2021. “Product Codefication Accuracy With Cosine Similarity And Weighted Term Frequency And Inverse Document Frequency (TF-IDF).” in Journal of Applied Engineering and Technological Science (JAETS), 62-69.

[17] Paletta, F. C., & 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).

[18] Schuster, M. E. 2008. Mercado de trabalho de tecnologia da informação: O Perfil dos profissionais demandado