UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
DEPARTAMENTO DE ENGENHARIAS E TECNOLOGIA
INTERDISCIPLINAR EM CIÊNCIA E TECNOLOGIA

Davi Emmanuel de Lima Rodrigues

Produtividade científica brasileira sob a perspectiva de gênero durante a pandemia de Covid-19: classificação e análise via aprendizado de máquina

PAU DOS FERROS
2023

Davi Emmanuel de Lima Rodrigues

Produtividade científica brasileira sob a perspectiva de gênero durante a pandemia de Covid-19: classificação e análise via aprendizado de máquina

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Ciência e Tecnologia.

Orientadora: Profa. Dra. Rosana Cibely Batista Rego

PAU DOS FERROS

2023

R696p Rodrigues, Davi Emmanuel.
Produtividade científica brasileira sob a perspectiva de gênero durante a pandemia de Covid-19: classificação e análise via aprendizado de máquina / Davi Emmanuel Rodrigues. - 2023.
41 f. : il.

Orientadora: Rosana Cibely Rego.
Monografia (graduação) - Universidade Federal Rural do Semi-árido, Curso de Ciência e Tecnologia, 2023.

1. COVID-19. 2. produção acadêmica. 3. classificação de gênero. 4. aprendizado de máquina. I. Rego, Rosana Cibely, orient. II. Título.

Ficha catalográfica elaborada por sistema gerador automáto em conformidade com AACR2 e os dados fornecidos pelo) autor(a).
Biblioteca Campus Pau dos Ferros
Bibliotecária: Erlanda Maria Lopes da Silva
CRB: 15/966

Davi Emmanuel de Lima Rodrigues

Produtividade científica brasileira sob a perspectiva de gênero durante a pandemia de Covid-19: classificação e análise via aprendizado de máquina

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Ciência e Tecnologia.

Defendida em: 06 / 10 / 2023.

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
ROSANA CIBELY BATISTA REGO
Data: 17/10/2023 13:40:33-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Rosana Cibely Batista Rego (UFERSA)
Presidente

gov.br Documento assinado digitalmente
HULIANE MEDEIROS DA SILVA
Data: 19/10/2023 13:35:25-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Huliane Medeiros Da Silva (UFERSA)
Membro Examinador

gov.br Documento assinado digitalmente
NATHALEE CAVALCANTI DE ALMEIDA LIMA
Data: 17/10/2023 16:21:22-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Náthalee Cavalcanti de Almeida Lima (UFERSA)
Membro Examinador

"*Tudo o que temos que decidir é o que fazer com o tempo que nos é dado*"

Gandalf

RESUMO

A pesquisa científica em todo o mundo enfrentou desafios significativos em decorrência da pandemia de COVID-19 e das medidas de distanciamento social necessárias. Esses desafios abrangem desde a acessibilidade a equipamentos até questões socioculturais e econômicas que podem ter impactado a pesquisa. Este estudo se concentra em analisar o impacto da pandemia de COVID-19 na pesquisa científica realizada no Brasil, com ênfase na produção de artigos científicos completos publicados em periõdicos no período de 2018 a 2021, considerando o gênero dos pesquisadores. Para coletar os dados necessários, foi desenvolvido um programa de extração de informações que coletou os nomes de pesquisadores brasileiros vinculados aos artigos. Além disso, foram aplicados diversos modelos de aprendizado de máquina, incluindo *SVM*, *BiLSTM* e *CNN*, para classificar o gênero dos autores. Os resultados obtidos revelaram que alguns desses modelos alcançaram uma impressionante taxa de precisão superior a 95% na identificação do gênero dos pesquisadores. Uma descoberta relevante foi a queda acentuada de 37,47% nas publicações de artigos científicos por pesquisadores brasileiros em 2021. Essa queda foi observada tanto em pesquisadores do gênero masculino quanto do feminino, mas os modelos de aprendizado de máquina aplicados mostraram que a redução foi mais acentuada para o gênero feminino na maioria dos casos. Essa tendência reflete desafios adicionais enfrentados pelas pesquisadoras devido à distribuição desigual de responsabilidades domésticas e de cuidados familiares. Esses resultados destacam a importância de compreender o impacto das crises, como a pandemia, na pesquisa científica e ressaltam a necessidade de medidas que promovam a equidade de gênero na academia.

Palavras-chave: COVID-19; produção acadêmica; classificação de gênero; aprendizado de máquina.

ABSTRACT

Scientific research activities, in general, have been affected due to the COVID-19 pandemic and the need for social distancing. The pandemic may have impacted scientific research due to factors such as access to equipment, sociocultural issues, and economic challenges. In this study, we conduct an analysis of the impact of COVID-19 on Brazilian scientific research by examining the number of full articles published in journals from 2018 to 2021, taking into account the gender of the researchers. To collect the necessary data, a web crawler was developed to extract the names of Brazilian researchers from the articles. Furthermore, several machine learning models, including SVM, BiLSTM, and CNN, were applied to classify the authors' gender. Remarkably, some models achieved a gender prediction accuracy of over 95% in many cases. A significant finding was a sharp 37.47% decline in the publications of articles by Brazilian researchers in 2021. This decline was observed for both male and female researchers, but machine learning models indicated that the reduction was more pronounced for the female gender in most cases. This trend highlights additional challenges faced by female researchers due to unequal distribution of domestic and caregiving responsibilities. These results underscore the importance of understanding the impact of crises, such as the pandemic, on scientific research and emphasize the need for measures to promote gender equity in academia.

Keywords: COVID-19; academic production; gender classification; machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A pandemia de COVID-19 trouxe consigo um cenário de desafios sem precedentes, afetando profundamente diversas esferas da vida em todo o mundo. No Brasil, não foi diferente, e a crise sanitária revelou-se particularmente impactante, com implicações significativas nas áreas de saúde, economia e educação (GONÇALVES et al., 2022; ORELLANA et al., 2021). À medida que a necessidade de distanciamento social e o imperativo de evitar aglomerações se tornaram predominantes, as atividades presenciais, incluindo o ensino e as atividades acadêmicas, foram abruptamente interrompidas. Nesse contexto, o ensino remoto emergiu como uma alternativa viável (ARCINIEGAS, 2021).

No entanto, a transição para o ensino remoto revelou as profundas disparidades sociais e econômicas que permeiam a sociedade brasileira. Para participar efetivamente do ensino remoto, um requisito fundamental é o acesso à internet de qualidade e a posse de equipamentos adequados, recursos que permanecem fora do alcance de muitos brasileiros (AZEVEDO e DE ALMEIDA NEVES, 2021; CASTIONI et al., 2021). Além disso, fatores socioeconômicos e culturais podem representar barreiras substanciais para a produtividade no trabalho remoto (AZEVEDO e DE ALMEIDA NEVES, 2021). É relevante destacar que as responsabilidades relacionadas ao cuidado de crianças, idosos e doentes, que recaem de forma desproporcional sobre as mulheres, frequentemente limitam a capacidade de cumprir as demandas do trabalho remoto de maneira satisfatória (MANCEBO, 2020).

Nesse contexto, surge a necessidade de investigar as mudanças e implicações na produtividade científica, especialmente no que se refere à publicação de artigos científicos, em meio à transição das atividades presenciais para as remotas. Este trabalho propõe-se a abordar essa questão, concentrando-se na realidade brasileira. Para tanto, realizamos a coleta de dados relacionados às pesquisas acadêmicas conduzidas no Brasil no período de 2018 a 2021, com base na Plataforma Lattes. No entanto, vale ressaltar que a Plataforma Lattes não fornece informações explícitas sobre o gênero dos autores dos trabalhos, tornando necessária a classificação de gênero como parte deste estudo.

A produtividade acadêmica sob a perspectiva de gênero é um tema crucial que merece destaque, especialmente em um contexto de crise como o da pandemia de COVID-19. Pesquisas anteriores já demonstraram que as dificuldades enfrentadas por pesquisadores durante a suspensão das atividades presenciais afetam de forma desigual homens e mulheres (KAFRUNI, 2020). As disparidades são evidentes não apenas entre aqueles que têm ou não

filhos, mas também em relação à capacidade de ambos os gêneros de se adaptarem ao trabalho remoto.

A classificação de gênero com base em nomes pode parecer uma tarefa simples, porém, quando lidamos com centenas ou milhares de nomes, torna-se uma tarefa complexa e potencialmente exaustiva (ZHAO, 2017). Automatizar esse processo por meio de modelos de *machine learning* e *deep learning* para a predição de gênero torna-se, assim, uma solução viável (DE SOUSA; DE OLIVEIRA SANTIAGO e DIAS, 2019; HORHIRUNKUL et al., 2021).

Existem diversos estudos que preveem o gênero a partir de imagens de rostos (ZHANG et al., 2017; VENUGOPAL; YADUKRISHNAN e NAIR, 2020), assim como pesquisas que propõem a previsão de gênero utilizando processamento de linguagem natural (VASHISTH e MEEHAN, 2020; KARIMI et al., 2016). Considerando a capacidade de classificação dos modelos de *machine learning* e *deep learning*, este trabalho se propõe a aplicar técnicas de inteligência computacional para analisar o efeito da pandemia de COVID-19 na publicação de artigos acadêmicos.

Diante desse contexto, a problematização central deste trabalho pode ser formulada da seguinte maneira:

Como a classificação automatizada do gênero dos autores de artigos acadêmicos, por meio de técnicas de inteligência computacional, pode contribuir para a compreensão do impacto da pandemia de COVID-19 na produtividade científica brasileira, considerando a perspectiva de gênero? Quais são os desafios e benefícios associados ao uso de modelos de *machine learning* e *deep learning* para a predição de gênero?

A justificativa para a realização deste trabalho é baseada na relevância e nas lacunas existentes no conhecimento sobre o impacto da pandemia de COVID-19 na produtividade científica brasileira, considerando a perspectiva de gênero. Embora tenham sido realizados estudos anteriores sobre os efeitos da transição para o ensino remoto e suas implicações na produtividade acadêmica (KAFRUNI, 2020; MANCEBO, 2020), pouca atenção tem sido dada à análise das disparidades de gênero nesse contexto.

A pandemia de COVID-19 causou uma mudança drástica nas atividades acadêmicas e científicas, levando à adoção generalizada do trabalho remoto e ao uso intensificado de tecnologias de comunicação e colaboração online. Essa transição teve impactos significativos

na forma como os pesquisadores conduzem suas atividades e interagem com suas comunidades acadêmicas (AZEVEDO e DE ALMEIDA NEVES, 2021).

Ao realizar a classificação automatizada do gênero dos autores de artigos acadêmicos por meio de técnicas de inteligência computacional, este estudo busca preencher essa lacuna de conhecimento, fornecendo informações valiosas sobre as implicações da pandemia na produtividade científica brasileira sob uma perspectiva de gênero. Compreender essas disparidades é essencial para promover a igualdade de oportunidades na academia e para informar políticas e práticas que visem mitigar os efeitos negativos da pandemia.

Além disso, ao aplicar técnicas de inteligência computacional, como *machine learning* e *deep learning*, na classificação do gênero, este estudo também contribuirá para o desenvolvimento e avanço dessas áreas de pesquisa, explorando sua aplicabilidade em contextos acadêmicos e científicos.

Portanto, a realização deste trabalho é justificada pela necessidade de compreender e abordar as disparidades de gênero na produtividade científica decorrentes da pandemia de COVID-19, a fim de promover a equidade de gênero e melhorar as condições de trabalho e pesquisa para todos os pesquisadores no Brasil.

1.1 OBJETIVO GERAL

Investigar o impacto da pandemia de COVID-19 na produtividade científica brasileira, considerando a perspectiva de gênero, por meio da classificação automatizada do gênero dos autores de artigos acadêmicos utilizando técnicas de inteligência computacional, com foco na análise das mudanças e implicações decorrentes da transição para o trabalho remoto.

1.2 OBJETIVOS ESPECÍFICOS

- Criar de um *crawler* para gerar uma base de dados acadêmicos recentes extraídos da Plataforma Lattes. Para a geração da base de dados, obteve-se o nome dos autores e títulos dos artigos publicados em revistas entre 2018 e 2021;
- Aplicar a técnica máquina de vetor de suporte (*Support Vector Machine - SVM*) e dos modelos de *deep learning BiLSTM* (*Bidirectional Long Short-Term Memory*) e *1D-CNN* (*One-dimensional Convolutional Neural Network*) para classificação de gênero dos pesquisadores com base nos artigos publicados a partir dos nomes dos pesquisadores;

1.3 PUBLICAÇÕES DERIVADAS DESTE TRABALHO

- REGO, Rosana Cibely B. et al. Brazilian scientific productivity from a gender perspective during the Covid-19 pandemic: classification and analysis via machine learning. **IEEE Latin America Transactions**, v. 21, n. 2, p. 302-309, 2023.

2 REVISÃO DA LITERATURA

2.1 IMPACTOS DA PANDEMIA NA PRODUTIVIDADE ACADÊMICA

A pandemia de COVID-19, que eclodiu em todo o mundo no início de 2020, teve um impacto sem precedentes em diversas áreas da sociedade. No contexto brasileiro, a crise de saúde pública revelou-se especialmente desafiadora, abalando não apenas o sistema de saúde, mas também a economia e o setor educacional (GONÇALVES et al., 2022; ORELLANA et al., 2021). Como parte das medidas de contenção do vírus, as atividades presenciais, incluindo as acadêmicas, foram rapidamente interrompidas, levando à implementação massiva do ensino remoto (ARCINIEGAS, 2021).

2.1.2 DISPARIDADE SOCIAL

Essa transição para o remoto evidenciou as disparidades sociais e econômicas latentes na sociedade brasileira. O acesso desigual à internet de alta qualidade e a posse de equipamentos adequados emergiram como obstáculos significativos para a participação efetiva no ensino remoto, uma realidade que permanece inacessível a muitos brasileiros (AZEVEDO e DE ALMEIDA NEVES, 2021; CASTIONI et al., 2021). Além disso, como evidencia Azevedo (2021), fatores socioeconômicos e culturais têm desempenhado um papel fundamental na determinação da capacidade de indivíduos em se adaptar ao trabalho remoto. Mancebo (2021) destaca que as responsabilidades de cuidado com crianças, idosos e doentes, que frequentemente recaem sobre as mulheres, têm limitado substancialmente suas possibilidades de cumprir as demandas do trabalho remoto de forma eficaz.

2.1.3 DISPARIDADE DE GÊNERO

O impacto da pandemia de COVID-19 na produtividade acadêmica tornou-se uma área de investigação crucial em um momento em que a educação e a pesquisa estão

enfrentando desafios sem precedentes. Um estudo anterior (KAFRUNI, 2020), conduzido durante a pandemia, revelou que a suspensão das atividades presenciais teve um impacto desigual nos pesquisadores, com diferenças significativas entre homens e mulheres. Essa pesquisa, baseada em uma amostra de 10.593 alunos de pós-graduação, indicou que 81% deles enfrentaram dificuldades na produção de suas teses e dissertações durante o período de suspensão das atividades presenciais.

Entre os homens que responderam à pesquisa, 36% daqueles que não tinham filhos afirmaram estar conseguindo trabalhar de forma remota. No entanto, entre os que possuíam filhos, apenas 17% relataram o mesmo nível de produtividade. Já entre as mulheres que participaram da pesquisa, os impactos foram mais acentuados. Das mulheres que não tinham filhos, apenas 33% disseram estar conseguindo trabalhar de forma remota, enquanto aquelas que tinham filhos relataram uma taxa ainda mais baixa, com apenas 10% conseguindo manter a mesma produtividade.

2.2 MODELOS DE APRENDIZADO DE MÁQUINA

2.2.1 REDES NEURAIS PROFUNDAS (DNNs)

As Redes Neurais Profundas, também conhecidas como *Deep Learning*, são modelos de aprendizado de máquina que consistem em várias camadas (profundas) de neurônios artificiais. Elas revolucionaram o campo devido à sua capacidade de aprender representações complexas a partir de dados brutos. *DNNs* têm aplicações em visão computacional, processamento de linguagem natural, reconhecimento de fala, entre outros. A profundidade das redes permite que elas capturem características hierárquicas, tornando-as especialmente úteis em tarefas complexas (LECUN, BENGIO e HINTON, 2015).

2.2.2 REDES NEURAIS RECORRENTES (RNNs)

As Redes Neurais Recorrentes são projetadas para lidar com sequências temporais, onde a ordem e a dependência temporal dos dados são críticas. Elas são amplamente usadas em tarefas como tradução automática, previsão de séries temporais e análise de sentimentos em texto. No entanto, as *RNNs* enfrentam problemas de gradientes que dificultam o aprendizado de dependências de longo prazo. Modelos mais avançados, como as *LSTM* (*Long Short-Term Memory*) e as *GRU* (*Gated Recurrent Unit*), foram desenvolvidos para superar essas limitações (HOCHREITER e SCHMIDHUBER, 1997).

2.2.3 REDES NEURAIS CONVOLUCIONAIS (CNNs)

As Redes Neurais Convolucionais são especialmente eficazes em tarefas de visão computacional. Elas utilizam camadas de convolução para aprender padrões locais em dados, tornando-as ideais para tarefas como classificação de imagens e detecção de objetos. *CNNs* revolucionaram a área de reconhecimento de imagem e permitiram o desenvolvimento de carros autônomos, sistemas de vigilância por vídeo e muito mais (KRIZHEVSKY, SUTSKEVER e HINTON, 2012).

Apesar de inicialmente desenvolvidas para tarefas de visão computacional, também demonstraram ser eficazes no processamento de linguagem natural (NLP) (CONNEAU et al., 2016; ZHANG e WALLACE, 2015). Enquanto a aplicação tradicional das *CNNs* envolve a extração de características de imagens, elas podem ser adaptadas para extrair recursos relevantes de sequências de texto, como sentenças e documentos. Isso é feito aplicando filtros convolucionais sobre janelas deslizantes de palavras ou caracteres, permitindo que as *CNNs* capturem informações contextuais e estruturais em textos, tornando-as uma escolha viável em tarefas de classificação de texto, análise de sentimentos e muito mais. A flexibilidade das *CNNs* as torna uma opção versátil em várias áreas do aprendizado de máquina, incluindo *NLP*.

2.2.4 MODELO DE MÁQUINA DE VETORES DE SUPORTE (SVM)

O *Support Vector Machine* (*SVM*) é uma técnica de aprendizado de máquina que se destaca por sua eficácia na classificação de dados por meio de aprendizado supervisionado. Sua abordagem envolve o uso de vetores de suporte, que desempenham um papel fundamental no processo de separação e agrupamento de dados com base em suas similaridades (KOWALCZYK, 2017). Sua fundação teórica sólida e capacidade de lidar com problemas de separação linear e não linear o tornaram uma ferramenta valiosa em diversos campos.

O *SVM* foi introduzido por Vapnik e Cortes em seu artigo seminal de 1995, que estabeleceu as bases teóricas do algoritmo. Eles enfatizaram a importância de encontrar um hiperplano de decisão que maximize a margem entre classes, conforme declarado no artigo: "O objetivo do *SVM* é encontrar o hiperplano ótimo que separa os exemplos de treinamento de diferentes classes, maximizando a margem." (CORTES e VAPNIK, 1995). Os *SVMs* não apenas oferecem alto desempenho, mas também fornecem interpretabilidade. Os vetores de

suporte e o hiperplano de decisão podem oferecer *insights* valiosos sobre a importância relativa das características nos dados, como enfatizado por Hsu et al. (2003) em seu guia prático: "Os *SVMs* são eficazes e ao mesmo tempo fornecem interpretabilidade, permitindo a identificação dos principais contribuintes para as decisões." (HSU et al., 2003). Em resumo, o *SVM* continua sendo uma ferramenta essencial em aprendizado de máquina devido à sua robustez teórica, capacidade de tratar problemas não lineares e ampla aplicabilidade em diversas áreas. Suas raízes sólidas na teoria e flexibilidade prática o tornam um modelo confiável e versátil para uma variedade de tarefas de análise de dados.

2.3 MÉTRICAS DE AVALIAÇÃO DOS MODELOS

A avaliação rigorosa de modelos de inteligência computacional desempenha um papel fundamental na garantia da qualidade e eficácia de suas aplicações (STRAUSS, JÚNIOR e FERREIRA, 2022). Nesse contexto, a utilização de métricas apropriadas torna-se crucial para fornecer uma visão precisa sobre o desempenho desses modelos.

Métricas como acurácia, precisão, recall e F1-score são fundamentais para medir a precisão global do modelo e sua capacidade de distinguir corretamente entre as diferentes classes.

As métricas utilizadas para avaliar o desempenho dos modelos utilizados neste trabalho foram:

Precisão (*Precision*): Indica a proporção de instâncias positivas identificadas corretamente em relação ao total de instâncias identificadas como positivas. A fórmula é:

$$Precision = \frac{TP}{(TP + FP)}$$

Onde *TP* são os verdadeiros positivos e *FP* são os falsos positivos.

Acurácia (*Accuracy*): Representa a proporção de instâncias corretamente classificadas em relação ao total de instâncias. A fórmula é:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Onde *TN* são os verdadeiros negativos e *FN* são os falsos negativos.

Recall (Revocação ou Sensibilidade): Indica a proporção de instâncias positivas identificadas corretamente em relação ao total de instâncias que são realmente positivas. A fórmula é:

$$Recall = \frac{TP}{TP + FN}$$

F1-*Score*: É a média harmônica entre precisão e *recall*, proporcionando um equilíbrio entre essas duas métricas. A fórmula é:

$$F1\text{-}Score = \frac{Precision * Recall}{Precision + Recall}$$

Essas métricas são especialmente úteis em problemas de classificação binária, onde você está tentando distinguir entre duas classes, como positivo e negativo.

2.3 CLASSIFICAÇÃO DE GÊNERO USANDO TÉCNICAS DE INTELIGÊNCIA COMPUTACIONAL

A classificação de gênero tem sido uma área de interesse em várias pesquisas anteriores, abrangendo diferentes contextos e abordagens, como referenciado nos seguintes estudos: (KARIMI et al., 2016; SEPTIANDRI, 2017; TO et al., 2020; REGO, SILVA E FERNANDES, 2021; TRIPATHI e FARUQUI, 2011).

Tripathi (2011) empreendeu uma análise abrangente ao utilizar a técnica de *machine learning* conhecida como Máquina de Vetores de Suporte (*SVM*) para prever o gênero com base em nomes indianos. O aspecto distintivo de seu estudo foi a aplicação da abordagem de *n-gram-suffixes* em vez de considerar a palavra completa. Essa estratégia envolve a análise de partes específicas das palavras, o que pode ser particularmente eficaz para identificar nuances de gênero em nomes. Outra investigação relevante, conduzida por Septiandri (2017), concentrou-se na previsão de gênero em nomes indonésios, explorando tanto nomes completos como apenas o primeiro nome. A análise comparativa dos resultados obtidos a partir desses dois estilos de dados proporcionou *insights* interessantes sobre as variações na precisão da classificação de gênero.

No estudo de Karimi (2016), os pesquisadores adotaram uma abordagem única, analisando o efeito do conceito astrológico de Gana em nomes pessoais. Essa análise culminou na criação de duas redes *Long Short Term Memory* (*LSTM*), que foram empregadas

para a classificação de gênero. Essa pesquisa destaca a diversidade de abordagens que podem ser aplicadas na resolução desse problema. Em um cenário diferente, To (2020) voltou-se para a classificação de nomes vietnamitas, comparando o desempenho do *SVM* com uma rede *LSTM*. Os resultados revelaram que o *SVM* superou outros modelos de aprendizado de máquina, enquanto o *LSTM* se destacou em relação aos modelos de *deep learning*. Um aspecto notável deste estudo foi a análise abrangente, considerando não apenas o primeiro nome, mas também o sobrenome e o nome completo. Finalmente, Rego (2021) explorou técnicas de *machine learning* e *deep learning* para classificar gênero em nomes brasileiros. Entre as várias abordagens testadas, o modelo *BiLSTM* se destacou com uma notável acurácia de 96,17%, enquanto a *Convolutional Neural Network* (*CNN*) também obteve resultados impressionantes, atingindo 95,78%. Esses resultados evidenciam a eficácia dessas técnicas na classificação de gênero com base em nomes.

Dentro do escopo deste trabalho, nosso foco recai sobre a aplicação de modelos de *machine learning* e *deep learning* na classificação de gênero de pesquisadores brasileiros, com especial atenção ao primeiro nome como variável de entrada. A avaliação de desempenho dos modelos é conduzida por meio de métricas rigorosas, incluindo acurácia, *recall*, precisão e F1-*score*, permitindo uma análise detalhada e comparativa de suas capacidades.

2.4 EXTRAÇÃO DE DADOS DA WEB

2.4.1 WEB SCRAPING

O *web scraping* é uma técnica essencial na extração de dados relevantes da *web* para uma variedade de aplicações, como pesquisa, análise de dados e desenvolvimento de produtos. Segundo Barbosa (2020), *web scraping* pode ser definido como uma “raspagem” de dados diretamente da *web*.

Nesse contexto, esta busca de dados na literatura destaca a necessidade de examinar o impacto da pandemia de COVID-19 na produtividade acadêmica, particularmente no cenário brasileiro, sob uma perspectiva de gênero. A transição para o ensino e pesquisa remotos, as disparidades de gênero e as técnicas de classificação de gênero baseadas em *machine learning* e *deep learning* constituem elementos essenciais deste estudo, que busca entender de que maneira a pandemia influenciou a produção acadêmica no Brasil.

2.4.2 WEB CRAWLING

O *web crawling*, ou rastreamento da web, é um componente fundamental dos motores de busca e de várias outras aplicações que requerem coleta de dados na internet. Essa técnica envolve a navegação sistemática pela *web* para coletar informações de diversas páginas da *web*, criando um índice ou banco de dados acessível para busca e recuperação (D'HAEN et al., 2016). A técnica de *web crawling* é usada em conjunto com a técnica de *web scraping*. Segundo Galdino (2020), "O *bot* que utiliza a técnica de *web scraping* é programado para efetuar requisições a um servidor web a partir de uma lista predefinida de *URLs* ou retornado pela técnica *web crawling*".

3 METODOLOGIA

Este estudo se caracteriza como uma pesquisa exploratória cujo objetivo principal é analisar os efeitos da pandemia de COVID-19 na produção científica do Brasil. Especificamente, busca-se examinar como essa crise impactou a produtividade acadêmica, da perspectiva de gênero. Para alcançar esse propósito, utilizamos técnicas de inteligência computacional para automatizar a identificação do gênero dos autores de artigos acadêmicos e construímos um *crawler* para coletar os dados de publicações acadêmicas disponíveis na plataforma Lattes. Além disso, concentramos nossa análise na investigação das mudanças e consequências que surgiram a partir da transição para o trabalho remoto.

3.1 COLETA DOS DADOS

A coleta de dados para este estudo representou uma etapa fundamental na obtenção de informações essenciais para a análise do impacto da pandemia de COVID-19 na produtividade científica brasileira, com foco na perspectiva de gênero. Esta pesquisa exploratória buscou compreender como a crise sanitária afetou a produção acadêmica do Brasil e como isso se refletiu em termos de gênero. Para atingir esse objetivo, optamos por utilizar a Plataforma Lattes, uma escolha amplamente justificada pelo seu uso generalizado entre pesquisadores de diversas áreas no Brasil, tornando-a uma fonte rica e representativa de dados acadêmicos.

O processo de coleta de dados começou com uma análise cuidadosa do funcionamento da Plataforma Lattes. Durante essa fase, examinamos detalhadamente a estrutura da plataforma, os mecanismos de pesquisa de currículos e a forma de obtenção das informações relevantes. Um ponto importante a ser destacado é que, devido à maneira como a plataforma

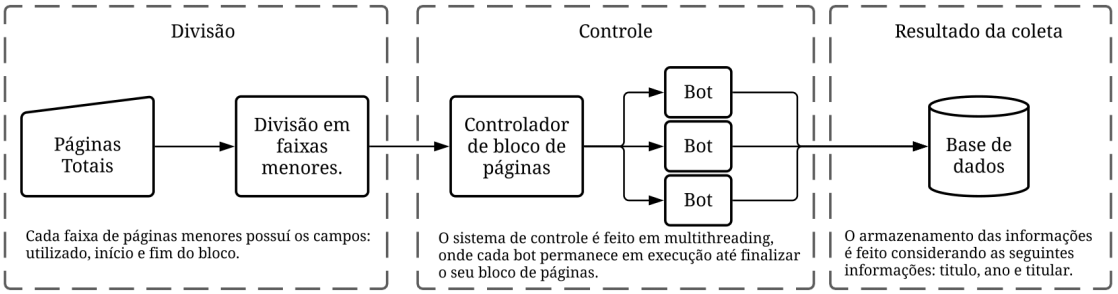
foi desenvolvida, nossa capacidade de coleta se limitou aos "artigos completos publicados em periódicos". Assim, concentramos nossa análise nessa categoria de publicações e extraímos informações cruciais, incluindo a data de publicação dos artigos (compreendendo o período de 2018 a 2021), os nomes dos autores, os títulos dos artigos, os níveis acadêmicos e as nacionalidades dos pesquisadores.

A coleta de dados, embora essencial, enfrentou desafios devido ao grande volume de currículos disponíveis na plataforma. O processo inicialmente se mostrou demorado, com uma média de 10 segundos gastos em cada currículo. Para otimizar essa tarefa, optamos por utilizar vários *bots* de coleta de dados de forma simultânea em um único computador. Essa abordagem permitiu que os *bots* operassem de maneira paralela, resultando em uma redução substancial no tempo de coleta, com uma média de apenas 4 segundos por currículo quando utilizamos 10 execuções simultâneas.

A coleta de dados foi organizada de forma sistemática. Iniciamos com os currículos de pesquisadores doutores e, em seguida, procedemos à coleta dos currículos de pesquisadores mestres. Além disso, distribuímos as páginas de coleta entre computadores disponíveis, cada um com suas próprias configurações, como AMD Athlon-3000G, Intel i3-7100u, Intel i3-8130U e Intel Celeron-3865u. Isso possibilitou a análise de um total impressionante de 98.907 currículos em um período de aproximadamente 36 horas.

O processo de coleta de dados foi subdividido em três etapas distintas: divisão, controle e resultado, conforme ilustrado na Figura 1. Na etapa de divisão, segmentamos as páginas a serem coletadas em blocos de tamanho pré-definido antes da execução. O estágio de controle envolveu a atribuição de cada bloco a um *crawler* para execução paralela, gerenciando de maneira eficiente a coleta dos dados. Após a conclusão da coleta de um bloco, realizamos uma verificação para determinar se havia mais segmentos disponíveis; caso contrário, o *bot* era encerrado. A fase final, denominada resultado, englobou o armazenamento dos dados coletados em um banco de dados, possibilitando a análise posterior. Esse processo meticuloso de coleta de dados é fundamental para a credibilidade e precisão das conclusões deste estudo.

Figura 1 - Diagrama de coleta de dados



Fonte: Autoria própria (2023).

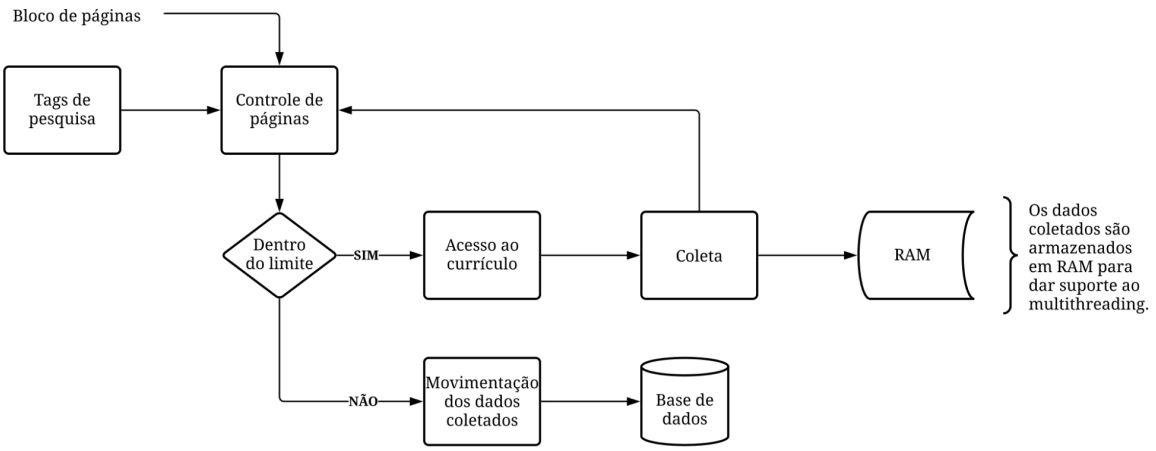
3.1.1 *BOT* DA COLETA DE DADOS

O *bot* é um software especialmente desenvolvido para a exploração da plataforma Lattes e a obtenção das informações necessárias sobre pesquisadores brasileiros. Dado que não existe uma base de dados própria com as informações necessárias para a análise dos impactos da COVID-19 na produção de artigos científicos, a utilização desse *bot* torna-se essencial para coletar e criar tal base.

Conforme ilustrado na Figura 2, o *bot* é iniciado com as tags de pesquisa e o conjunto de páginas a serem coletadas, permitindo o acesso aos currículos dos pesquisadores. A partir disso, ele analisa minuciosamente as informações presentes nas páginas, verificando se há registros de "artigos completos publicados em periódicos" dentro dos anos definidos para a pesquisa. Caso essa informação seja encontrada, ela é armazenada para análise posterior.

O resultado da coleta na Plataforma Lattes revelou um total de 455.794 trabalhos acadêmicos publicados no período de 2018 a 2021, distribuídos da seguinte forma: 114.030 em 2018, 117.246 em 2019, 138.136 em 2020 e 86.382 em 2021. Vale destacar que, ao desconsiderar os casos excepcionais, ou seja, as pessoas com produções muito acima do intervalo interquartílico, a média de trabalhos por pesquisador foi de aproximadamente 6 trabalhos. Esse dado é essencial para compreender a distribuição geral da produção científica no período de estudo.

Figura 2 - Diagrama de funcionamento do *bot*



Fonte: Autoria própria (2023).

3.1.2 TRATAMENTO DOS DADOS

Algoritmos de aprendizado de máquina necessitam de dados numéricos para a realização de seu treinamento e classificação. Com isso, a base de dados utilizada (que consiste de dados textuais) necessita passar por um processo de codificação. Inicialmente, é necessária a realização do tratamento dos dados, removendo possíveis acentos ou caracteres especiais. Em seguida, é realizada a etapa de codificação dos dados, onde primeiramente é criado um glossário com cada letra presente nos dados e, a partir disso, é desenvolvido um vetor com o mesmo tamanho do glossário, em que cada posição representa uma letra da palavra e é definido como 1, caso não exista tal letra no nome a posição é definida como 0, conforme mostrado na Figura 3. Esta técnica de codificação é chamada de *one-hot encoding*.

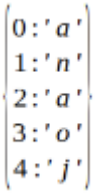
Figura 3 - Exemplo *one-hot encoding*

j :	0	0	0	0	0	0	0	0	1	0	0	0	0	0	...	0
o :	0	0	0	0	0	0	0	0	0	0	0	0	0	1	...	0
a :	1	0	0	0	0	0	0	0	0	0	0	0	0	0	...	0
n :	0	0	0	0	0	0	0	0	0	0	0	0	1	0	...	0
a :	1	0	0	0	0	0	0	0	0	0	0	0	0	0	...	0

Fonte: Autoria própria (2023).

O *one-hot encoding* desempenhou um papel crucial na codificação dos nomes utilizados durante o treinamento e classificação dos modelos *BiLSTM* e *CNN*. No entanto, quando se trata da aplicação dos dados no *SVM*, é necessária a adoção de uma técnica diferente, uma vez que o *one-hot encoding* gera um vetor de valores, enquanto o *SVM* requer a entrada de valores unitários. Para abordar essa diferença, recorremos à biblioteca *Natural Language Toolkit* (*NLTK*), que oferece um módulo específico para lidar com modelos de *SVM*. Para usar a *NLTK*, começamos criando um dicionário contendo as informações relevantes para cada dado. Neste contexto, a informação é representada pela sequência inversa das letras do nome, limitada a 10 letras, conforme mostrado na Figura 4. Posteriormente, associamos esse dicionário aos valores de classificação, no caso, a base de dados usada para treinar o modelo.

Figura 4 - *NLTK*



Fonte: Autoria própria (2023).

A codificação das classes de saída, ou seja, as categorias de gênero (feminino/masculino), segue um processo distinto. Dado que estamos lidando com uma classificação binária, onde existem apenas dois valores possíveis, simplesmente atribuímos a cada classe de classificação os valores 0 ou 1, conforme apropriado. Uma vez concluída a codificação dos dados, procedemos à divisão de acordo com as necessidades específicas de cada modelo. Esse processo de pré-processamento é fundamental para garantir que os dados estejam em um formato compatível com os requisitos de cada algoritmo de aprendizado de máquina.

3.2 MODELOS DE APRENDIZADO DE MÁQUINA

Nesta seção, dedicaremos nosso foco à apresentação detalhada das técnicas fundamentais que desempenharam um papel crucial na classificação de gênero dos dados que foram coletados. Esta seção proporcionará uma compreensão abrangente das técnicas

utilizadas em nossa análise de classificação de gênero, permitindo uma visão clara de como cada uma delas foi aplicada para alcançar nossos objetivos de pesquisa.

3.2.1 MÁQUINA DE VETOR DE SUPORTE (SVM)

No contexto do *SVM*, uma questão crucial é a identificação do hiperplano que melhor divide as classes de dados. A determinação desse hiperplano, que define as margens de separação entre as classes, requer a escolha criteriosa de um *kernel* apropriado. Quando os dados de classificação exibem padrões que podem ser linearmente separados, um *kernel* linear é suficiente para essa tarefa. No entanto, em situações em que os padrões são não lineares e complexos, torna-se imperativo utilizar um *kernel* não linear, uma vez que problemas não lineares frequentemente apresentam uma maior probabilidade de serem tornados linearmente separáveis em um espaço de alta dimensão (COVER, 1965).

No âmbito da análise do problema de classificação textual de gênero, foram conduzidas simulações envolvendo diferentes funções *kernels* para determinar qual se adequa melhor às características dos dados. Os resultados dessas simulações, que incluem métricas de avaliação como acurácia e F1-*score*, para cada modelo SVM com seu respectivo *kernel*, são detalhadamente apresentados na Tabela 1. Esses resultados desempenham um papel essencial na escolha da estratégia de classificação mais eficaz para o contexto específico da tarefa em questão.

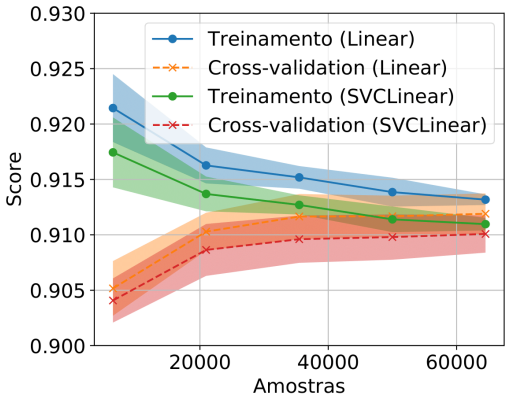
Tabela 1 - Acurácia dos modelos SVM

Acurácia (%)	**F1-*Score***	***Kernel***
95,43	0,9493	*NuRBF*
95,07	0,9455	*RBF*
91,32	0,9047	*Linear*
91,31	0,9047	*Poly*
91,10	0,9017	*LinearSVC*
88,60	0,8740	*NuPoly*
88,45	0,8736	*Sigmoid*
86,59	0,8439	*NuSigmoid*

Fonte: Autoria própria (2023)

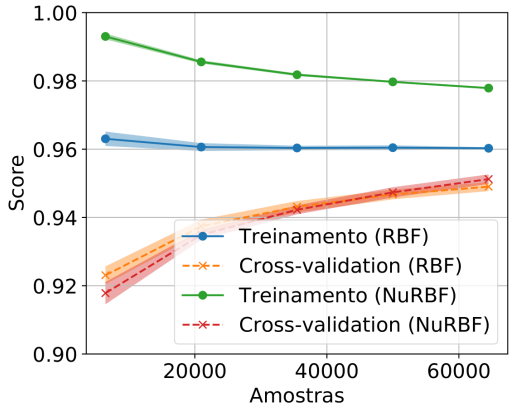
Dentre todos os modelos de classificação avaliados, aquele que demonstrou a maior precisão de classificação foi o modelo com o *kernel NuRBF*. Além disso, para uma análise mais abrangente dos modelos com diferentes *kernels*, apresentamos as curvas de aprendizado nas Figuras 5, 6 e 7.

Figura 5 - Curvas de aprendizagem *SVM* com *kernels* linear e *SVC*-linear



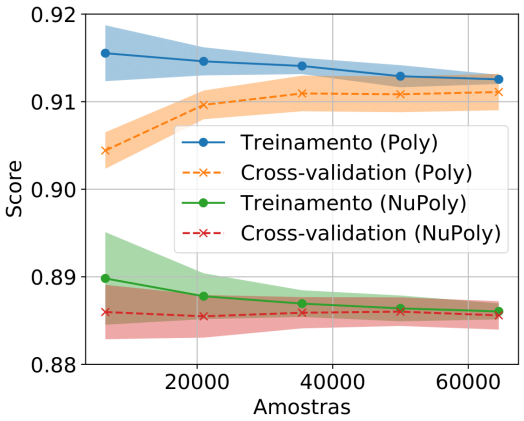
Fonte: Autoria própria (2023)

Figura 6 - Curvas de aprendizagem *SVM* com *kernels* RBF e NuRBF



Fonte: Autoria própria (2023)

Figura 7 - Curvas de aprendizagem *SVM* com *kernels* Poly e NuPoly



Fonte: Autoria própria (2023)

As curvas de aprendizado retratam a relação entre a pontuação (*score*) do treinamento e a pontuação da validação cruzada (*cross-validation*) em função do número variável de amostras de treinamento. Observa-se que os modelos representados nas Figuras 5 e 7 apresentam desafios relacionados tanto ao erro devido ao viés (bias) quanto ao erro devido à variância. Além disso, à medida que mais dados de treinamento são incorporados, as pontuações de treinamento e validação cruzada tendem a convergir, sugerindo que esses modelos podem não se beneficiar significativamente com a adição de mais dados.

No entanto, na Figura 6, observamos um cenário diferente, onde os modelos exibem baixa variância e baixo viés. Além disso, à medida que o número de amostras aumenta, o desempenho do modelo com o *kernel NuRBF* melhora notavelmente. Portanto, com base nessas análises, concluímos que o modelo mais adequado para a tarefa de classificação de gênero é o modelo que utiliza o *kernel NuRBF*. Ele se destaca por sua capacidade de generalização e melhoria contínua de desempenho à medida que mais dados de treinamento são disponibilizados.

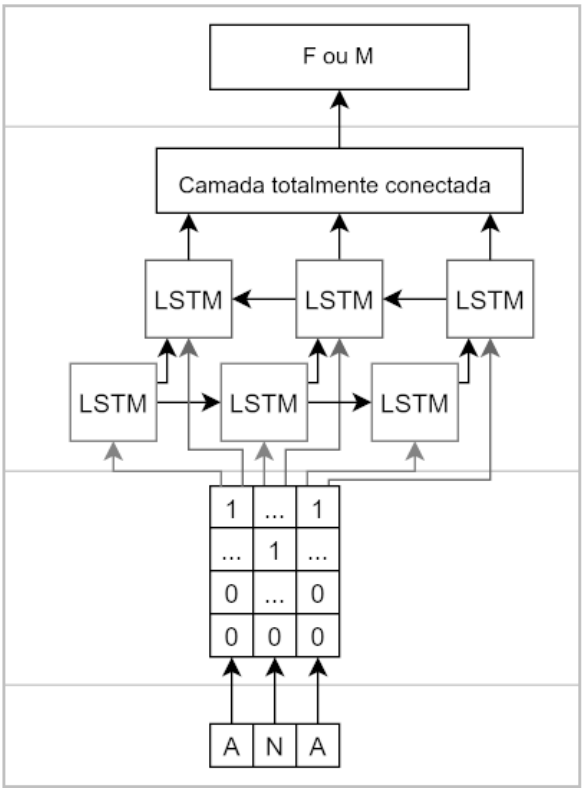
3.2.2 BIDIRECTIONAL LONG SHORT-TERM MEMORY (BiLSTM)

A arquitetura conhecida como *Long Short Term Memory* (*LSTM*) é um tipo de rede neural recorrente (*Recurrent Neural Network - RNN*) criada com o propósito de superar desafios encontrados na *RNN* convencional, tais como o desaparecimento ou explosão do

gradiente (GOODFELLOW, BENGIO e COURVILLE, 2016; JELODAR et al., 2020). Uma *LSTM* bidirecional, ou *BiLSTM*, é uma variação que processa a sequência de dados tanto no sentido direto quanto no inverso, o que a diferencia da *LSTM* tradicional (GOODFELLOW, BENGIO e COURVILLE, 2016).

Neste trabalho, a implementação do modelo *BiLSTM* foi realizada com base em um modelo base previamente desenvolvido por Rego (2021) e também com a consideração de informações da literatura. Isso resultou em um modelo composto por: uma camada *BiLSTM* com 20 neurônios de entrada e 128 neurônios de saída, uma camada de saída com 128 neurônios de entrada e 1 neurônio de saída, seguida de uma camada de ativação utilizando a função Sigmoid. Seu funcionamento é apresentado na Figura 8, onde, inicialmente a informação é codificada através da técnica *one-hot encoding* para que, posteriormente, seja inserida nas células do *LSTM* em um sentido da informação e, em seguida, inseridos nas células do *LSTM* com o sentido oposto, juntamente com a informação de saída do *LSTM* anterior, resultando no *BiLSTM* e permitindo a classificação da informação desejada.

Figura 8 - Arquitetura da rede *BiLSTM*



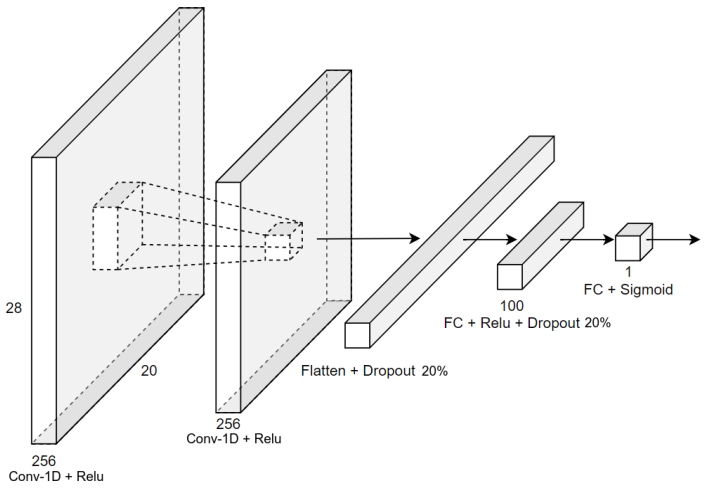
Fonte: Autoria própria (2023)

3.2.3 REDE NEURAL CONVOLUCIONAL UNIDIMENSIONAL (1D-CNN)

As Redes Neurais Convolucionais, mais conhecidas como *CNNs*, são modelos de *deep learning* amplamente utilizados para tarefas de classificação de imagens (LOPEZ e KALITA, 2017). No entanto, é notável que, como demonstrado em trabalhos anteriores (ZHANG e WALLACE, 2015; REGO, SILVA E FERNANDES, 2021; LOPEZ e KALITA, 2017), essas redes são igualmente aplicáveis a problemas de processamento de linguagem natural (*Natural Language Processing - NLP*). Para abordar desafios em *NLP*, existem diversas estratégias, dentre as quais se destaca a utilização de vetores de palavras pré-treinados (CHEN, 2015). Outra abordagem eficaz é a aplicação de redes neurais convolucionais em nível de caracteres (ZHANG e WALLACE, 2015; REGO, SILVA E FERNANDES, 2021). Nesse contexto, aproveitando a versatilidade das *CNNs* em tarefas de *NLP*, optamos por implementar uma arquitetura unidimensional para a classificação de gênero.

A arquitetura proposta, ilustrada na Figura 9, compreende duas camadas convolucionais unidimensionais com a função de ativação *ReLu* (unidade linear retificada). Em seguida, há uma camada que realiza o processo de *flatten*, convertendo os dados para uma dimensão unidimensional, e aplicando um *dropout* de 20% para regularização. Além disso, uma camada totalmente conectada com função de ativação *ReLu* e *dropout* de 20% é incorporada. Por fim, a arquitetura culmina com uma última camada totalmente conectada que utiliza a função de ativação sigmóide. Esse conjunto de camadas compõe a rede neural convolucional unidimensional projetada para a tarefa de classificação de gênero.

Figura 9 - Arquitetura da CNN



Fonte: Autoria própria (2023)

3.3 TREINAMENTO DOS MODELOS

O processo de treinamento dos modelos teve início com a utilização da base de dados disponibilizada pelo brasil.io, a qual pode ser acessada através do link (https://brasil.io/dataset/genero-nomes/nomes). Essa base de dados é composta por 100787 nomes brasileiros, obtidos a partir do CENSO de 2010, e apresenta uma distribuição onde 54.82% dos nomes são femininos, enquanto 45.18% são masculinos. Além dos modelos *SVM, BiLSTM e CNN*, também foram implementados outros classificadores, incluindo Extra Trees, Decision Tree, Algoritmo K-Nearest Neighbors (Knn), Naive Bayes, Random Forest, Gradient Boosting, Light Gradient Boosting, Logistic Regression, Classificador Ridge e Ada Boost.

Para a preparação dos dados utilizados nos modelos de *machine learning*, foi realizado um embaralhamento dos registros e a subsequente divisão aleatória em subconjuntos, com 80% dos dados destinados ao treinamento e 20% aos testes. Importante destacar que, durante esse processo, a proporção das classes foi mantida, preservando a frequência de rótulos das classes feminina e masculina. Para os modelos de *deep learning*, que compreendem *BiLSTM* e *CNN*, foi necessária a definição de uma função de perda para avaliar o desempenho durante o treinamento. Dado que o problema em questão é de classificação binária (1 ou 0, M ou F), optou-se pela função de perda de entropia cruzada binária, que é apropriada para essa categoria de tarefa. Portanto, a função de perda utilizada é expressa da seguinte forma:

$$L(p, q) = -\frac{1}{N}\left[\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(q(y_i))\right]$$

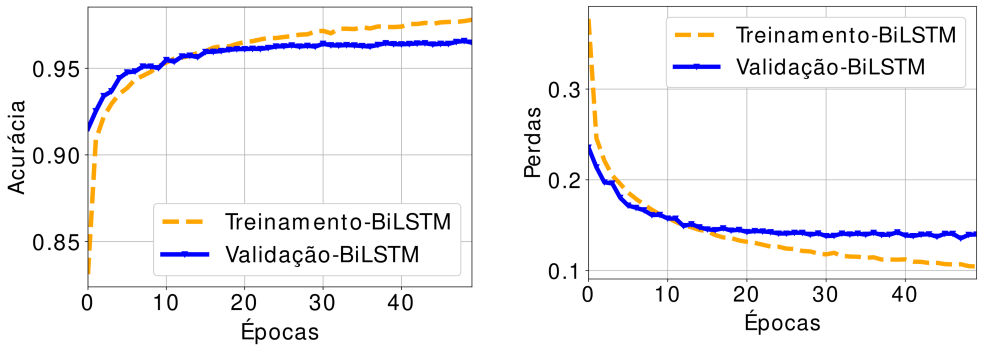
Nessa equação, "y" representa o rótulo, sendo 0 para indicar feminino e 1 para masculino. A variável "p(y)" representa a probabilidade prevista de que o gênero seja masculino para todos os "N" pontos considerados. Por sua vez, "$q(y_i)$" é a probabilidade prevista de que o gênero seja feminino, e é definida como o complemento de "$p(y_i)$", ou seja, "$q(y_i) = 1 - p(y_i)$".

Durante o processo de treinamento de algoritmos de *deep learning*, é altamente recomendável a alocação de uma parcela dos dados para a validação do modelo. Esse procedimento permite que o modelo efetue ajustes em suas previsões à medida que avança no processo de treinamento, reduzindo a probabilidade de *overfitting* e melhorando a capacidade

de generalização das classificações (HAWKINS, 2004). Nesse contexto, o conjunto de dados foi submetido a um processo de embaralhamento e dividido aleatoriamente em subconjuntos, sendo alocados 60% dos dados para treinamento, 20% para validação e 20% para teste. Essa divisão cuidadosa visa garantir que o modelo seja treinado, validado e testado de forma eficiente, equilibrando o aprendizado com a avaliação. Além disso, para mitigar os riscos de *overfitting*, aplicamos a técnica conhecida como *early stopping* (GOODFELLOW, BENGIO e COURVILLE, 2016). Isso significa que o treinamento é interrompido mais cedo caso sejam observados sinais de que o modelo está super ajustando os dados de treinamento.

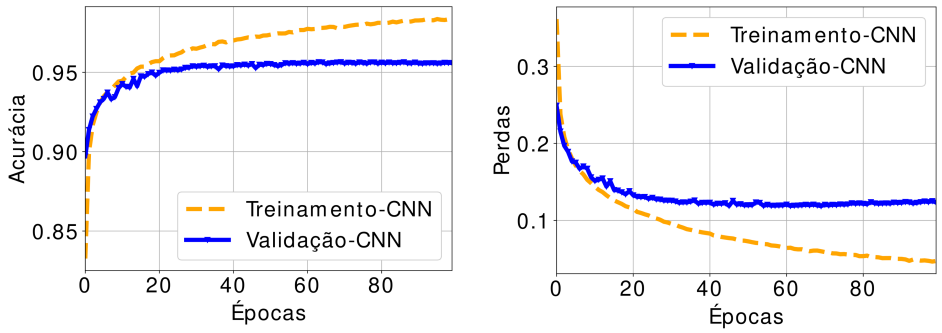
As Figuras 10 (referentes à *BiLSTM*) e 11 (referentes à *CNN*) apresentam as curvas de acurácia e perda dos modelos em relação às épocas de treinamento, proporcionando uma visão clara do desempenho e aprendizado dos modelos ao longo do processo de treinamento.

Figura 10 - Acurácia e perdas em relação às épocas (BiLSTM)



Fonte: Autoria própria (2023)

Figura 11 - Acurácia e perdas em relação as épocas (CNN)



Fonte: Autoria própria (2023)

A Tabela 2 fornece uma visão abrangente das métricas de desempenho, incluindo acurácia, precisão, recall e F1-*Score*, referentes a cada um dos modelos implementados. Essas métricas são cruciais para avaliar completamente a eficácia dos modelos treinados neste estudo. Entre todos os modelos avaliados, o *BiLSTM* se destaca como aquele que registra os melhores valores em todas as métricas consideradas. Em outras palavras, o *BiLSTM* demonstrou a maior capacidade de classificar corretamente o gênero dos pesquisadores. Além disso, os modelos *SVM* (com *kernel NuRBF*) e *CNN* também apresentaram resultados positivos.

Vale ressaltar que o *Naive Bayes* se destacou por uma menor sensibilidade ou *recall*, ou seja, ele mostrou uma menor habilidade em identificar todas as situações de classe verdadeira como valor esperado. No entanto, cada modelo tem suas vantagens e limitações, e o desempenho deve ser avaliado com base nos objetivos específicos da tarefa. Dessa forma, com base nos resultados obtidos, optamos por utilizar as técnicas *BiLSTM*, *CNN* e *SVM*, uma vez que elas demonstraram as melhores métricas para classificar o gênero dos nomes brasileiros coletados. Esses modelos foram empregados para inferir a produtividade científica brasileira no período de 2018 a 2021 com base nas análises realizadas.

Tabela 2 - Métricas de desempenho dos modelos

Modelo	**Acurácia**	***Recall***	**Precisão**	**F1-*score***
Extra Trees	0,9482	0,9351	0,9498	0,9424
Random Forest	0,9460	0,9311	0,9487	0,9398
LightGBM	0,9222	0,9129	0,9152	0,9140
Decision Tree	0,9210	0,9114	0,9139	0,9126
KNN	0,9034	0,8649	0,9171	0,8902
Logistic Regression	0,8672	0,8279	0,8725	0,8496
Ridge Classifier	0,8604	0,7946	0,8855	0,8375
Gradient Boosting	0,8339	0,6864	0,9283	0,7891
Ada Boost	0,8263	0,7335	0,8629	0,7927
Naive Bayes	0,7076	0,3715	0,9559	0,5350

MLP	0,8698	0,8444	0,8642	0,8492
RNN	0,9400	0,9336	0,9335	0,9320
GRU	0,9500	0,9452	0,9442	0,9425
SVM (NuRBF)	0,9543[1]	0,9571[1]	0,9597[1]	0,9584[1]
BiLSTM	0,9652[1]	0,9612[1]	0,9618[1]	0,9614[1]
CNN	0,9558[1]	0,9529[1]	0,9508[1]	0,9518[1]

Fonte: Autoria própria (2023)

Após uma análise detalhada, observou-se que os nomes que geram mais erros nos classificadores são aqueles que possuem um *ratio* inferior a 1. O *ratio* representa a relação entre a frequência com que um nome é associado a homens ou mulheres (sendo que a categoria com a maior frequência determina o gênero do nome) e a soma total das frequências do nome para homens e mulheres. A Tabela 3 apresenta alguns exemplos desses nomes para uma melhor compreensão.

Tabela 3 - Nomes classificados de forma errada pelos modelos

Nome	**Freq. Feminino**	**Freq. Masculino**	***Ratio***	**Gênero (CENSO)**	**Gênero (Classificadores)**
Leovir	88	61	0,59	F	M
Jenair	196	161	0,55	F	M
Rutinei	59	51	0,54	F	M
Ildair	85	53	0,62	F	M

Fonte: Autoria própria (2023)

4 RESULTADOS

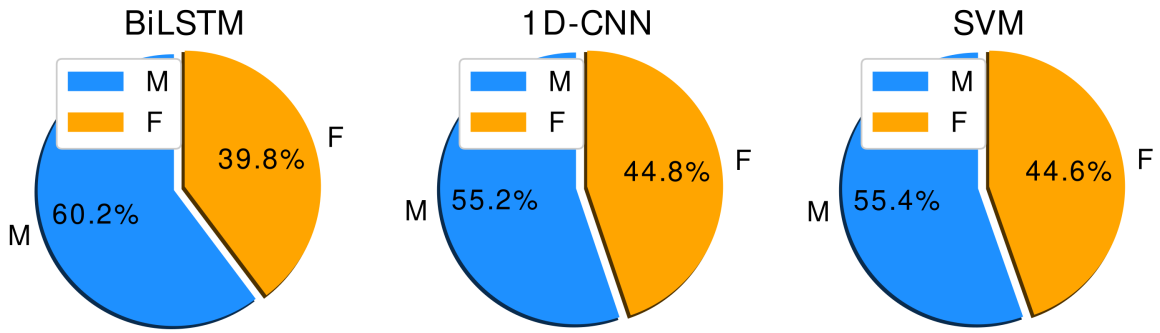
Após o treinamento e teste dos modelos, procedeu-se a uma segunda etapa de processamento, que envolveu a classificação dos dados provenientes da Plataforma Lattes. As Figuras 12 (a), (b) e (c) exibem os resultados obtidos por meio dos diversos modelos para cada gênero, no período compreendido entre 2018 e 2021.

[1] Melhores resultados

Os resultados apresentados indicam que os modelos *1D-CNN* e *SVM* produziram resultados bastante similares, com uma diferença de apenas 0.2% na classificação, conforme evidenciado nas Figuras (b) e (c), respectivamente.

Essa discrepância se deve principalmente ao fato de que os modelos *1D-CNN* e *SVM* exibem métricas bastante semelhantes, conforme destacado na Tabela 2. No entanto, a classificação de gênero realizada pelo modelo *BiLSM* demonstrou uma diferença mais significativa entre os sexos dos pesquisadores, com 60.2% identificados como do sexo masculino e apenas 39.8% como do sexo feminino, durante o período de 2018 a 2021, como ilustrado na Figura (a).

Figura 12 - Classificação de gênero entre 2018 e 2021 (a) usando o modelo BiLSTM (b) usando o modelo 1D-CNN (c) Usando o modelo SVM



(a) (b) (c)

Fonte: Autoria própria (2023)

Esses dados ressaltam de maneira clara a disparidade de gênero que está presente na produção científica.

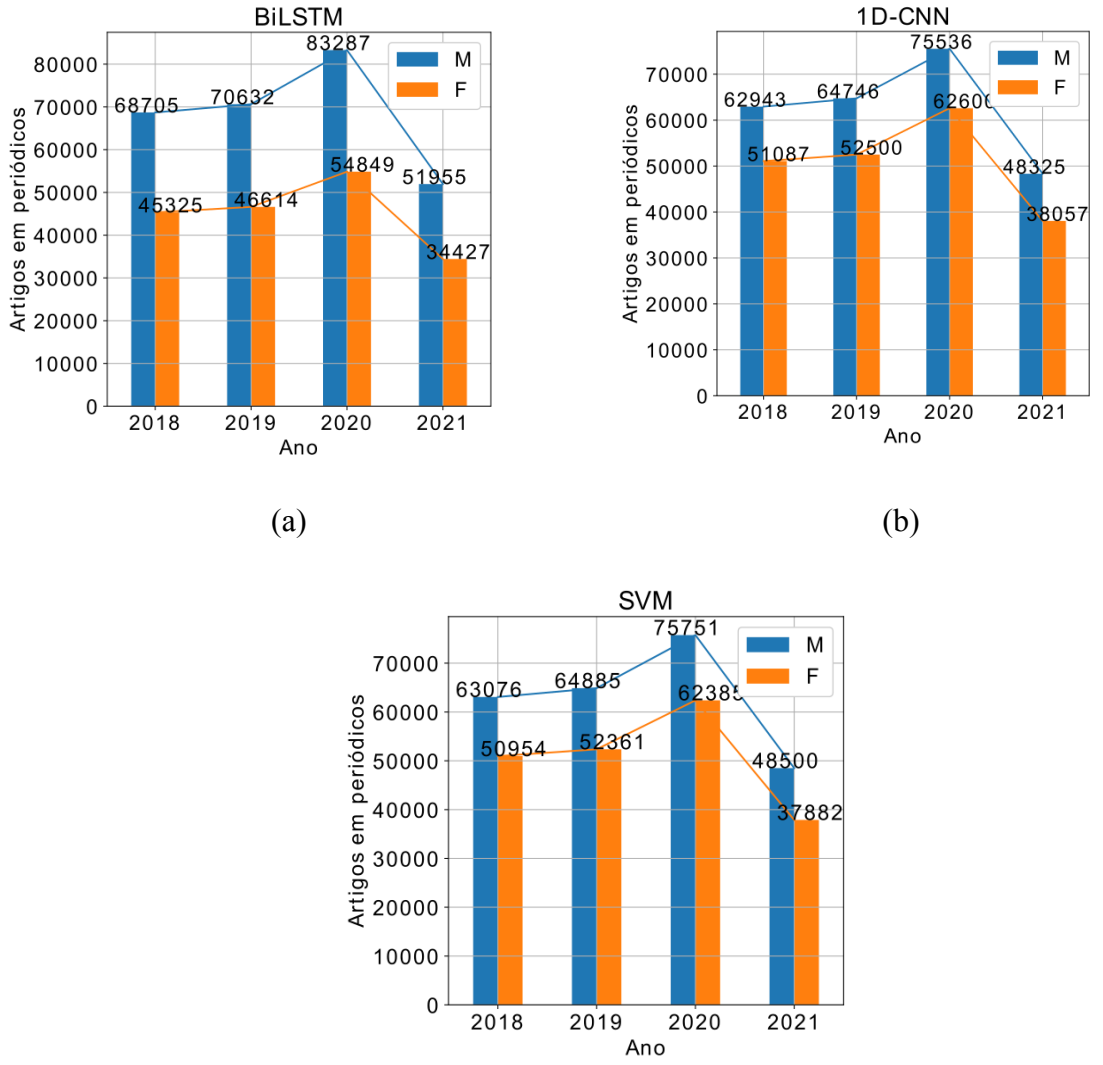
Uma análise mais aprofundada das publicações de artigos no período de 2018 a 2021 é apresentada nas Figuras 13 (a), (b) e (c). Como pode ser observado nessas figuras, o número de trabalhos científicos estava em constante crescimento até o ano de 2020, independente do gênero dos autores. No entanto, a partir de 26 de outubro de 2021, houve uma acentuada queda nesses números, como evidenciado pelos gráficos.

Em termos percentuais, entre 2020 e 2021, ocorreu uma significativa redução de 37.47% no volume de trabalhos científicos publicados por pesquisadores de ambos os sexos, conforme indicado pelos dados classificados com o *BiLSTM*, como evidenciado na Figura 11

(a). Em relação à produção de acordo com o gênero, a queda foi de 37.62% no número de trabalhos publicados por pesquisadores e de 37.23% no caso das pesquisadoras.

Analisando os resultados obtidos com a *CNN*, conforme mostrado na Figura 13 (b), observa-se uma diminuição de 36.02% no número de trabalhos científicos publicados por pesquisadores e uma queda de 39.20% na produção de pesquisadoras durante o mesmo período. Quanto aos dados classificados com o *SVM*, como ilustrado na Figura 13 (c), houve uma diminuição de 35.97% no número de trabalhos científicos publicados por pesquisadores e uma queda de 39.27% na produção de pesquisadoras de 2020 para 2021.

Figura 13 - Produtividade científica na perspectiva de gênero: (a) usando o modelo BiLSTM (b) usando o modelo 1D-CNN (c) Usando o modelo SVM



Fonte: Autoria própria (2023)

Através da análise da Tabela 4, podemos examinar a discrepância na produção acadêmica entre homens e mulheres. Como evidenciado nas Figuras 12 (a), (b) e (c), os homens apresentam consistentemente uma produção significativamente superior. Isso é refletido em diferenças sempre positivas. Por exemplo, com base nos resultados do *BiLSTM*, em 2020, os homens publicaram um número 20.59% maior de trabalhos em comparação com as mulheres.

Quando consideramos os modelos *CNN* e *SVM*, a maior disparidade ocorreu no segundo ano da pandemia, ou seja, em 2021, enquanto a menor diferença foi observada em 2020. No entanto, no caso do *BiLSTM*, a maior diferença ocorreu no primeiro ano da pandemia, com a menor sendo registrada no segundo ano, em 2021. Isso indica uma variação nas tendências de produção ao longo dos anos, com diferentes modelos mostrando diferentes pontos de destaque na disparidade de gênero.

Tabela 4 - Diferença na produção por gênero masculino sobre o feminino

Modelo	2018	2019	2020	2021	Média (%)	Desvio Padrão
BiLSTM	20,50	20,49	20,59	20,29	20,47	0,13
CNN	10,40	10,44	9,36	11,89	10,52	1,03
SVM	10,63	10,68	9,68	12,29	10,82	1,07

Fonte: Autoria própria (2023)

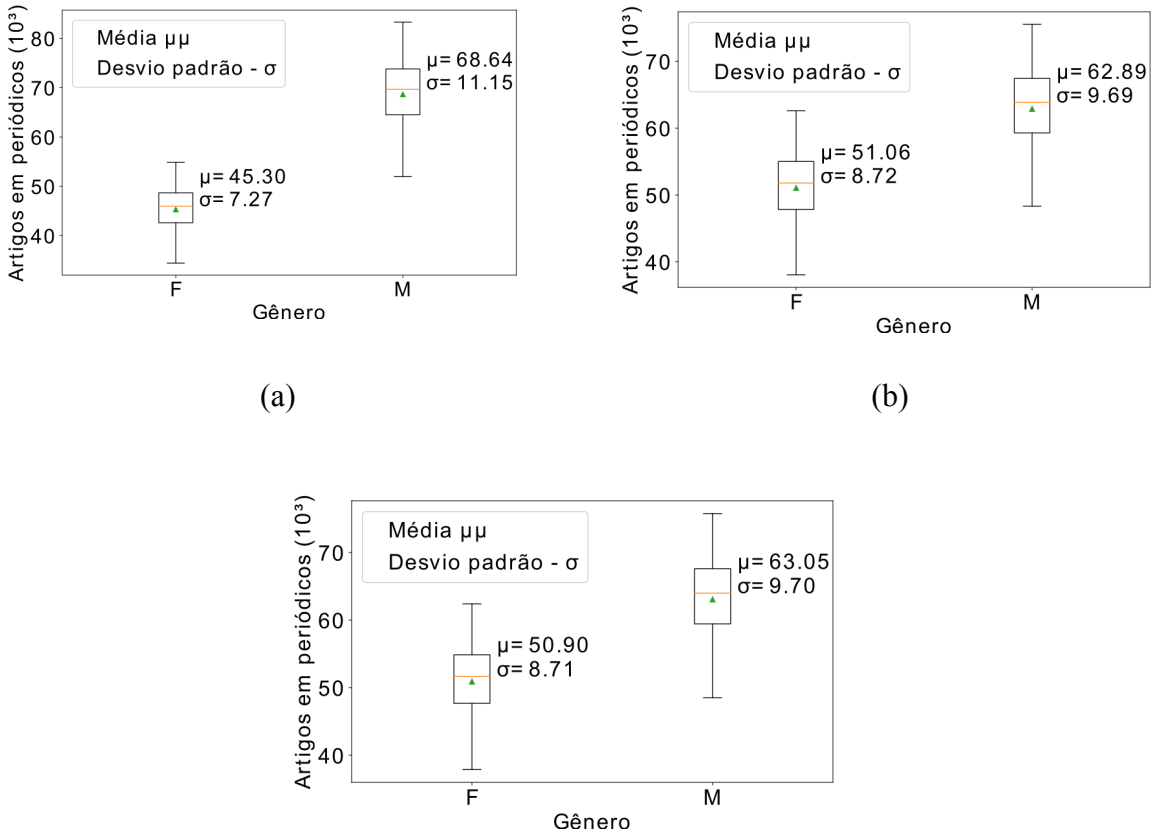
Uma análise mais aprofundada da distribuição dos dados, classificados pelos diferentes modelos (*BiLSTM, 1D-CNN* e *SVM*), pode ser realizada por meio dos *boxplots* comparativos, que são apresentados nas Figuras 12 (a), (b) e (c). Esses *boxplots* fornecem uma visão abrangente da relação entre os artigos publicados em periódicos e os gêneros ao longo dos últimos quatro anos.

Ao examinarmos as Figuras 14 (a), (b) e (c), podemos notar que os dados exibem uma distribuição simétrica para ambos os gêneros. Além disso, é possível concluir que os artigos cujos autores são do gênero masculino apresentam uma maior variabilidade em comparação com os artigos do gênero feminino. A dispersão entre os gêneros é mais pronunciada quando se considera a classificação obtida por meio do modelo de *deep learning BiLSTM*, como mostrado na Figura 12 (a).

Por outro lado, as classificações obtidas com os modelos *1D-CNN* e *SVM* revelam uma menor dispersão entre os gêneros, como pode ser observado nas Figuras 14 (b) e (c), respectivamente. Essa menor dispersão sugere uma maior consistência na produção acadêmica entre os gêneros.

Essas dispersões destacam a complexa dualidade enfrentada pelas mulheres, equilibrando as responsabilidades domésticas e profissionais, conforme confirmado pela pesquisa de Parents de Neumann (2020).

Figura 14 - Produtividade científica na perspectiva de gênero: (a) usando o modelo BiLSTM (b) usando o modelo 1D-CNN (c) Usando o modelo SVM



Fonte: Autoria própria (2023).

5 CONSIDERAÇÕES FINAIS

Este estudo abordou os impactos da pandemia de COVID-19 na produção de trabalhos científicos no contexto brasileiro. A pesquisa começou com a coleta dos dados necessários e,

posteriormente, desenvolveu e analisou modelos de machine learning, alcançando uma precisão superior a 95% na classificação dos dados. Com base nisso, prosseguimos com a análise dos resultados obtidos em cada modelo. Observamos que, entre os anos de 2018 e 2020, houve um aumento significativo no número de publicações científicas, independentemente do gênero dos autores, atingindo o pico em 2020. No entanto, em 2021, ocorreu uma queda notável, levando a níveis inferiores aos de 2018 para ambos os gêneros.

A análise das porcentagens de queda em 2021 revelou uma redução mais acentuada na produção acadêmica das pesquisadoras quando os dados foram classificados pelo modelo *CNN* e pelo *SVM*. Por outro lado, quando os dados foram classificados pelo modelo *BiLSTM*, a queda foi mais acentuada entre os pesquisadores do gênero masculino. É importante ressaltar que a plataforma Lattes é alimentada diretamente pelos pesquisadores, o que implica a possibilidade de dados mais recentes ainda não terem sido atualizados. Portanto, as estatísticas apresentadas refletem a produtividade científica no período de 2018 a 2021, com base nos dados disponíveis na plataforma Lattes.

Essa análise fornece *insights* valiosos sobre as tendências de produção acadêmica durante o período da pandemia, destacando os desafios enfrentados por pesquisadores de ambos os gêneros e a importância de considerar esses aspectos na avaliação da produtividade científica no Brasil. Além disso, ressalta a necessidade de políticas e medidas que promovam a equidade de gênero na pesquisa científica.

Considerando as nuances reveladas por esta pesquisa, surgem oportunidades significativas para estudos futuros, oferecendo uma visão mais abrangente e aprofundada:

- A pesquisa pode ser expandida além do limite temporal estabelecido (até 2021) para capturar os desenvolvimentos pós-pandêmicos, permitindo uma compreensão mais completa das tendências de produção científica.
- Uma abordagem mais holística pode ser adotada ao incorporar variáveis externas, como fatores econômicos, sociais e políticas públicas, para uma análise mais contextualizada dos impactos na produção acadêmica.
- Além das análises quantitativas, uma pesquisa qualitativa pode ser empreendida para capturar as experiências pessoais dos pesquisadores, fornecendo insights valiosos sobre os desafios enfrentados e possíveis soluções.

- Explorar outras fontes de dados além da Plataforma Lattes, como plataformas acadêmicas adicionais e fontes complementares, pode enriquecer a compreensão da produção científica.

Essas sugestões de trabalhos futuros buscam não apenas expandir o conhecimento existente, mas também oferecer orientações práticas para moldar políticas e práticas futuras, promovendo uma comunidade científica mais resiliente e equitativa.

6 REFERÊNCIAS

ARCINIEGAS, Daniel Fernando Terraza et al. Students' Attention Monitoring System in Learning Environments based on Artificial Intelligence. **IEEE Latin America Transactions**, v. 20, n. 1, p. 126-132, 2021.

AZEVEDO, Vanessa Ragone; DE ALMEIDA NEVES, Pedro. Desigualdades educacionais à luz da Covid-19: disparidades do meio rural e urbano. **Revista de Desenvolvimento e Políticas Públicas**, v. 5, n. 1, p. 25-54, 2021.

BARBOSA, Ana Beatriz Gomes; CAVALCANTI, Alexsandro Bezerra. Web Scraping e Análise de dados. **Anais... V CONAPESC, s/n. Campina Grande: Editora Realize**, 2020.

CASTIONI, Remi et al. Universidades federais na pandemia da Covid-19: acesso discente à internet e ensino remoto emergencial. **Ensaio: Avaliação e políticas públicas em educação**, v. 29, p. 399-419, 2021.

CHEN, Yahui. **Convolutional neural network for sentence classification**. Dissertação de mestrado - University of Waterloo, 2015.
CONNEAU, Alexis et al. Very deep convolutional networks for text classification. **arXiv preprint arXiv:1606.01781**, 2016.

CORTES, Corinna; VAPNIK, Vladimir. Support-vector networks. **Machine learning**, v. 20, p. 273-297, 1995.

COVER, Thomas M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. **IEEE transactions on electronic computers**, n. 3, p. 326-334, 1965.

D'HAEN, Jeroen et al. Integrating expert knowledge and multilingual web crawling data in a lead qualification system. **Decision Support Systems**, v. 82, p. 69-78, 2016.

DE SOUSA, Sérgio José; DE OLIVEIRA SANTIAGO, Monique; DIAS, Thiago Magela Rodrigues. Uma estratégia para identificação de gênero em repositórios de dados abertos utilizando um modelo de rede neural artificial. **Ciência da Informação**, v. 48, n. 3, 2019.

GALDINO, Igor Martins; DE LIMA GALLINDO, Erica; MOREIRA, Mário WL. Utilização de Bots para Obtenção Automática de Dados Públicos usando as Técnicas de Web Crawling e Web Scraping. In: **Anais do VIII Workshop de Computação Aplicada em Governo Eletrônico**. SBC, 2020. p. 172-179.

GÓES, Geraldo Sandoval; MARTINS, Felipe dos Santos; NASCIMENTO, José Antônio Sena do. Trabalho remoto no Brasil em 2020 sob a pandemia do Covid-19: quem, quantos e onde estão?. 2021.

GONÇALVES, Clarissa Petrachini et al. The impact of COVID-19 on the Brazilian Power Sector: operational, commercial, and regulatory aspects. **IEEE Latin America Transactions**, v. 20, n. 4, p. 529-536, 2022.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep learning**. MIT press, 2016.

HAWKINS, Douglas M. The problem of overfitting. **Journal of chemical information and computer sciences**, v. 44, n. 1, p. 1-12, 2004.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long short-term memory. **Neural computation**, v. 9, n. 8, p. 1735-1780, 1997.

HORHIRUNKUL, Chattarin et al. Thai Name Gender Classification using Deep Learning. In: **2021 25th International Computer Science and Engineering Conference (ICSEC)**. IEEE, 2021. p. 295-300.

HSU, Chih-Wei et al. A practical guide to support vector classification. 2003.

JELODAR, Hamed et al. Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. **IEEE Journal of Biomedical and Health Informatics**, v. 24, n. 10, p. 2733-2742, 2020.

KAFRUNI, Larissa Santos. Alunos de pós-graduação e os impactos na produtividade acadêmica durante o isolamento social da Covid-19. 2020.

KARIMI, Fariba et al. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In: **Proceedings of the 25th International conference companion on World Wide Web**. 2016. p. 53-54.

KOWALCZYK, Alexandre. Support vector machines succinctly. **Syncfusion Inc**, 2017.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, v. 25, 2012.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **nature**, v. 521, n. 7553, p. 436-444, 2015.

LEKAMGE, T. C.; FERNANDO, T. G. I. Finding the gender of personal names and finding the effect of gana on personal names with long short term memory. In: **2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)**. IEEE, 2019. p. 1-8.

LOPEZ, Marc Moreno; KALITA, Jugal. Deep Learning applied to NLP. **arXiv preprint arXiv:1703.03091**, 2017.

MANCEBO, Deise. Trabalho remoto na Educação Superior brasileira: efeitos e possibilidades no contexto da pandemia. **Revista USP**, n. 127, p. 105-116, 2020.

NEUMANN, Adriana et al. Produtividade acadêmica durante a pandemia: efeitos de gênero, raça e parentalidade. **Levantamento realizado pelo Movimento Parent in Science durante o isolamento social relativo à Covid-19. Parent In Science**, 2020.

ORELLANA, Jesem Douglas Yamall et al. Excesso de mortes durante a pandemia de COVID-19: subnotificação e desigualdades regionais no Brasil. **Cadernos de saúde pública**, v. 37, p. e00259120, 2021.

REGO, Rosana CB; SILVA, Verônica ML; FERNANDES, Victor M. Predicting Gender by First Name Using Character-level Machine Learning. **arXiv preprint arXiv:2106.10156**, 2021.

REGO, Rosana Cibely B. et al. Brazilian scientific productivity from a gender perspective during the Covid-19 pandemic: classification and analysis via machine learning. **IEEE Latin America Transactions**, v. 21, n. 2, p. 302-309, 2023.

SEPTIANDRI, Ali Akbar. Predicting the gender of indonesian names. **arXiv preprint arXiv:1707.07129**, 2017.

STRAUSS, Edilberto; JÚNIOR, Manoel Villas Bôas; FERREIRA, Wagner Luiz Lobo. A IMPORTÂNCIA DE UTILIZAR MÉTRICAS ADEQUADAS DE AVALIAÇÃO DE PERFORMANCE EM MODELOS PREDITIVOS DE MACHINE LEARNING. **Projectus**, v. 7, n. 2, p. 52-62, 2022.

TO, Huy Quoc et al. Gender prediction based on vietnamese names with machine learning techniques. In: **Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval**. 2020. p. 55-60.

TRIPATHI, Anshuman; FARUQUI, Manaal. Gender prediction of Indian names. In: **IEEE Technology Students' Symposium**. IEEE, 2011. p. 137-141.

VASHISTH, Pradeep; MEEHAN, Kevin. Gender classification using twitter text data. In: **2020 31st Irish Signals and Systems Conference (ISSC)**. IEEE, 2020. p. 1-6.

VENUGOPAL, Anand; YADUKRISHNAN, O. V.; NAIR, Remya. A SVM based Gender Classification from Children Facial Images using Local Binary and Non-Binary Descriptors. In: **2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)**. IEEE, 2020. p. 631-634.

ZHANG, Ke et al. Age group and gender estimation in the wild with deep RoR architecture. **IEEE Access**, v. 5, p. 22492-22503, 2017.

ZHANG, Ye; WALLACE, Byron. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. **arXiv preprint arXiv:1510.03820**, 2015.

ZHAO, Hua; KAMAREDDINE, Fairouz. Advance gender prediction tool of first names and its use in analysing gender disparity in Computer Science in the UK, Malaysia and China. In: **2017 International Conference on Computational Science and Computational Intelligence (CSCI)**. IEEE, 2017. p. 222-227.