Brazilian Scientific Productivity from a Gender Perspective during the Covid-19 Pandemic: Classification and Analysis via Machine Learning

G Nascimento, D Rodrigues, R Rego, S Nascimento, and V Silva

***Abstract*—Scientific research activities, in general, have been affected due to the COVID-19 pandemic and the need for distancing. In this paper, an analysis of the impact of COVID-19 on Brazilian scientific research is made, examining the number of complete manuscripts published in the period from 2018 to 2021, considering the researcher's gender. A *crawler* is implemented to extract the names of Brazilian researchers from the articles, and some machine learning models (SVM, BiLSTM, and CNN) are applied to classify the authors' gender. Some models are able to accurately predict gender in more than 95% of cases. In addition, we verified that in 2021 there was a drop of 37.47% in the publications of articles by Brazilian researchers. The results indicate that there was a greater drop in publications for females in most machine learning models applied, corroborating differences in the distribution of household activities and family care between the two genders.**

***Index Terms*—COVID-19, scientific production, gender classification, machine learning.**

I. Introdução

O Brasil foi e vem sendo um dos países mais afetados pela pandemia de COVID-19 [1], [2]. Segundo o Instituto de Pesquisa Econômica Aplicada (IPEA), em 2020, 11% dos trabalhadores brasileiros estavam exercendo suas funções de forma remota [3]. Com a recomendação de evitar aglomerações, minimizando a disseminação do vírus, as atividades escolares e acadêmicas de forma presencial foram prontamente suspensas. Buscando dar continuidade ao ensino sem risco de contaminação, o ensino a distância (EAD) foi aplicado [4]. Essa solução acentuou as desigualdades sociais do Brasil, pois para o ensino remoto ser possível é necessário ter acesso a internet de qualidade e equipamentos que estão fora da realidade de muitos brasileiros [5], [6]. É importante citar que fatores socioeconômicos e culturais podem ser empecilhos para que alguns indivíduos consigam ser produtivos trabalhando de forma remota [5]. Nesse ponto, merece destacar que as demandas com cuidados de crianças, idosos e doentes, que se exigem de forma particular às mulheres, limitam em muito as possibilidades de cumprimento do trabalho remoto a contento [7].

Para verificar as possíveis mudanças e implicações na produtividade cientifica relacionado a publicações de artigos científicos, dadas as mudanças das atividades antes presenciais para remotas, este trabalho é proposto. Para isso, foi realizada uma coleta de dados referentes às pesquisas acadêmicas realizadas no Brasil entre os anos de 2018 e 2021. Os dados dos trabalhos publicados são obtidos a partir da Plataforma Lattes. No entanto, a referida Plataforma não mostra de forma explícita uma das informações necessárias para verificar a produção científica de acordo com o gênero. Dessa forma, para verificar como a pandemia da COVID-19 afetou a produtividade cientifica brasileira na perspectiva de gênero, é preciso realizar a classificação de gênero dos autores dos trabalhos publicados.

A produtividade acadêmica na perspectiva de gênero é um tema importante que deve-se fomentar o debate. No trabalho [8], foi realizada uma pesquisa online com 10593 alunos de pós-graduação, dos quais 81% afirmaram estar tendo dificuldades na produção da suas teses/dissertações durante o período de suspensão das atividades presenciais. Dentre os homens que responderam à pesquisa, 36% dos quais não têm filhos disseram estar conseguindo trabalhar de forma remota, enquanto entre os que possuem filhos, apenas 17% afirmaram o mesmo. Para as mulheres que responderam a pesquisa, a suspensão das das atividades presencias teve um impacto maior, dentre as que responderam e não possuem filhos, 33% afirmaram estar conseguindo trabalhar de forma remota, entre as que possuem filhos, apenas 10% afirmaram o mesmo.

Classificar o gênero a partir de um nome pode parecer trivial, porém pode ser uma atividade exaustiva quando se trata de centenas ou até mesmo milhares de nomes [9]. Além disso, em aplicações como investigações de psicologia, questões de pesquisa antropológica e sociológica, inferir ou classificar o gênero de uma pessoa é necessário. Nesse caso, uma das soluções para facilitar esse processo é a automatização a partir da utilização de modelos de *machine learning e deep learning* para a predição de gênero [10], [11]. Muitos trabalhos preveem o gênero a partir de imagens de rostos de pessoas, como [12]–[15]. Outros pesquisadores propuseram a previsão de gênero usando processamento de linguagem natural [16]–[18]. Com base na capacidade de classificação dos modelos de *machine e deep learning*, este trabalho propõe a aplicação de técnicas de inteligência computacional para analisar o efeito da pandemia da COVID-19 na publicação de artigos acadêmicos

Gabriel da S. Nascimento, Instituto Federal da Paraíba, nascimento.gabriel@academico.ifpb.edu.br

Davi Emmanuel de Lima Rodrigues, Universidade Federal Rural do Semi-Árido, davi.rodrigues@alunos.ufersa.edu.br

R. C. B. Rego, Departamento de Engenharias e Tecnologia, Universidade Federal Rural do Semi-Árido, rosana.rego@ufersa.edu.br

Samara Martins Nascimento, Universidade Federal Rural do Semi-Árido, samara.nascimento@ufersa.edu.br

Verônica Maria Lima Silva, Universidade Federal da Paraíba, veronica.lima@ci.ufpb.br

no cenário brasileiro de acordo com o gênero. Assim, as principais contribuições deste trabalho são:

i) A criação 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;
ii) Aplicação da 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;
iii) A classificação de gênero e análise das informações coletadas da perspectiva da aplicação de técnicas de aprendizado de máquina.

II. Trabalhos Relacionados

A classificação de gênero é um problema abordado em algumas pesquisas, tais como [17], [19]–[22] sendo aplicada em cenários distintos. Para predizer o gênero com base nos nomes indianos, Tripathi [22] utilizou a técnica de *machine learning* SVM. Tripathi utilizou a abordagem de n-gram-suffixes ao invés da palavra inteira a qual determina um valor fixo para o tamanho palavra a ser classificada. Diferentemente, [19] apresentou um modelo para predição de nomes indonésios a partir de nomes completos e a partir do primeiro nome, analisando a diferença nos resultados de cada estilo de dados.

No trabalho [17], é analisado o efeito de Gana (trata-se de um conceito astrológico) em nomes pessoais e o utilizando para a classificação de gênero, resultando na criação de duas *Long Short Term Memory* (LSTM). Diferentemente, buscando a classificação de nomes vietnamitas, no trabalho [20] é apresentado uma comparação entre o SVM e uma rede LSTM. O SVM demonstrou o melhor resultado entre os modelos de *machine learning* e o LSTM apresentou resultados superiores a todos os outros modelos de *deep learning*. Os resultados foram analisados utilizando o primeiro nome, o sobrenome e o nome completo. Os autores disponibilizam ainda como resultado uma API web com base nos modelos treinados. Já no trabalho [21], técnicas de *machine learning* e *deep learning* são utilizadas em nomes brasileiros para classificação do gênero. Entre as técnicas utilizadas algumas apresentaram excelentes resultados, é possível citar o BiLSTM, com uma acurácia de 96.17%, e a CNN com 95.78%. Portanto, neste trabalho, são utilizados os modelos de *machine learning* para classificação de gênero de pesquisadores brasileiros utilizando o primeiro nome. Para análise dos modelos, métricas como a acurácia, *recall*, precisão e F1-*score* são utilizadas.

III. Coleta de Dados

A coleta de dados deste trabalho foi realizada a partir da Plataforma Lattes em 26 de outubro de 2021. A escolha pela Plataforma se deu pelo fato de ser amplamente utilizada por pesquisadores brasileiros de diversas áreas para organização e compartilhamento de seus currículos. Para isso, foi realizada a análise do funcionamento da Plataforma e o desenvolvimento de um *crawler* (ou *bot*) para a obtenção dos dados.

Na fase de análise do funcionamento da Plataforma Lattes, foi verificada a estrutura e execução de cada parte referente a pesquisa dos currículos e obtenção das informações necessárias. A partir disso, percebeu-se que devido a forma como a plataforma foi desenvolvida, apenas tornou-se possível a coleta de ”artigos completos publicados em periódicos”. Desta forma, a análise foi realizada com base nos artigos completos publicados em periódicos. Foram coletados dados como data de publicação dos trabalhos (entre 2018 a 2021), nome dos autores, título do artigo, nível acadêmico e nacionalidade.

A busca apresentou uma quantidade significativa de currículos, tornando a coleta algo demorado a ser concluído (com uma média de tempo de 10 segundos por currículo). A partir disso, foi definida a utilização de diversos *bots* de forma simultânea, ou seja, execução de forma paralela em um mesmo computador. Essa execução resultou em um ganho de tempo a depender da quantidade de *crawlers* utilizados, chegando a uma média de 4 segundos por currículo ao utilizar 10 execuções.

A coleta teve início com currículos de pesquisadores doutores, e posteriormente, de pesquisadores mestres. Além disso, as páginas para a coleta foram divididas entre os computadores disponíveis (sendo suas configurações: AMD athlon-3000G, Intel i3-7100u, Intel i3-8130U, Intel celeron-3865u), possibilitando a realização da análise de 98907 currículos em cerca de 36 horas. A coleta dos dados foi dividida em três partes, conforme mostrado na Fig. 1, sendo elas: a divisão, o controle e o resultado. A divisão é a parte inicial da coleta, onde é feita a segmentação das páginas a serem coletadas, as dividindo em blocos de acordo com um tamanho pré-definido antes da execução. O controle recebe os blocos de páginas e os gerencia, designando cada bloco a um *crawler*, no qual sua execução é feita de forma paralela, ao finalizar a coleta de um bloco é feito a verificação da existência de outros segmentos disponíveis, caso inexistente, é feito o encerramento do *bot*. O resultado se refere a parte final da execução, na qual os dados são armazenados em uma base de dados.

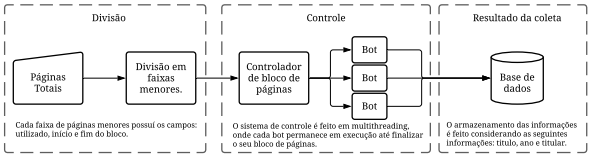


Fig. 1. Diagrama da coleta de dados.

A. Crawler Multithreading

O *crawler* trata-se de um programa desenvolvido para a navegação no Lattes e a coleta das informações desejadas acerca de pesquisadores brasileiros. Tendo em vista que as informações para a análise dos efeitos da Covid-19 na publicação de artigos científicos não possui uma base de dados própria, é necessário a utilização do *bot* para coleta e criação da mesma. Assim como apresentado na Fig. 2, o *bot*

é inicializado com as *tags* de pesquisa e o bloco de páginas a serem coletadas, permitindo o acesso aos currículos. A partir disso, é feita a análise das informações presentes na página, verificando a existência de "artigos completos publicados em periódicos"dentro dos anos definidos para a pesquisa, em caso da existência da informação, a mesma é armazenada.

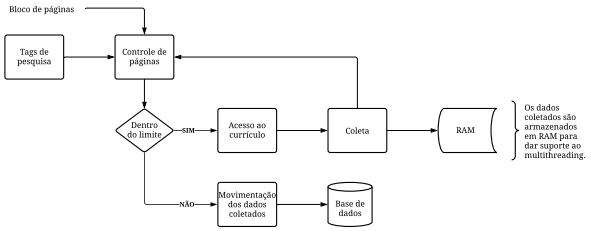


Fig. 2. Diagrama do crawler.

O resultado da coleta na Plataforma Lattes obteve 455794 trabalhos acadêmicos entre os anos de 2018 a 2021, sendo 114030 (2018), 117246 (2019), 138136 (2020) e 86382 (2021). Desconsiderando os *outliers*, isto é, pessoas com produções muito acima do intervalo interquartílico, teria-se uma média de aproximadamente 6 trabalhos por pessoa. O *crawler* e os dados coletados podem ser acessado no repositório [23].

B. Tratamento dos Dados

Algoritmos de *machine learning* 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 expressão (1). Esta técnica de codificação é chamada de *one-hot encoding*.

$$\begin{array}{c} j: \\ a: \\ e: \\ l: \end{array} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots & 0 \end{bmatrix} \quad (1)$$

O *one-hot encoding* foi utilizado para a codificação dos nomes aplicados no treinamento e classificação dos modelos de BiLSTM e CNN. Para a aplicação dos dados no SVM é necessário a utilização de outra técnica, pois o *one-hot* gera um vetor de valores enquanto o SVM aceita apenas valores unitários. Assim, foi utilizada a biblioteca *Natural Language Toolkit* (NLTK), que possui um módulo próprio para trabalhar com modelos de SVM. Para a utilização da NLTK, é criado um dicionário com as informações desejadas acerca de cada dado (neste caso, a informação é cada letra do nome de forma inversa com o limite de 10 letras), em seguida, o dicionário é agrupado com o seu valor de classificação (no caso da base de dados para o treinamento do modelo). A codificação da informação de saída, isto é, as classes (feminino/masculino) é feita por um processo diferente. Por se tratar de uma classificação binaria (onde se tem apenas dois valores possíveis), basta definir cada o valor de classificação como 0 ou 1. Finalizando a codificação dos dados, é feita a divisão de acordo com as necessidades de cada modelo.

IV. Modelos de Aprendizado de Máquina

Nesta seção, será apresentado um breve resumo da técnica SVM e dos modelos de *deep learning* BiLSTM e CNN utilizados para classificação de gênero dos dados coletados.

A. Support Vector Machine

O SVM trata-se de uma técnica de *machine learning* que utiliza vetores de suporte para classificação de dados através do aprendizado supervisionado. Esses vetores servem como "divisores de áreas" separando e agrupando os dados de acordo com suas semelhanças [24]. No SVM os vetores servem para definir o hiperplano e a questão mais importante é como identificar o hiperplano que separa as classes. Para determinação do hiperplano que define as margens de separação das classes é necessário definir um *kernel*. Se o problema de classificação possui padrões linearmente separáveis, então o uso de um *kernel* linear é suficiente. Entretanto, se os padrões são não lineares, é necessário fazer o uso de um *kernel* não linear, pois um problema não linear tem maior probabilidade de ser linearmente separável em um espaço de alta dimensão [25].

Para analisar o *kernel* que melhor se adequa ao problema de classificação textual de gênero, simulações com diferentes funções *kernels* foram realizadas. Na Tabela I tem-se os dados de acurácia e F1-Score de cada modelo SVM com sua respectiva função *kernel*. Dentre todos os modelos de classificação, o modelo que classificou mais corretamente foi o do *kernel* Nu RBF. Além disso, para uma melhor análise dos modelos com diferentes *kernels* as curvas de aprendizagens são mostradas nas Fig. 3 (a), (b) e (c). A curva de aprendizado está mostrando a relação da pontuação ou *score* do treinamento versus o *score* dos testes com validação cruzada (*cross validation*) para um modelo com um número variável de amostras de treinamento. Observa-se que os modelos apresentados nas Fig. 3 (a) e (c) sofrem tanto de erro devido ao viés (*bias*), quanto de erro devido à variância. Além disso, como as pontuações do treinamento e da validação cruzada convergem à medida que mais dados são adicionados, os modelos provavelmente não se beneficiarão de mais dados.

TABELA I
SVM - Acurácia dos modelos.

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

Na Fig. 3 (b), observa-se que os modelos apresentam baixa variância e *bias*. Além disso, com o aumento das amostras o modelo com *kernel* NuRBF melhora sua performance. Portanto, o modelo que mais se adequá a classificação de gênero é o modelo com *kernel* NuRBF.

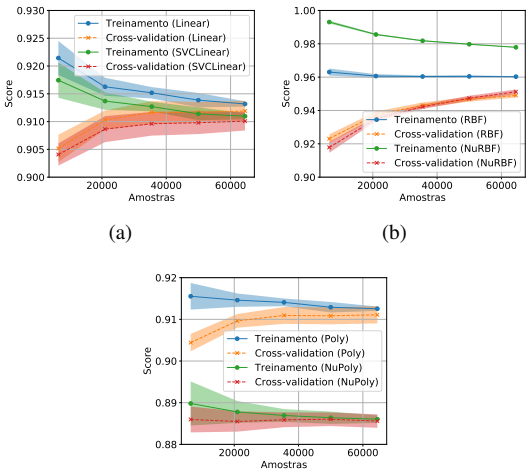


Fig. 3. Curvas de aprendizagem (a) SVM com *kernels* linear e SVC-linear, (b) SVM com *kernels* RBF e NuRBF, (c) SVM com *kernels* Poly e NuPoly.

B. LSTM Bidirecional

Long Short Term Memory (LSTM) é um tipo rede neural recorrente (*Recurrent Neural Network* - RNN), que foi desenvolvida com o intuito de resolver problemas da RNN clássica (como o *vanishing/exploding* do gradiente) [26], [27]. Uma LSTM bidirecional ou BiLSTM processa a sequência de informações em ambas as direções para trás ou para frente, tornando uma BiLSTM diferente da LSTM regular. [26]. A implementação do modelo BiLSTM, neste trabalho, se deu a partir de um modelo base desenvolvido por [21] e análise da literatura, permitindo a obtenção de um modelo com: uma camada BiLSTM com 20 neurônios de entradas e 128 neurônios de saída, uma camada de saída com 128 neurônios de entrada e 1 de saída e, por fim, uma camada de ativação com a função Sigmoid.

C. Rede Neural Convolucional Unidimensional

A Rede Neural Convolucional, ou simplesmente CNN, é um modelo de *deep learning* muito utilizado na classificação de imagens [28]. Mas nos trabalhos de [21], [28], [29], é possível ver que elas são perfeitamente aplicáveis em problemas de processamento de linguagem natural (*Natural Language Processing* - NLP). Para esses problemas, existem algumas abordagens, sendo uma delas a utilização de vetores de palavras pré-treinados [30]. Outra solução é a utilização redes neurais convolucionais a nível de caractere [21], [29]. Neste contexto, dada a capacidade da CNN em NLP, foi realizada a implementação da arquitetura unidimensional para classificação de gênero. A arquitetura proposta possui duas camadas convolucionais unidimensional com função de ativação ReLu (unidade linear retificada), uma camada para realizar o processo de *flatten*, isto é, converter os dados para dimensão 1D juntamente com a aplicação de um *dropout* de 20%. Além disso, ainda tem-se uma camada totalmente conectada com função de ativação ReLu e *dropout* de 20%. Ao final, tem-se uma ultima camada totalmente conectada com função de ativação *sigmoid*.

V. Resultados

Para a obtenção dos resultados foi necessário realizar algumas etapas. A etapa de treinamento será apresentada na seção V-A, enquanto as etapas de classificação e resultados serão apresentados na seção V-B.

A. Treinamento dos Modelos

O treinamento dos modelos foi realizado a partir da base de dados do brasil.io (https://brasil.io/dataset/genero-nomes/nomes), que consiste de 100787 nomes brasileiros obtidos no CENSO de 2010, onde 54.82% são nomes femininos e 45.18% são nomes masculinos. Além dos modelos SVM, BiLSTM e CNN, foram implementados também os classificadores *extra trees, decision tree, k-nearest neighbors algorithm (KNN), Naive Bayes, random forest, gradient boosting, light gradient boosting, logistic regression, classificador ridge e Ada boost*. Para os modelos de *machine learning*, os dados foram embaralhados e divididos em subconjuntos aleatórios em 80% para treino e 20% para teste, preservando a frequência dos rótulos das classes.

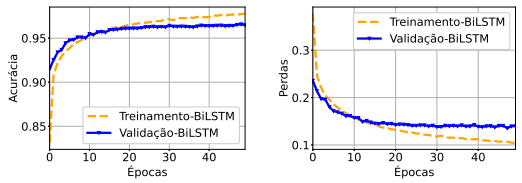
Os modelos de *deep learning* precisam de uma função de perda para calcular o desempenho do modelo durante o treinamento. Como o problema, neste trabalho, é uma classificação binária (1 ou 0, M ou F), a função de perda de entropia cruzada binária foi selecionada. Portanto, a função de perda é dada por

$$L(p,q) = -\frac{1}{N}\left[\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1-y_i)\cdot log(q(y_i))\right], \quad (2)$$

onde y é o rótulo (0 para feminino e 1 para masculino), $p(y)$ é a probabilidade prevista de o gênero ser masculino para todos os N pontos e $q(y_i) = 1 - p(y_i)$ é a probabilidade prevista de o gênero ser feminino.

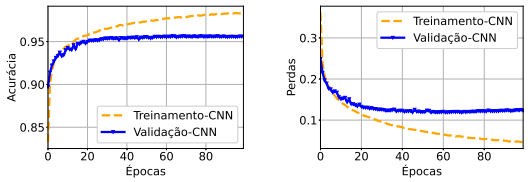
Durante o treinamento de algoritmos de *deep learning* é recomendado a utilização de parte dos dados para a validação do modelo, permitindo que o mesmo faça correções em sua predição durante o processo, permitindo uma menor possibilidade de *overfitting* e não generalizar a classificação [31]. A partir disso, o conjunto de dados foi embaralhado e divididos em subconjuntos aleatórios utilizado 60% para treino, 20% para validação e 20% para teste. Além disso, para evitar o *overfitting* é aplicado a técnica *early stopping* [26]. Na Fig. 4 (BiLSTM) e Fig. 5 (CNN) são apresentadas as curva de acurácia e perdas dos modelos em relação as épocas.

Na tabela II são apresentadas as métricas referentes a acurácia, precisão, recall e f1-score em cada um dos modelos implementados. Através dessas métricas é possível avaliar



(a) (b)

Fig. 4. Treinamento do modelo: (a) acurácia em relação as épocas (b) perdas em relação as épocas.



(a) (b)

Fig. 5. Treinamento do modelo: (a) acurácia em relação as épocas (b) perdas em relação as épocas.

completamente a eficácia dos modelos treinados. Dentre todos os modelos, o BiLSTM é o modelo com os melhores valores em todos as métricas. Isto é, o BiLSTM foi o modelo que classificou mais corretamente o gênero. Ainda tem-se também o SVM com *kernel* Nu RBF e CNN com bons resultados. Pode-se observar que o Naive Bayes foi técnica que apresentou uma menor sensibilidade ou *recall*, isto é, dentre todas as situações de classe verdadeira como valor esperado, quantas estão corretas. Portanto, com base nesses resultados as técnicas BiLSTM, CNN e SVM serão utilizadas, pois apresentaram melhores métricas para classificar o gênero dos nomes brasileiros coletados, e assim inferir a produtividade cientifica brasileira durante os anos de 2018 a 2021.

TABELA II
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	0.9543	0.9571	0.9597	0.9584
BiLSTM	0.9652	0.9612	0.9618	0.9614
CNN	0.9558	0.9529	0.9508	0.9518

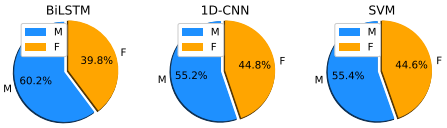
Após uma análise minuciosa, foi verificado que os nomes que causam mais erros nos classificadores são nomes que apresentam um *ratio* inferior a 1. Na Tabela III é possível visualizar alguns exemplos de nomes. Para mais detalhes a planilha completa pode ser acessada em [23].

TABELA III
NOMES CLASSIFICADOS DE FORMA ERRADA SEGUNDO A BASE DE DADOS DO CENSO.

Nome	Ratio	Gênero (CENSO)	Gênero (Classificadores)
Leovir	0.59	F	M
Jenair	0.55	F	M
Rutinei	0.54	F	M
Ildair	0.69	F	M

B. Aplicação dos Modelos

Com os modelos treinados e testados, foi realizada uma segunda tarefa de processamento que se refere à classificação dos dados, obtidos a partir da Plataforma Lattes. Nas Fig. 6 (a), (b) e (c), pode-se observar os resultados obtidos por meio dos diferentes modelos para cada gênero, entre os anos de 2018 a 2021. Os resultados mostrados indicam que os modelos 1D-CNN e SVM apresentaram um resultado similar, isto é, uma diferença de 0.2% na classificação de acordo com as Fig. 6 (b) e (c), respectivamente. Isso ocorre principalmente por os modelos 1D-CNN e SVM possuem métricas similares conforme mostrado na Tabela II. Já a classificação dos gêneros com o modelo BiLSM apresentou uma discrepância maior entre os sexos dos pesquisadores com 60.2% sendo do sexo masculino e apenas 39.8% sendo do sexo feminino, considerandos o período de 2018 a 2021, conforme visto na Fig 6 (a). Esses dados evidenciam a disparidade de gênero presente na produção científica.



(a) (b) (c)

Fig. 6. Classificação de gênero entre 2018 e 2021: (a) usando o modelo BiLSTM, (b) usando o modelo 1D-CNN e (c) usando o modelo SVM.

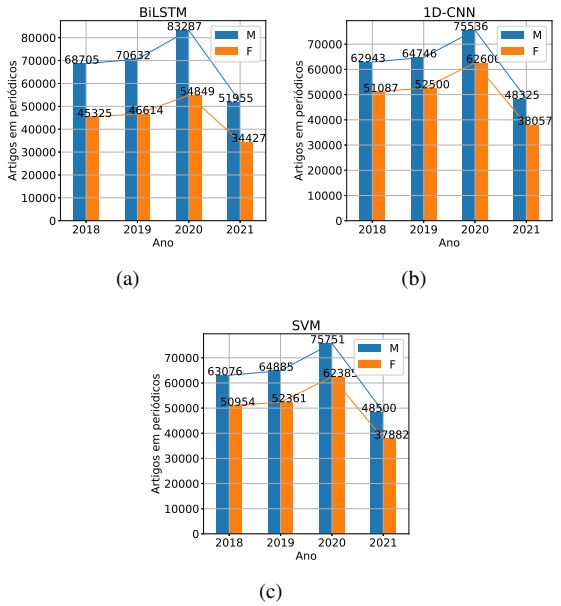
Uma visualização mais detalhada das publicações de artigos entre os anos de 2018 a 2021 é exibida nas Fig. 7 (a), (b) e (c). Conforme mostrado nas Fig. 7 (a), (b) e (c), o número de trabalhos científicos vinham crescendo até 2020, independentes do gênero dos autores. Mas, até 26 de outubro de 2021, esses números tiveram uma queda acentuada. Em termos de porcentagem, de 2020 para 2021 houve uma queda de 37.47% no número de trabalhos científicos publicados por pesquisadores de ambos os sexos, de acordo com os dados classificados com BiLSTM, conforme mostrado na Fig. 7 (a). Com relação a produção por gênero, houve uma queda de 37.62% na publicação de trabalhos por pesquisadores e uma queda 37.23% para pesquisadoras. Já observando a Fig. 7 (b), para os dados classificados com a CNN, houve uma queda de 36.02% no número de trabalhos científicos publicados por pesquisadores, e de 39.20% nos publicados por pesquisadoras. De acordo com a Fig. 7 (c), para os dados classificados com

o SVM, houve uma queda de 35.97% no número de trabalhos científicos publicados por pesquisadores, e de 39.27% nos publicados por pesquisadoras.

Com tabela IV, é possível analisarmos a diferença na produção acadêmica dos homens frente as mulheres. Conforme as Fig. 6 (a), (b) e (c), os homens apresentam um produção significativamente maior. Dessa forma, a diferença é sempre positiva. Por exemplo, de acordo com o BiLSTM, em 2020 os homens publicaram 20.59% a mais que as mulheres. Na CNN e no SVM a maior diferença ocorreu no segundo ano da pandemia (2021), e a menor em 2020, já para o BiLSTM a maior ocorreu no primeiro ano de pandemia e a menor no segundo (2021).

TABELA IV
DIFERENÇA NA PRODUÇÃO DO GÊNERO MASCULINO SOBRE O FEMININO.

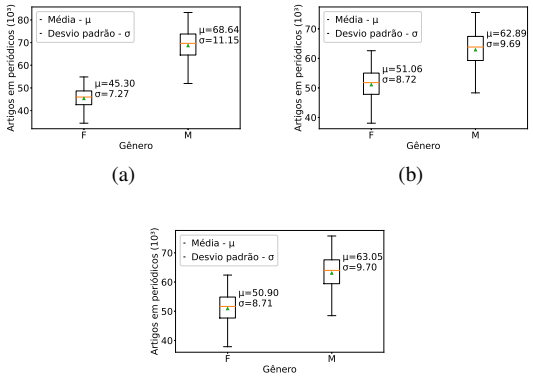
Modelo	2018	2019	2020	2021	Média (%)	Desvio Padrão
BiLSTM	20,50	20,49	20,59	20,29	20,47	0,13
CNN	10,40	10,44	9,36	11,89	10,52	1,03
SVM	10,63	10,68	9,68	12,29	10,82	1,07



(a) (b) (c)

Fig. 7. Produtividade científica na perspectiva de gênero: (a) usando o modelo BiLSTM, (b) usando modelo 1D-CNN e (c) usando SVM.

A distribuição dos dados classificados com os diferentes modelos (BiLSTM, 1DCNN e SVM) pode ser melhor analisada via os *boxplots* comparativos, exibidos nas Fig. 8 (a), (b) e (c). Os *boxplots* fornecem uma análise entre os artigos publicados em periódicos versus gêneros ao longo dos últimos quatro anos. Analisando as Fig. 8 (a), (b) e (c), observa-se que os dados apresentam uma distribuição simétrica para ambos os gêneros. Além disso, pode-se concluir que os artigos com autores com gênero masculino apresentam uma maior variabilidade quando comparado com os do gênero feminino. Uma dispersão maior entre os gêneros pode ser observada com a classificação obtida através do modelo de *deep learning* BiLSTM apresentado na Fig. 8 (a). Já a classificação obtida com os modelos 1D-CNN e SVM apresentam uma dispersão menor entre os gêneros, como pode ser observado na Fig. 8 (b) e (c), respectivamente. As dispersões apresentadas demonstram a dualidade conflituosa entre casa e trabalho que as mulheres enfrentam, confirmando a pesquisa da Parents de [32].



(a) (b) (c)

Fig. 8. Produtividade científica na perspectiva de gênero: (a) modelo BiLSTM, (b) modelo 1D-CNN e (c) modelo SVM.

VI. CONCLUSÃO

Neste trabalho, foi analisado os efeitos da pandemia do COVID-19 na publicação de trabalhos científicos brasileiros. Inicialmente, foi realizada a coleta dos dados necessários. Posteriormente, foi realizado o desenvolvimento e análise de modelos de *machine learning*, obtendo valores acima de 95%. Dessa forma, foi realizada a classificação das informações. A partir disto, foi feita a análise dos resultados em cada modelo, na qual é possível verificar que entre 2018 e 2020 houve, independente do gênero, um crescimento no número de publicações, sendo 2020 o ano com maior publicações. Entretanto, 2021 apresentou uma queda, levando a níveis menores que 2018 para ambos os gêneros. A partir dos valores de 2021, é realizada a análise da porcentagem de quedas em cada gênero, resultando em uma queda maior para pesquisadores do gênero feminino, quando a classificação foi realizada pelo modelo CNN e pelo SVM. Já quando os dados foram classificados pelo BiLSTM a queda foi maior para pesquisadores do gênero masculino. Vale ressaltar ainda que a plataforma Lattes é alimentada pelo pesquisador, dessa forma, existe uma probabilidade de um dado mais recente ainda não ter sido atualizado. Portanto, o que as estatísticas apresentadas mostram é a produtividade científica durante 2018 a 2021 na perspectiva dos dados da plataforma Lattes.

REFERÊNCIAS

[1] C. P. Gonçalves, D. S. Ramos, P. S. Rosa, M. H. Balan, B. Bezerra, M. Cavalieri, and R. F. de Mello, "The impact of covid-19 on the brazilian power sector: operational, commercial, and regulatory aspects," *IEEE Latin America Transactions*, vol. 20, no. 4, pp. 529–536, 2022.

[2] J. D. Y. Orellana, G. M. d. Cunha, L. Marrero, R. I. Moreira, I. d. C. Leite, and B. L. Horta, "Excesso de mortes durante a pandemia de covid-19: subnotificação e desigualdades regionais no brasil," *Cadernos de Saúde Pública*, vol. 37, p. e00259120, 2021.

[3] G. S. Góes, F. d. S. Martins, and J. A. S. Nascimento, "Trabalho remoto no brasil em 2020 sob a pandemia do covid-19: quem, quantos e onde estão?" Disponível em: https://www.ipea.gov.br/cartadeconjuntura/index.php/2021/07/trabalho-remoto-no-brasil-em-2020- sob- a-pandemia-do-covid-19-quem-quantos-e-onde-estao/., 2021.
[4] D. F. T. Arciniegas, M. Amaya, A. P. Carvajal, P. A. Rodriguez-Marin, L. Duque-Muñoz, and J. D. Martinez-Vargas, "Students' attention monitoring system in learning environments based on artificial intelligence," *IEEE Latin America Transactions*, vol. 20, no. 1, pp. 126–132, 2021.
[5] V. R. Azevedo and P. de Almeida Neves, "Desigualdades educacionais à luz da covid-19: disparidades do meio rural e urbano," *Revista de Desenvolvimento e Políticas Públicas*, vol. 5, no. 1, pp. 25–54, 2021.
[6] R. Castioni, A. A. S. d. Melo, P. M. Nascimento, and D. L. Ramos, "Universidades federais na pandemia da covid-19: acesso discente à internet e ensino remoto emergencial," *Ensaio: Avaliação e políticas públicas em educação*, vol. 29, pp. 399–419, 2021.
[7] D. Mancebo, "Trabalho remoto na educação superior brasileira: efeitos e possibilidades no contexto da pandemia," *Revista USP*, no. 127, pp. 105–116, 2020.
[8] L. S. Kafruni, "Alunos de pós-graduação e os impactos na produtividade acadêmica durante o isolamento social da covid-19," in *XXXII Salão de Iniciação Científica*, 2020.
[9] H. Zhao and F. Kamareddine, "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, pp. 222–227.
[10] S. J. de Sousa, M. d. O. Santiago, and T. M. R. Dias, "Uma estratégia para identificação de gênero em repositórios de dados abertos utilizando um modelo de rede neural artificial," *Ciência da Informação*, vol. 48, no. 3, mar. 2020. [Online]. Available: http://revista.ibict.br/ciinf/article/view/4908
[11] C. Horhirunkul, S. Vasupongayya, S. Sae-wong, S. Suwanmanee, and T. Angchuan, "Thai name gender classification using deep learning," in *2021 25th International Computer Science and Engineering Conference (ICSEC)*. IEEE, 2021, pp. 295–300.
[12] K. Zhang, C. Gao, L. Guo, M. Sun, X. Yuan, T. X. Han, Z. Zhao, and B. Li, "Age group and gender estimation in the wild with deep ror architecture," *IEEE Access*, vol. 5, pp. 22 492–22 503, 2017.
[13] A. Venugopal, O. Yadukrishnan, and R. Nair T., "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)*, 2020, pp. 631–634.
[14] S. Mittal and V. S. Rajput, "Gender and age based census system for metropolitan cities," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2020, pp. 1094–1097.
[15] A. Kuehlkamp and K. Bowyer, "Predicting gender from iris texture may be harder than it seems," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 904–912.
[16] P. Vashisth and K. Meehan, "Gender classification using twitter text data," in *2020 31st Irish Signals and Systems Conference (ISSC)*, 2020, pp. 1–6.
[17] T. Lekamge and T. Fernando, "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)*, vol. 250, 2019, pp. 1–8.
[18] F. Karimi, C. Wagner, F. Lemmerich, M. Jadidi, and M. Strohmaier, "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, pp. 53–54.
[19] A. A. Septiandri, "Predicting the gender of indonesian names," 2017.
[20] H. Q. To, K. V. Nguyen, N. L.-T. Nguyen, and A. G.-T. Nguyen, "Gender prediction based on vietnamese names with machine learning techniques," *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, Dec 2020. [Online]. Available: http://dx.doi.org/10.1145/3443279.3443309
[21] R. C. Rego, V. M. Silva, and V. M. Fernandes, "Predicting gender by first name using character-level machine learning," *arXiv preprint arXiv:2106.10156*, 2021.
[22] A. Tripathi and M. Faruqui, "Gender prediction of indian names," in *IEEE Technology Students' Symposium*. IEEE, 2011, pp. 137–141.
[23] G. Silva, D. Emmanuel, and R. Rosana. (2022, 7) Scientific productivity: Gender classification . [Online]. Available: https://github.com/TheGabrielSN/gender-classification-scientific-productivity
[24] A. Kowalczyk, *Support vector machines succinctly*. Syncfusion, Inc., 2017, vol. volume.
[25] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Transactions on Electronic Computers*, vol. EC-14, no. 3, pp. 326–334, 1965.
[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
[27] H. Jelodar, Y. Wang, R. Orji, and S. Huang, "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*, vol. 24, no. 10, pp. 2733–2742, 2020.
[28] M. M. Lopez and J. Kalita, "Deep learning applied to nlp," *arXiv preprint arXiv:1703.03091*, 2017.
[29] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, pp. 649–657, 2015.
[30] Y. Chen, "Convolutional neural network for sentence classification," Master's thesis, University of Waterloo, 2015.
[31] D. M. Hawkins, "The problem of overfitting," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 1, pp. 1–12, 2004, pMID: 14741005. [Online]. Available: https://doi.org/10.1021/ci0342472
[32] A. Neumann *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.

Gabriel da S. Nascimento Graduando em Engenharia de Computação pelo Instituto Federal de Educação, Ciências e Tecnologia da Paraíba. Possuí ensino técnico em Telecomunicações pela Escola Técnica Redentorista, além de conhecimentos na área de programação (Python, C/C++). Atua na área de desenvolvimento de software de otimização com python 3. Interesse em pesquisas nas áreas de inteligencia artificial, sistemas embarcados e desenvolvimento de software.

Davi Emmanuel de Lima Rodrigues Graduando em Ciência e Tecnologia pela Universidade Federal Rural do Semi-Árido. Atuando no projeto de pesquisa: PEH30001-2021 - Produtividade Científica na perspectiva de gênero durante a pandemia do Covid-19 (UFERSA). Possui conhecimentos na área de programação (Python, C, Java e Javascript). Interesse em pesquisas nas áreas de inteligência artificial, desenvolvimento WEB e mobile.

Rosana Cibely B. Rego Professora no Departamento de Engenharias e Tecnologia na Universidade Federal Rural do Semi-Árido e Doutora em Engenharia Elétrica e de Computação pela Universidade Federal do Rio Grande do Norte (2022), com pesquisas na área de Controle Inteligente, controle Neural, redes neurais, aprendizado de máquina. Certificada pela Huawei ICT Academy em inteligência artificial. Possui conhecimentos na área de programação (C/C++, Java, Python, Fortran, MatLab/Scilab).

Samara Martins Nascimento Doutora em Ciência da Computação, pela Universidade Federal do Ceará. Professora Adjunta na Universidade Federal Rural do Semi-Árido (UFERSA). É uma das líderes dos Grupos de Pesquisas Laboratório de Inovações em Software (LIS) e Laboratório de Inteligência Computacional (CiLab). As suas principais áreas de interesse são Banco de Dados, Big Data, Data Streams, Data Warehouse, Gerenciamento de Dados, Qualidade de Software e Métricas de Software.

Verônica M L Silva Possui graduação em Engenharia de Computação pela Universidade Federal do Ceará (2011). Desde 2015 é professora da Universidade Federal do Semi-Árido Rural (UFERSA) e doutora em Engenharia Elétrica pela Universidade Federal de Campina Grande (UFCG), 2019. Seus interesses de pesquisa incluem sistemas digitais, analógicos-conversores digitais, conversores analógicos para informação e sistemas embarcados, inteligência artificial.