METODOLOGIA DE BUSCA AUTOMATIZADA DE DADOS NO PODER PÚBLICO: uma abordagem para análise de dados e aprendizado de máquina em políticas públicas

Antonio Ordones Neto – antonio.ordones@aluno.unb.br

Hamilton Batista de Sousa Silva – silva.hamilton@aluno.unb.br

Resumo

Este artigo examina uma metodologia criada pela ONG Transparência Brasil, voltada para a exploração do uso da inteligência artificial (IA) no setor público, particularmente no que diz respeito à coleta de dados. A investigação tem como meta avaliar a viabilidade e a potencial replicação dessa abordagem em diferentes áreas, incluindo a de políticas públicas. Pretende-se ainda oferecer aos administradores e acadêmicos uma perspectiva mais transparente e compreensível sobre a análise de dados e o emprego da aprendizagem de máquina em tais contextos. A pesquisa conclui sublinhando a importância da manipulação e processamento de informações como recursos sólidos na tomada de decisões, enriquecendo a compreensão na área e abrindo caminho para aplicações futuras no âmbito público.

Palavras-chave: avaliação metodológica. inteligência artificial. banco de dados. tratamento de dados. políticas públicas.

Abstract

This article examines a methodology developed by the NGO Transparência Brasil, aimed at exploring the use of artificial intelligence (AI) in the public sector, particularly regarding data collection. The investigation seeks to assess the feasibility and potential replication of this approach in various fields, including public policy. It also aims to provide administrators and academics with a more transparent and comprehensible perspective on data analysis and the application of machine learning in such contexts. The research concludes by emphasizing the importance of information manipulation and processing as solid resources in decision-making, enriching understanding in the field, and paving the way for future applications in the public domain.

Keywords: methodological evaluation. artificial intelligence. database. data processing. public policies.

1 INTRODUÇÃO

Apesar de já serem reconhecidos os benefícios potenciais das tecnologias de Inteligência Artificial (IA) na prestação de serviços públicos, é necessário considerar os possíveis impactos negativos dessas ferramentas no que diz respeito aos direitos individuais, à privacidade, à proteção contra discriminação, ao acesso à justiça e às liberdades de expressão, associação e reunião.

Nesse contexto, a Organização não governamental (ONG) Transparência Brasil[1] elaborou o Relatório "Recomendações de Governança: uso de inteligência artificial pelo poder público" (2021) no qual foi empregada metodologia que teve como objetivo encontrar ferramentas de IA em uso no setor público de forma automatizada, e, assim, fornecer insumos para recomendações a partir da análise dos casos.

Com base nos dados coletados, foram analisadas as ferramentas de IA utilizadas pelos Poderes Executivo, Legislativo e Judiciário brasileiros, identificando possíveis impactos negativos nos direitos dos cidadãos e apresentando as principais preocupações da sociedade civil em relação ao uso dessas tecnologias. Ao final, são oferecidas recomendações de governança para a aplicação de sistemas de IA no setor público.

Considerando esse contexto, o presente artigo buscará avaliar a metodologia implementada pela Transparência Brasil, isto é, a sequência de ações executadas para a realização de tarefas e automações, incluindo processamento de dados, manipulação de arquivos, execução de comandos, interação com sistemas e outras operações que foram usadas na pesquisa. Serão executados os códigos desenvolvidos pela organização, em parceria com a Northwestern University, que foram disponibilizados gratuitamente na plataforma GitHub[2].

A partir da avaliação metodológica proposta neste artigo, busca-se demonstrar aos gestores e pesquisadores que a integração de áreas tão distintas é possível, mesmo com apenas uma base introdutória em programação, computação ou estatística. A análise concentra-se nas práticas metodológicas subjacentes à aplicação de inteligência artificial e tecnologias correlatas e no seu potencial inovador no campo das políticas públicas.

Dessa forma, a iniciativa deste trabalho buscar enriquecer o debate sobre políticas públicas no cenário nacional, bem como propiciar aos profissionais da área a oportunidade de

[1] A Transparência Brasil é uma ONG que possui a missão de promover a transparência e o controle social do poder público, contribuindo para a integridade e o aperfeiçoamento das instituições, das políticas públicas e do processo democrático.
[2] GitHub é uma plataforma de hospedagem de código-fonte e arquivos. Ele permite que programadores, ou qualquer usuário cadastrado na plataforma contribuam em projetos a partir de qualquer lugar do mundo.

aprimorar habilidades analíticas, capacitando-os para identificar soluções eficientes, interpretar dados e discernir padrões e tendências, a partir de uma perspectiva sistêmica.

1.1 Objetivo

Constituem-se objetivos da pesquisa:

a) Realizar a avaliação da metodologia empregada na pesquisa do Transparência Brasil para a coleta e análise de dados
b) Fornecer aos gestores e pesquisadores de políticas públicas uma compreensão acerca do emprego das ferramentas utilizadas; e
c) Demonstrar como o uso de ferramentas analíticas pode auxiliar na tomada de decisão do âmbito de políticas públicas.

1.2 Metodologia

O foco deste trabalho é avaliar a metodologia de coleta e análise dados, empregada na pesquisa do Transparência Brasil (2021), particularmente quanto à sua possível aplicabilidade e repetibilidade por gestores públicos que operam fora do domínio da Tecnologia da Informação (TI). A análise da eficácia envolve o exame da coleta de dados, a adequação das ferramentas utilizadas e a confiabilidade dos resultados gerados.

A aplicabilidade, neste cenário, refere-se à extensão em que a metodologia pode ser replicada em diferentes situações. O artigo propõe uma análise dessa aplicabilidade, buscando determinar se a metodologia é versátil o suficiente para ser utilizada em outras áreas da gestão governamental.

A repetibilidade, critério fundamental nesta avaliação, se relaciona à capacidade da metodologia de ser consistentemente replicada em diferentes contextos e por diferentes atores. Esta característica assume particular importância na gestão pública, considerando a variedade de contextos. A avaliação da repetibilidade examina se a metodologia é suficientemente bem definida e estruturada para permitir que gestores públicos de outras áreas além da TI possam implementá-la obtendo resultados satisfatórios.

A ênfase na repetibilidade por atores fora do campo da TI realça a necessidade de métodos que sejam não apenas tecnicamente sólidos, mas também acessíveis e compreensíveis para um amplo espectro de profissionais. Tal abordagem facilita a colaboração interdisciplinar e a inovação no setor público.

Nesse sentido, serão reproduzidas as diversas etapas tecnológicas e metodológicas de coleta, processamento, classificação e análise de informações, conforme detalhadas na Tabela 1, a fim de avaliar sua repetibilidade por gestores públicos em contextos variados.

Tabela 1 – Procedimentos a serem executados com base na metodologia do Relatório "Recomendações de Governança: uso de inteligência artificial pelo poder público" do Transparência Brasil

1. Identificação de ferramentas e linguagens
2. Execução de *scripts*
3. Extração de Texto dos sites selecionados
4. Triagem e Classificação manual dos dados (endereços dos sites)
5. Implementação de um Modelo Preditivo para avaliar a relevância dos textos extraídos
6. Análise dos Resultados

2 DESENVOLVIMENTO

Inicialmente foram analisados os códigos da coleta de dados automatizada, disponibilizados no repositório do *Github*[3] pela Transparência Brasil. Identificou-se, então, que foi utilizada a linguagem de programação foi Python, que é de fácil aprendizagem e possui boa clareza.

Importa destacar que, embora a linguagem de programação Python seja reconhecida pela sua simplicidade e clareza, há vantagens e desvantagens para indivíduos leigos fora do campo da TI. As vantagens incluem a disponibilidade de uma vasta gama de recursos online para aprendizagem, uma comunidade ativa e solidária, e uma sintaxe intuitiva que facilita a compreensão inicial. No entanto, o aprendizado e a aplicação efetiva podem requerer tempo e esforço, possivelmente tornando a abordagem menos acessível para aqueles que não possuem uma base introdutória em tecnologia da informação.

Para trabalhar com códigos em Python, é necessário instalar a linguagem de programação[4] e escolher um editor de texto adequado para escrever e executar os códigos. Existem várias opções disponíveis, como Visual Studio Code, PyCharm, Atom e Sublime Text, entre outros. Cada um desses programas oferece recursos que facilitam a escrita e o teste de

[3] https://github.com/Transparencia-Brasil/algoritmos-brasil
[4] https://www.python.org/downloads/

códigos, e a escolha deve ser feita de acordo com as necessidades e preferências do usuário. É nesse editor que os códigos serão abertos, editados e executados.

O primeiro script a ser executado consta do arquivo chamado "buscador.py". Esse arquivo realiza buscas automáticas no Google com base em uma lista de palavras-chaves fornecidas, retornando o endereço do site, títulos e descrição de cada resultado.

A pesquisa aqui proposta utilizou as mesmas 65 palavras-chaves (Anexo I) da pesquisa do Transparência Brasil, limitando-se os resultados a sites hospedados nos domínios dos Três Poderes e do Ministério Público (gov.br, jus.br, leg.br e mp.br).

Ao final da coleta dos dados, foram obtidos um total de 9.436 registros, excluindo-se as duplicatas, conforme resumido no Anexo 2. Os dados foram salvos em uma planilha.

Assim, após a coleta do código do buscador, é necessário fazer uma triagem de todas as URLs e reservar aquilo que for de interesse e conexo para a pesquisa. Essa análise, tal qual foi feita na pesquisa do Transparência Brasil, foi manual:

> "Com base nesses resultados, foram manualmente avaliados 6.195 endereços de sites governamentais (URLs) de acordo com o conteúdo que apresentavam para julgar se estavam relacionados ao tema ou não." (TRANSPARÊNCIA BRASIL, 2020, pag. 5)

Essa triagem, no presente trabalho, consistiu em criar uma nova coluna no arquivo consolidado e classificar cada registro coletado como “Relevant” ou “Not Relevant”. Após essa classificação, foi realizada a extração dos conteúdos de texto dos sites usando os códigos "extrator.py" e "scraper.py".

A fim de complementar a análise e com o intuito de delimitar ainda mais a amostra pesquisada, foi desenvolvido um código próprio[5] para segmentar os dados de acordo com a presença de um conjunto de palavras específicas. No caso específico, filtrou-se com base em palavras-chave da área de segurança pública.

Por fim, executou-se o código "estimador.py”, que contém um modelo preditivo[6], utilizado para prever resultados a partir da análise de um conjunto de dados fornecido. O modelo utilizado se baseia no algoritmo de regressão logística, que é uma técnica de aprendizado de máquina usada para classificação de dados.

Os textos extraídos e suas respectivas classificações ("Relevant" ou "Not relevant") foram apresentados ao modelo, que a partir desse conjunto de dados foi capaz de prever a relevância de novos textos em uma escala de 0(irrelevante) a 1(relevante).

---

[5] O código está disponível em https://github.com/AntonioOrdones/Segmenta-o_Seguran-a-P-blica.git

[6] Um modelo preditivo é uma ferramenta/código usada para fazer previsões ou estimativas sobre algo desconhecido, baseando-se em informações conhecidas.

3 CONSIDERAÇÕES FINAIS

A relevância da análise de dados e do emprego de métodos computacionais ultrapassa o campo tradicional da Tecnologia da Informação. Essas abordagens têm se mostrado úteis em contextos tão abrangentes quanto o das políticas públicas.

Nesse cenário, ressalta-se que a união de métodos de áreas que à primeira vista parecem desconectadas, como a gestão pública e a TI gera novas oportunidades de aprimoramento da eficácia da gestão/governança pública.

Reforça-se que a tecnologia não é uma entidade isolada, mas um elemento integrado a diversos aspectos da vida moderna, incluindo a Administração Pública. Neste contexto, a linguagem de programação Python mostrou-se uma ferramenta útil e acessível para os gestores fora do campo da TI. Essa aplicação ofereceu vantagens como flexibilidade na análise de grandes volumes de dados e facilidade de integração com outras tecnologias. No entanto, também apresentou desafios, como a necessidade de treinamento específico e complexidades no manuseio de determinadas tarefas.

A aplicação prática dessas técnicas no setor público vai além de uma mera possibilidade teórica, tornando-se uma exigência vital para lidar com os intricados desafios da sociedade contemporânea, uma vez que a inclusão de métodos de TI em políticas públicas, pode promover uma administração pública mais transparente, eficiente e inovadora.

Dessa maneira, conclui-se que é possível a replicação da metodologia analisada, mas também enfatiza a necessidade de integrar a análise de dados na prática dos gestores públicos de setores não relacionados à TI. Tal adoção pode levar a uma gestão mais eficaz, transparente e inovadora, em harmonia com as exigências e complexidades da atualidade, embora seja fundamental que os gestores possuam qualificação mínima em conceitos de TI para enfrentar possíveis dificuldades.

REFERÊNCIAS

TRANSPARÊNCIA BRASIL. **Recomendações de Governança**: uso de inteligência artificial pelo poder público. 2021. Disponível em: https://www.transparencia.org.br/downloads/publicacoes/Recomendacoes_Governanca_Uso_IA_PoderPublico.pdf. Acesso em 07 jun. 2023.

Anexo 1 – palavras-chave

Algoritmo;
análise AROUND(3) automatizada;
Análise AROUND(3) preditiva;
avaliação AROUND(3) automática;
cálculo AROUND(3) automático;
cálculo AROUND(3) automatizado;
calibragem AROUND(3) automática;
classificação AROUND(3) automática;
classificação AROUND(3) automatizada;
classificação AROUND(3) modelo;
classificação AROUND(3) numérica;
Computação;
Computacional;
equação AROUND(3) classificação;
equação AROUND(3) pontuação;
equação AROUND(3) Ranking;
filtragem AROUND(3) automática;
filtragem AROUND(3) automatizada;
fórmula AROUND(3) classificação;
fórmula AROUND(3) ranking;
Informática;
matriz AROUND(3) cálculo;
matriz AROUND(3) classificação;
metodologia AROUND(3) classificação;
metodologia AROUND(3) rating;
modelagem AROUND(3) preditiva;
pontuação AROUND(3) automática;
scoring AROUND(3) automatizado;
simulação AROUND(3) automatizada;
desenvolvido AROUND(6) automatizado;
sistema AROUND(6) algoritmo;
sistema AROUND(6) automatizado;
plataforma AROUND(3) algoritmo;
plataforma AROUND(3) automatizado;
machine learning;
aprendizado AROUND(2) máquina;
inteligência artificial;
modelo AROUND(3) estatístico;
modelagem AROUND(3) estatística;
sistema AROUND(3) inteligente;
processamento de linguagem natural;
NLP;
deep learning;
redes neurais;
Regressão;
Tensorflow;
Keras;
Bigdl;
Microsoft Cognitive Toolkit;
PyTorch;
OLS;
Regressão logística;
regressão linear;
Random forest;
Árvore de decisão;
Decision tree;
Support vector machine;
Análise de cluster;
Robô;
Assistente virtual;
Bot;
e-serviço;
processo AROUND(3) automatizado;
sistema AROUND(3) automatizado; e
decisão AROUND(3) automatizada.

Anexo 2 – Resumo dos resultados do código de busca

Termo de pesquisa	gov.br	jus.br	leg.br	mp.br	Total
Algoritmo	40	66	42	50	198
análise AROUND(3) automatizada	65	52	28	25	170
Análise AROUND(3) preditiva	54	44	5	17	120
Análise de cluster	96	85	82	66	329
aprendizado AROUND(2) máquina	23	52	12	31	118
Árvore de decisão	29	93	95	93	310
Assistente virtual	87	83	42	71	283
avaliação AROUND(3) automática	93	57	23	18	191
bigdl	2				2
bot	4	39	49	66	158
cálculo AROUND(3) automático	93	92	29	28	242
cálculo AROUND(3) automatizado	30	42	7	16	95
calibragem AROUND(3) automática	31	7			38
classificação AROUND(3) automática	91	49	23	19	182
classificação AROUND(3) automatizada	13	14	3	14	44
classificação AROUND(3) modelo	89	77	52	53	271
classificação AROUND(3) numérica	83	32	12	15	142
Computação	20	88	84	58	250
computacional	90	31	59	77	257
decisão AROUND(3) automatizada	78	62	21	34	195
Decision tree	56	73	30	51	210
deep learning	13	16	8	11	48
desenvolvido AROUND(6) automatizado	35	22	5	11	73
e-serviço	96	34	95	96	321
equação AROUND(3) classificação	57	2		1	60
equação AROUND(3) pontuação	24				24
equação AROUND(3) Ranking	11				11
filtragem AROUND(3) automática	43	34	18	13	108
filtragem AROUND(3) automatizada	23	2	1	1	27
fórmula AROUND(3) classificação	93	17	12	9	131
fórmula AROUND(3) ranking	20	8			28
Informática	98	92	93	96	379
inteligência artificial	86	75	71	55	287

keras	10	6	6	4	26
machine learning	20	17	11	9	57
matriz AROUND(3) cálculo	87	21	9	5	122
matriz AROUND(3) classificação	86	39	15	25	165
metodologia AROUND(3) classificação	89	26	7	19	141
metodologia AROUND(3) rating	2		3	3	8
Microsoft Cognitive Toolkit	64			3	67
modelagem AROUND(3) estatística	14	10	9	24	57
modelagem AROUND(3) preditiva	59	3	4	1	67
modelo AROUND(3) estatístico	89	78	33	36	236
NLP	13	53	37	28	131
OLS	29	49	42	47	167
plataforma AROUND(3) algoritmo	25	33	13	8	79
plataforma AROUND(3) automatizado	83	24	6	7	120
pontuação AROUND(3) automática	58	19	6	7	90
processamento de linguagem natural	56	51	63	55	225
processo AROUND(3) automatizado	80	77	63	54	274
PyTorch	8	1			9
Random forest	69	14	7	19	109
redes neurais	36	60	72	46	214
regressão	65	94	71	8	238
regressão linear	46	69	51	47	213
Regressão logística	49	79	56	71	255
robô	77	91	60	71	299
scoring AROUND(3) automatizado	2				2
simulação AROUND(3) automatizada	25	3	1		29
sistema AROUND(3) automatizado	1	20	20	14	55
sistema AROUND(3) inteligente	79	89	92	58	318
sistema AROUND(6) algoritmo	84	68	58	25	235
sistema AROUND(6) automatizado	4	17	13	21	55
Support vector machine	11	15	12	17	55
tensorflow	5	10	1		16
TOTAL GERAL	**3191**	**2576**	**1842**	**1827**	**9436**