MONITOR ALEPE: SOLUCIONANDO PROBLEMAS DE CLASSIFICAÇÃO COM USO DE INTELIGÊNCIA ARTIFICIAL

Ícaro Bernardes[1]

Leonardo F. Nascimento[2]

RESUMO: O relacionamento dos chefes do executivo com as respectivas assembleias legislativas é definidor do sucesso ou insucesso de determinadas políticas e propostas. Neste sentido, a otimização deste diálogo demanda um conhecimento dos anseios e interesses de deputados e senadores por parte dos governos estaduais e/ou federal. O presente artigo tem como objetivo apresentar uma solução que envolve: a) a extração automatizada das proposições da Assembléia Legislativa do Estado de Pernambuco - ALEPE; b) o uso da API do ChatGPT para classificar automaticamente tais proposições de acordo com os eixos prioritários do governo de Pernambuco; c) criação de uma plataforma em R/Shiny com os dados obtidos. Soluções que integram Inteligência Artificial na gestão pública vão se tornar cada vez mais frequentes nos próximos anos. É preciso que esta incorporação seja tecnicamente elaborada ao mesmo tempo que assegure o controle dos gestores sobre os limites e possibilidades do seu uso.

Palavras-chave: Gestão pública, Inteligência Artificial, Assembléia Legislativa do Estado de Pernambuco, chatGPT

INTRODUÇÃO

Diariamente, atores políticos vão até a tribuna das assembleias municipais, estaduais e no parlamento federal para debater temáticas de interesse da gestão pública. Os discursos proferidos por vereadores, deputados e senadores[3] têm como objetivo legislar, fiscalizar e controlar o uso da *res* pública. Em termos mais amplos, as falas proferidas por tais atores compõem o substrato da vida política que se expressa sob a forma de conflitos, disputas e valores. Nas últimas décadas, o conjunto desses discursos passaram a estar disponíveis em formato eletrônico. Essa mudança trouxe a possibilidade de tratarmos o texto como dado

[1] Secretaria de Planejamento do Estado de Pernambuco - BIT Analytics. E-mail: icaro.coutinho@seplag.pe.gov.br.

* Os autores gostariam de agradecer à toda equipe do Instituto de Gestão da SEPLAG/PE.

[2] Secretaria de Planejamento do Estado de Pernambuco - Laboratório de Humanidades Digitais da Universidade Federal da Bahia. E-mail: leofn@seplag.pe.gov.br.

[3] E, também, por governadores de Estado e membros do Tribunal de Justiça do Estado, do Ministério Público Estadual e do Tribunal de Contas do Estado.

(GRIMMER, ROBERTS & STEWART, 2022) e com isso um conjunto de possibilidades metodológicas se colocou à disposição de pesquisadores e gestores.

Por outro lado, o relacionamento dos chefes do executivo com as respectivas assembleias legislativas é definidor do sucesso ou insucesso de determinadas políticas e propostas. Neste sentido, a otimização deste diálogo demanda um conhecimento dos anseios e interesses de deputados e senadores por parte dos governos estaduais e/ou federal. O presente artigo tem como objetivo apresentar uma solução que envolve: a) a extração automatizada das proposições da Assembléia Legislativa do Estado de Pernambuco - ALEPE; b) o uso da API do ChatGPT para classificar automaticamente tais proposições de acordo com os eixos prioritários do governo de Pernambuco; e, por fim, c) a criação de uma plataforma em R/Shiny com os dados obtidos para facilitar o acesso rápido e simplificado de tudo que está sendo proposto.

Soluções que integram Inteligência Artificial na gestão pública vão se tornar cada vez mais frequentes nos próximos anos. É preciso que esta incorporação seja tecnicamente elaborada e, ao mesmo tempo, que ela assegure o controle dos gestores sobre os limites e possibilidades do uso destas ferramentas, "contemplando aspectos técnicos, éticos e de formação de recursos humanos" (CÓBE et al, 2020, p.39). Na primeira seção, nós detalhamos o significado das proposições para o poder executivo, o desafio de usar modelos de NLP em português e apresentamos o problema de classificação automatizada segundo eixos temáticos do governo. Em seguida, é descrita a ferramenta "Monitor Alepe"

AS PROPOSIÇÕES DA ALEPE, *NATURAL LANGUAGE PROCESSING* (NLP) E O PROBLEMA DA CLASSIFICAÇÃO EM EIXOS TEMÁTICOS

De acordo com o regimento da Assembleia Legislativa de Pernambuco, as proposições dos deputados e comissões podem assumir a forma de: proposta de emenda à Constituição; projetos de lei; projetos de resolução; projetos de decreto legislativo; indicações; requerimentos; e emendas, subemendas e substitutivos[4]. Essas demandas geralmente são custosas para serem lidas diariamente, embora cada proposição tenha um determinado grau de relevância social e política. Por exemplo, pedidos de conserto de estradas, de mais segurança em determinados bairros e de fornecimento de água, constituem pautas importantes que nem sempre o executivo consegue, dentro de sua atividade de monitoramento, acompanhar.

[4] https://legis.alepe.pe.gov.br/texto.aspx?tiponorma=4&numero=1891&complemento=0&ano=2023&tipo=&url= acessado em set 2023.

Com o objetivo de sanar tal lacuna que o Monitor Alepe foi elaborado. Inicialmente, os autores pensaram que se tratava de um problema de modelagem de tópicos.

> O modelagem de tópicos é um tipo de metodologia de análise de *big data* para descobrir tópicos abstratos que ocorrem repetidamente em uma coleção de documentos. Ao escrever um artigo, um autor tem uma palavra-chave específica em mente, esta palavra-chave é repetida ao longo do artigo. A coleção de palavras-chave é modelada como uma mistura finita sobre um conjunto subjacente de probabilidades de tópico, e então um tópico latente em um documento específico é retornado. (BLEI & JORDAN, 2003; CHO, 2019)

No entanto, lidar com os tópicos "latentes" que "emergem" não resolveria completamente a tarefa de deliberação acerca das resoluções dos temas. Em outras palavras, é preciso que, por exemplo, as proposições urgentes sobre "educação" sejam enviadas para as equipes de gestores especializadas neste assunto. Há, deste modo, um fluxo de trabalho que precisa ser estabelecido entre as demandas apresentadas na ALEPE e a tomada de decisão das equipes da gestão do executivo. Neste sentido, achamos mais adequado classificar as proposições de acordo com os eixos estratégicos do programa de governo[5] que denominamos na plataforma de "eixos temáticos". Isso foi feito através da API[6] do chatGPT que descreveremos com mais detalhes na seção seguinte. Antes disso, para concluir, existem alguns desafios do uso do chatGPT no português.

Tecnicamente falando, o chatGPT é uma variante do modelo GPT (*Generative Pre-trained Transformer*) desenvolvido pela OpenAI. O núcleo do GPT é a arquitetura *Transformer*. Esta arquitetura foi introduzida em 2017 no artigo "Attention is All You Need" (VASWANI et al. 2017) e revolucionou o campo do Processamento de Linguagem Natural (NLP). A língua inglesa, devido ao seu domínio no mundo acadêmico e tecnológico, serve como a base primária para muitos *datasets* em NLP. Isso pode levar a uma baixa performance quando esses modelos são aplicados diretamente a textos em português, dado que a estrutura gramatical, a

[5] Cidades Sustentáveis e Resilientes Ciência; Tecnologia e Inovação; Clima e Meio Ambiente; Competitividade e Dinamismo Econômico; Cultura e Economia Criativa; Educação, Conhecimento e Inovação Gestão; Transparência e Colaboração; Inclusão Social e Direitos Humanos; Políticas para Mulheres; Saúde e Qualidade de Vida; Segurança Cidadã; Turismo; Zona Rural Mais Forte. Cf. https://static.poder360.com.br/2022/10/Plano-de-governo-Raquel-Lyra.pdf acessado em set 2023.

[6] API é a sigla para "Interface de Programação de Aplicações" (do inglês, "Application Programming Interface"). Trata-se de um conjunto de regras e especificações que permitem que aplicativos se comuniquem entre si. Em outras palavras, uma API permite que um software "fale" com outro software, facilitando a integração e o funcionamento conjunto de diferentes sistemas.

morfologia e o uso do idioma podem diferir significativamente do inglês. Embora não tenhamos dados sobre o treino do chatGPT versão 3.5 turbo que utilizamos, o GPT-3 consiste principalmente em inglês (93%) e 7% de conteúdo em outros idiomas. Esse cenário ressalta a importância de desenvolver e refinar os modelos específicos para o português, a fim de garantir que as nuances e características únicas da língua sejam capturadas com precisão.

O MONITOR ALEPE

A construção da ferramenta compreendeu três etapas: 1) a elaboração de um *script*[7] de *web scraping*[8] na linguagem R de programação para extração no *website* da ALEPE[9] de todas as proposições; 2) em seguida, nós enviamos *prompts* ao ChatGPT com os resultados extraídos e com os eixos do atual governo de Pernambuco; 3) por fim, nós produzimos um painel em Shiny/R para exploração das proposições classificadas e com um pequeno resumo para facilitar a exploração. A seguir, nós detalharemos cada uma destas etapas técnicas[10].

Web Scraping na linguagem R para extração no website da ALEPE

As diversas proposições proferidas por deputados, comissões, membros e órgãos da administração pública estadual são constantemente atualizadas no website da ALEPE. Por meio de um *script* de *web scraping* em R do portal (https://www.alepe.pe.gov.br/proposicoes/). Inicialmente, através de requisições POST, são extraídas informações (metadados) sobre as proposições: autor, data de publicação, nome da proposição e url para acesso ao conteúdo completo. No portal, as proposições são ordenadas por data e nome, mas são publicadas em ritmos irregulares. Isto é, proposições são adicionadas em ordem aleatória e por vezes com atraso. Assim, a atualização dos metadados é feita explorando documentos desde a data da coleta até dois dias anteriores à data mais recente no banco. Duplicatas são descartadas imediatamente nessa etapa. Essa tarefa é sucedida pela extração dos textos mais recentes.

O diálogo do *script* elaborado em R com o servidor é determinante para que o *web scraping* flua sem interrupções. Nesse sentido, nós adotamos três medidas para manter um bom diálogo

---

[7] Um script é uma coleção de tarefas escritas em uma determinada linguagem de programação para serem executadas de maneira automatizada e mais veloz que a execução por um operador humano. (NASCIMENTO, 2017, p.3)

[8] Trata-se de uma técnica de extração de dados online através de linguagens de programação através da qual podemos manipular websites, capturar seu conteúdo e organizá-lo sob a forma de uma base de dados. "A raspagem, entretanto, não é apenas uma técnica, mas também envolve uma forma particular de lidar com a informação e o conhecimento: é também uma prática analítica." (MARRES & WELTEVREDE, 2013, p.317)

[9] https://www.alepe.pe.gov.br/proposicoes/ acessado em set 2023.

[10] Os *scripts* de *web scraping* e envio de *prompts* ao ChatGPT podem ser acessados no repositório do projeto (https://github.com/IcaroBernardes/monitorALEPE).

com o portal da ALEPE. Primeiro, verificamos que o *robots.txt*[11] do *website* autorizava a manipulação do conteúdo com uso de script, Além disso, nós empregamos um rudimentar *backing off* de 1 segundo para não sobrecarregar o servidor com as requisições. Por fim, as funções de extração são encapsuladas em funções-advérbio[12] (que modificam outras funções) que mantém o fluxo do script mesmo em caso de erros.

O uso do ChatGPT em tarefas de classificação

Após a extração da base atualizada, outro script é invocado para submeter os textos mais recentes ao ChatGPT para resumo e uma classificação com múltiplos rótulos (Multi-Label Classification) dentro dos eixos temáticos. O seguinte *prompt* foi enviado ao GPT: *"Leia o texto a seguir: {texto}. Apresente em um parágrafo, precedido pela expressão 'Resumo: ', um resumo do texto com no máximo 200 caracteres. Em outro parágrafo precedido pela expressão 'Temas: ' apresente uma lista do grau de pertencimento do texto aos temas a seguir: {temas}. Apresente apenas os nomes de dois dos temas com alto grau de pertencimento separados por |."*. No *prompt, {texto}* se refere ao texto da proposição (limitado aos 10.000 primeiros caracteres, quando preciso) e *{temas}* se refere aos eixos temáticos. O prompt foi enviado no corpo de uma requisição POST ao API do ChatGPT com os parâmetros listados na Tabela 1.

Parâmetro	Valor
model	"gpt-3.5-turbo"
temperature	0
max_tokens	300
top_p	1.0
frequency_penalty	0.0
presence_penalty	0.0

Tabela 1: parâmetros da requisição ao ChatGPT

Por fim, produzimos um painel de exploração das proposições em Shiny. Ele facilita a navegação nos dados permitindo seleção temporal (data), categórica (autor, nome da proposição, demandam ação e temas) e textual (caixa de busca). As proposições filtradas pela seleção são exibidas à direita. Para melhor performance e visualização apenas 5 são exibidas por vez. O menu de navegação inferior dá acesso às páginas contendo todos os textos. O botão

[11] cf. https://www.alepe.pe.gov.br/robots.txt acessado em set 2023.
[12] cf. https://purrr.tidyverse.org/reference/faq-adverbs-export.html acessado em set 2023.

superior esquerdo permite ordenar as proposições por data, autor ou nome. Também é possível efetuar o *download* do resultado da seleção em pdf.

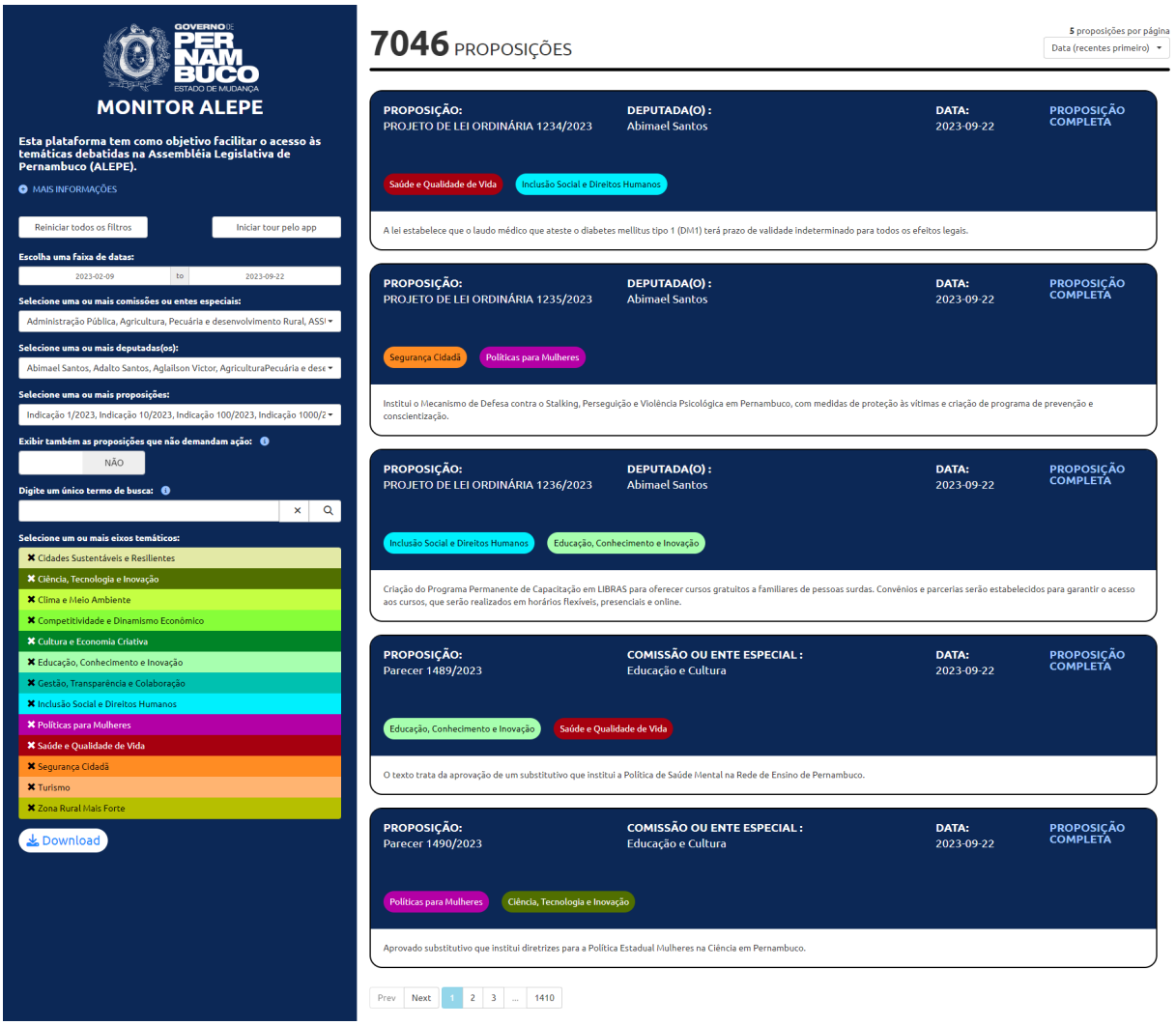

Figura 1: painel de exploração de dados das proposições da ALEPE

APRESENTAÇÃO E DISCUSSÃO DE RESULTADOS

Após a extração mais recente, foram obtidas 7672 proposições que correspondem ao período de 09/02/2023 até 22/09/2023. As Figuras de 2 até 7 evidenciam os entraves à classificação célere das proposições.

O primeiro desafio é o grande volume de textos vindos da ALEPE. A Figura 2 mostra a quantidade diária de proposições produzidas. Em vermelho está a média diária de propostas (aproximadamente 80 proposições). Essa quantidade supera o número de cadeiras (49) e de comissões (13) da Assembleia do estado. Isso ocorre, pois é comum um ente realizar múltiplas proposições em um mesmo dia.

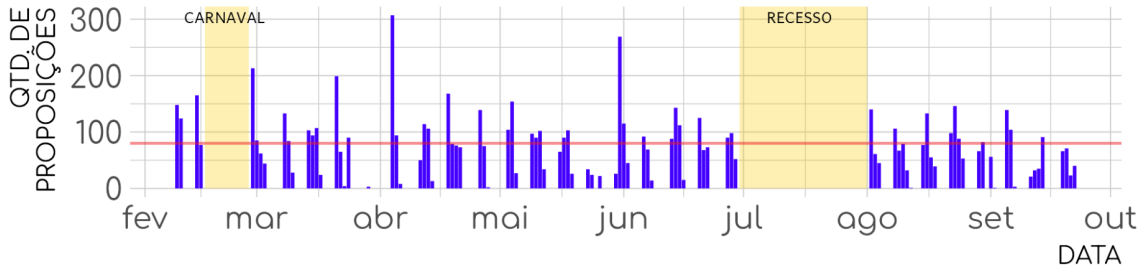

Figura 2: frequência diária das proposições

O grande volume de textos torna a leitura dispendiosa, ainda que grande parte dos mesmos tenha pequeno comprimento como se vê nas Figuras 3 e 4. Metade dos textos tem 504 caracteres ou menos (destacado em vermelho na Figura 3). Isso equivale a 7 linhas em um típico texto digital com 75 caracteres por linha. Em contraponto, a mediana do volume total de caracteres escritos diariamente é pouco superior a 100K (destacado em vermelho na Figura 4). Algo como mais de 1300 linhas de texto.

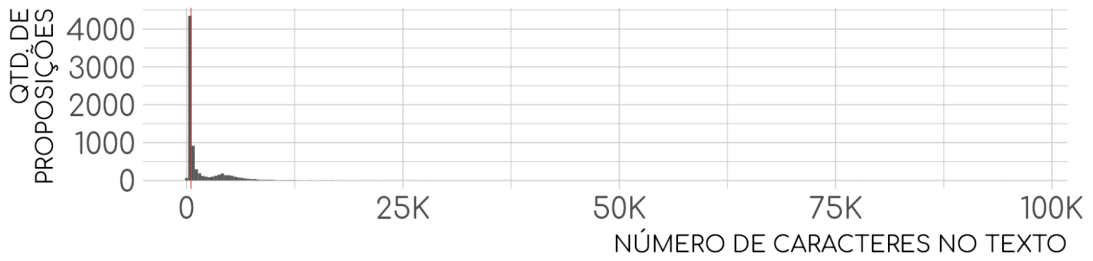

Figura 3: distribuição da quantidade de caracteres das proposições

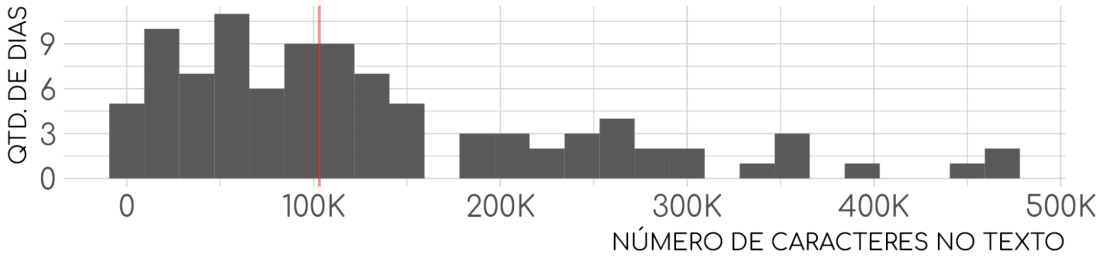

Figura 4: distribuição da quantidade total diária de caracteres das proposições

O segundo desafio é a grande variabilidade temática dos textos, a despeito da concentração da produção em poucos autores. A Figura 5 mostra como se distribuem as proposições por autoria. Podemos ver que as comissões, entes usualmente focados em certos temas, produzem bem menos textos que os deputados. Na Figura 6 vemos a concentração da autoria. Apenas 10 autores produziram metade de todas as proposições. Ademais, o autor mais produtivo gerou 18% de todas as proposições. É importante notar que apenas um dentre os dez autores mais produtivos tem um provável foco temático (comissão de Administração Pública).

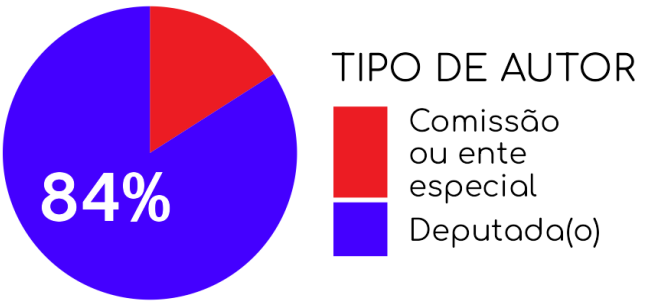

Figura 5: distribuição das proposições por tipo de autor

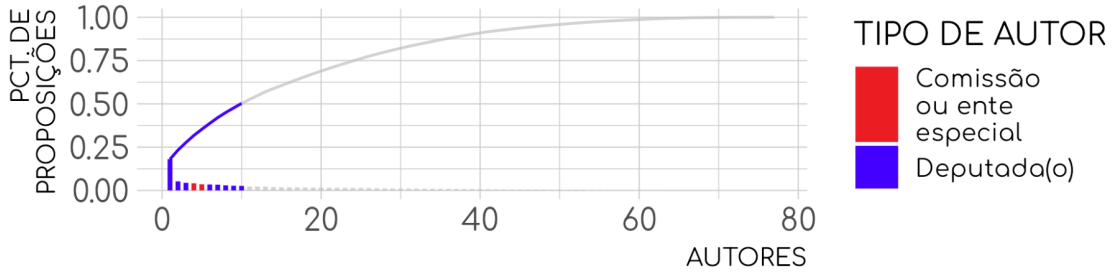

Figura 6: distribuição das proposições por autores mais frequentes

O terceiro desafio é distinguir quais proposições afetam o Executivo e, portanto, são nosso objeto de interesse. A discriminação de tais proposições foi feita de forma conservadora, isto é, apenas os textos claramente fora do escopo foram filtrados. O crivo usado foi a presença do conjunto de expressões "voto(s) de {termo}" ou "votação de {termo}", onde *{termo}* são palavras como aplauso, congratulações, pesar, etc. Notadamente elas são usadas para representar apoios simbólicos. Na Figura 7 vemos que cerca de 10% das proposições contém tais expressões. Um maior repertório de expressões pode ampliar esse número.

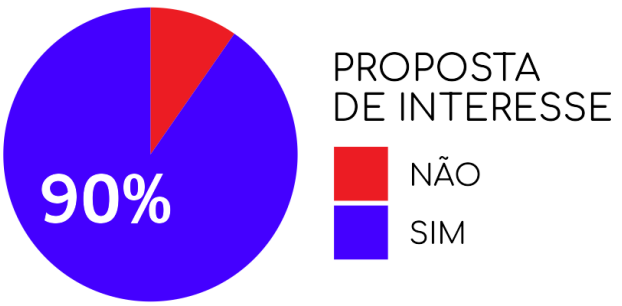

Figura 7: distribuição das proposições por interesse do Executivo

O uso do ChatGPT nos fornece o impulso inicial para abandonar a inércia diante dessa hercúlea tarefa. Em resposta a um mesmo *prompt*, obtemos resumo e temas mais aderentes. A

Figura 8 mostra com que frequência os eixos temáticos foram associados a alguma proposição. Nota-se que cerca de 70% das proposições foi associada aos seguintes eixos: gestão, transparência e colaboração (25,0% das proposições); saúde e qualidade de vida (14,5%); inclusão social e direitos humanos (14,5%); segurança cidadã (10,5%); cidades sustentáveis e resilientes (7,3%).

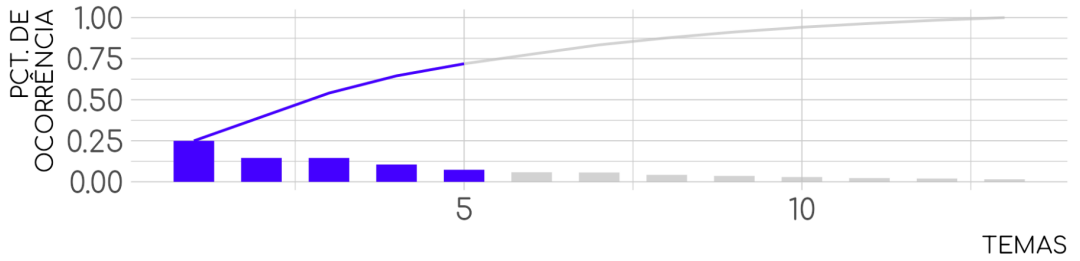

Figura 8: distribuição da ocorrência dos eixos temáticos nas proposições

CONSIDERAÇÕES FINAIS

É imperativo que os gestores públicos e líderes do poder executivo estabeleçam um diálogo contínuo e monitorem as deliberações do poder legislativo, visando sustentar uma interação construtiva com essa esfera governamental. As proposições legislativas representam um dos principais mecanismos formais através dos quais a "Casa do Povo" articula suas intenções e comunicações. Entretanto, a vastidão dos conteúdos textuais, a diversidade temática e a existência de documentos que não influenciam diretamente a gestão representam obstáculos para uma rápida e eficaz categorização destas proposições.

Soluções tecnológicas, como o ChatGPT e painéis desenvolvidos em Shiny/R, oferecem uma abordagem inicial mais acessível, especialmente quando comparadas ao investimento necessário para treinar modelos de NLP em português e criar painéis usando ferramentas proprietárias, como o Power BI. O "Monitor ALEPE", por sua vez, apresenta as proposições de maneira concisa, facilitando pesquisas temporais, categorizações e análises textuais, além de permitir a seleção e *download* das proposições relevantes.

Futuros desenvolvimentos na ferramenta apontam para o uso das categorias definidas pelo GPT em nova rodada de classificação. Seja para a classificação de futuras proposições através de uma listagem de palavras-chave pertencentes ao corpus de cada tema, seja para o treinamento de um modelo partindo de um conjunto de proposições cujo rótulo será validado.

REFERÊNCIAS

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993–1022, 2003.

CHO, Hae-Wol. Topic Modeling. **Osong Public Health and Research Perspectives**, v. 10, n. 3, p. 115–116, jun. 2019.

CÓBE, Raphael M. O. et al. Rumo a uma política de Estado para inteligência artificial. **Revista USP**, n. 124, p. 37–48, 19 mar. 2020.

GRIMMER, Justin; ROBERTS, Margaret E.; STEWART, Brandon M. **Text as Data**. [S.l.]: Princeton University Press, 2022.

MARRES, N. & WELTEVREDE, E. Scraping the Social? Journal of Cultural Economy, v. 6, n. 3, p. 313–335, 1 ago. 2013, p.317.

NASCIMENTO, Leonardo. Combinando webscraping em R e ATLAS.ti na pesquisa em ciências sociais: as possibilidades e desafios da sociologia digital. Disponível em: https://www.researchgate.net/publication/317343570_Combinando_webscraping_em_R_e_ATLASti_na_pesquisa_em_ciencias_sociais_as_possibilidades_e_desafios_da_sociologia_digital (acessado em sep 2023).

VASWANI, Ashish et al. **Attention Is All You Need**. . [S.l.]: arXiv. Disponível em: <http://arxiv.org/abs/1706.03762>. Acesso em: 30 set. 2023. , 12 jun. 2017

ANEXO

TERMO DE SUBMISSÃO DE ARTIGO, DE AUTORIZAÇÃO PARA PUBLICAÇÃO, CESSÃO DE DIREITOS AUTORAIS, DE PARTICIPAÇÃO ONLINE, DECLARAÇÃO DE ORIGINALIDADE E INEDITISMO*

Eu, Leonardo Fernandes Nascimento, inscrito no CPF sob o nº. 78092060525, cujo endereço laboral é Rua da Aurora, 1377 - Santo Amaro, Recife - PE, 50040-090, telefone(s) 7199898-4141 *e-mail* leofn@seplag.pe.gov.br, filiado à Instituição Secretaria de Planejamento do Estado de Pernambuco, na condição de Gerente Geral de Projetos Especiais, submeto ao 16º CONGRESSO DE GESTÃO PÚBLICA DO RIO GRANDE DO NORTE (CONGESP) o artigo intitulado "MONITOR ALEPE: SOLUCIONANDO PROBLEMAS DE CLASSIFICAÇÃO COM USO DE INTELIGÊNCIA ARTIFICIAL", para avaliação e publicação no site do 16º CONGESP, estou ciente que caso o meu trabalho seja aprovado pelo Comitê Científico assumo as seguintes responsabilidades:

1. Comparecerei para sua apresentação *online*, no dia e hora previamente comunicado e autorizo a publicação do material utilizado em minha apresentação no site do evento, assim como o uso de sons e imagens na internet.
2. Autorizo também o recebimento de mensagens via *WhatsApp* com informações relativas ao meu trabalho científico e/ou minha participação no evento.
3. Declaro que o trabalho é original e não contém nenhuma forma de plágio, estando o autor ciente da sua responsabilidade expressa pelo uso de textos e imagens de terceiros, quando tal uso exigir autorização.
4. Caso o texto seja aprovado e selecionado, responsabilizo-me pelo seu teor, ciente de que a publicação implica transferência dos direitos autorais ao 16º CONGESP, nas versões eletrônicas e publicações impressas, conforme permissivo constante do artigo 49 da Lei de Proteção de Direitos Autorais (Lei 9.610, de 19/02/98), e que a não observância desse compromisso submeterá o infrator a sanções e penas previstas no mesmo diploma legal.

Recife, 30 de Setembro de 2023.

Leonardo Fernandes Nascimento

Documento assinado digitalmente
gov.br
LEONARDO FERNANDES NASCIMENTO
Data: 30/09/2023 12:41:43-0300
Verifique em https://validar.iti.gov.br

ANEXO

TERMO DE SUBMISSÃO DE ARTIGO, DE AUTORIZAÇÃO PARA PUBLICAÇÃO, CESSÃO DE DIREITOS AUTORAIS, DE PARTICIPAÇÃO ONLINE, DECLARAÇÃO DE ORIGINALIDADE E INEDITISMO*

Eu, Ícaro Bernardes dos Santos Coutinho, inscrito no CPF sob o nº. 85379026591, cujo endereço laboral é Rua da Aurora, 1377 - Santo Amaro, Recife - PE, 50040-090, telefone(s) 7198313-4538 *e-mail* icaro.coutinho@seplag.pe.gov.br, filiado à Instituição Secretaria de Planejamento do Estado de Pernambuco, na condição de Gerente de Projetos Especiais, submeto ao 16º CONGRESSO DE GESTÃO PÚBLICA DO RIO GRANDE DO NORTE (CONGESP) o artigo intitulado "MONITOR ALEPE: SOLUCIONANDO PROBLEMAS DE CLASSIFICAÇÃO COM USO DE INTELIGÊNCIA ARTIFICIAL", para avaliação e publicação no site do 16º CONGESP, estou ciente que caso o meu trabalho seja aprovado pelo Comitê Científico assumo as seguintes responsabilidades:

1. Comparecerei para sua apresentação *online*, no dia e hora previamente comunicado e autorizo a publicação do material utilizado em minha apresentação no site do evento, assim como o uso de sons e imagens na internet.
2. Autorizo também o recebimento de mensagens via *WhatsApp* com informações relativas ao meu trabalho científico e/ou minha participação no evento.
3. Declaro que o trabalho é original e não contém nenhuma forma de plágio, estando o autor ciente da sua responsabilidade expressa pelo uso de textos e imagens de terceiros, quando tal uso exigir autorização.
4. Caso o texto seja aprovado e selecionado, responsabilizo-me pelo seu teor, ciente de que a publicação implica transferência dos direitos autorais ao 16º CONGESP, nas versões eletrônicas e publicações impressas, conforme permissivo constante do artigo 49 da Lei de Proteção de Direitos Autorais (Lei 9.610, de 19/02/98), e que a não observância desse compromisso submeterá o infrator a sanções e penas previstas no mesmo diploma legal.

Recife, 30 de Setembro de 2023.

Ícaro Bernardes dos Santos Coutinho

gov.br
Documento assinado digitalmente
ICARO BERNARDES DOS SANTOS COUTINHO
Data: 30/09/2023 12:45:18-0300
Verifique em https://validar.iti.gov.br