INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS FLORIANO
TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

PEDRO LUCAS ALVES DE ASSIS CARDOSO

EPIDEMIOLOGIA DIGITAL APLICADA A ARBOVIROSES NO PIAUÍ: PROPOSTA E AVALIAÇÃO DE UMA METODOLOGIA DE EXTRAÇÃO GEOLOCALIZADA DE MENÇÕES EM FONTES ABERTAS

FLORIANO – PI
2026

PEDRO LUCAS ALVES DE ASSIS CARDOSO

EPIDEMIOLOGIA DIGITAL APLICADA A ARBOVIROSES NO PIAUÍ: PROPOSTA E AVALIAÇÃO DE UMA METODOLOGIA DE EXTRAÇÃO GEOLOCALIZADA DE MENÇÕES EM FONTES ABERTAS

Trabalho de Conclusão de Curso (artigo científico) apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Floriano.

Orientador: Prof. Me. Ronaldo Pires Borges

FLORIANO – PI
2026

EPIDEMIOLOGIA DIGITAL APLICADA A ARBOVIROSES NO PIAUÍ: PROPOSTA E AVALIAÇÃO DE UMA METODOLOGIA DE EXTRAÇÃO GEOLOCALIZADA DE MENÇÕES EM FONTES ABERTAS

Digital epidemiology applied to arboviruses in Piauí: proposal and evaluation of a methodology for the geolocated extraction of mentions from open sources

Pedro Lucas Alves de Assis Cardoso[1]
Prof. Me. Ronaldo Pires Borges[2]

RESUMO

Este trabalho propôs, implementou em caráter de prova de conceito e avaliou criticamente uma metodologia de Epidemiologia Digital. O objetivo foi identificar e geolocalizar menções a arboviroses em fontes públicas online no estado do Piauí, com ênfase na caracterização dos desafios locais e na análise da viabilidade técnica e das limitações da abordagem. A pesquisa é aplicada, de natureza exploratória e descritiva, conduzida segundo a lógica da pesquisa de desenvolvimento, com abordagem quali-quantitativa. A metodologia articulou três etapas. A primeira foi a coleta automatizada de dados (*web scraping*) em notícias e boletins oficiais. A segunda foi o Reconhecimento de Entidades Nomeadas com a biblioteca spaCy, seguido da associação entre doença e município por co-ocorrência em sentença. A terceira materializou o fluxo em um protótipo funcional, dotado de painel com mapa interativo. A avaliação foi quantitativa e qualitativa. No plano quantitativo, as menções extraídas foram comparadas a um conjunto de referência (Padrão-Ouro) pelas métricas de precisão, revocação e medida F1. No plano qualitativo, analisaram-se os erros e as limitações observadas. Como referencial teórico, recorreu-se sobretudo a Eysenbach, Salathé, Nadeau e Sekine, e Jurafsky e Martin. A pesquisa justifica-se pelas limitações dos sistemas oficiais de

[1] Graduando em Tecnologia em Análise e Desenvolvimento de Sistemas, Instituto Federal do Piauí - Campus Floriano. E-mail: caflo.2024114tads0028@aluno.ifpi.edu.br
[2] Professor Mestre, Instituto Federal do Piauí - Campus Floriano. E-mail: ronaldo.borges@ifpi.edu.br

notificação, sujeitos a atraso e subnotificação, e pela lacuna de instrumentos que explorem fontes abertas no contexto piauiense. Os resultados indicaram alta precisão (94,9%) e revocação limitada (42,5%), com medida F1 de 58,7%. Esses números evidenciam a viabilidade da abordagem como sinal complementar. Ao mesmo tempo, revelam desafios locais relevantes, como a instabilidade da coleta automatizada, o ruído linguístico e a natureza heurística da geolocalização. Conclui-se que o estudo contribui menos como solução acabada e mais como caracterização crítica das condições, possibilidades e limites do uso de Epidemiologia Digital sobre fontes abertas para a vigilância de arboviroses no Piauí.

Palavras-chave: epidemiologia digital; processamento de linguagem natural; reconhecimento de entidades nomeadas; *web scraping*; arboviroses.

ABSTRACT

This work proposed, implemented as a proof of concept, and critically evaluated a Digital Epidemiology methodology. The goal was to identify and geolocate mentions of arboviruses in online public sources in the state of Piauí, emphasizing the characterization of local challenges and the analysis of the technical feasibility and limitations of the approach. It is an applied, exploratory, and descriptive study, conducted under a development research rationale with a qualitative-quantitative approach. The methodology combined three stages. The first was automated data collection (*web scraping*) from news outlets and official bulletins. The second was Named Entity Recognition with the spaCy library, followed by the association between disease and municipality through co-occurrence within a sentence. The third materialized the workflow in a functional prototype featuring a dashboard with an interactive map. Evaluation was both quantitative and qualitative. Quantitatively, the extracted mentions were compared against a reference set (gold standard) using precision, recall, and F1 score. Qualitatively, errors and observed limitations were analyzed. The theoretical framework drew mainly on Eysenbach, Salathé, Nadeau and Sekine, and Jurafsky and Martin. The research is justified by the limitations of official notification systems, subject to delay and underreporting, and by the gap of instruments exploring open sources in the Piauí context. The results showed high precision (94.9%) and limited recall (42.5%), with an F1 score of 58.7%. These figures reveal the feasibility of the approach as a complementary signal. At the same time, they expose relevant local challenges, such as the instability of automated collection, linguistic noise, and the heuristic nature of geolocation. It is concluded that the study contributes less as a finished solution and more as a critical characterization of the conditions, possibilities, and limits of using Digital Epidemiology over open sources for arbovirus surveillance in Piauí.

Keywords: digital epidemiology; natural language processing; named entity recognition; *web scraping*; arboviruses.

Data de aprovação: 09/07/2026

1 INTRODUÇÃO

No mundo contemporâneo, a sociedade vivencia a era do *Big Data*, caracterizada pela produção massiva e contínua de informações em ambientes virtuais (Mayer-Schönberger; Cukier, 2014). Diariamente, plataformas de notícias e redes sociais acumulam um vasto repositório de dados não estruturados. Esses dados refletem o cotidiano, os interesses e, inevitavelmente, o estado de saúde da população. Tal fenômeno viabilizou o surgimento de novas abordagens analíticas. Por meio da mineração automatizada desses registros digitais, a Ciência da Computação passou a oferecer soluções para problemas complexos da saúde pública (Salathé, 2018).

No âmbito da vigilância sanitária, o modelo tradicional de monitoramento, baseado em notificações clínicas oficiais, é indispensável, mas responde de forma relativamente lenta ao surgimento de surtos. Em paralelo, a infovigilância, ou Epidemiologia Digital, surge como um paradigma complementar, ao explorar o fato de que surtos epidêmicos frequentemente repercutem na internet antes mesmo de serem contabilizados pelos sistemas oficiais (Brownstein; Freifeld; Madoff, 2009).

Este trabalho explora a interseção entre técnicas de Inteligência Artificial e a gestão em saúde, aplicando-as ao cenário das arboviroses (dengue, zika e chikungunya) no estado do Piauí. O recorte local não é trivial. O estado combina um vasto interior, cobertura desigual de conectividade, marcada variação linguística regional e produção jornalística concentrada em poucos veículos. Esses fatores tornam a extração automatizada de informação geolocalizada um problema particular, e não uma simples transposição de soluções concebidas para grandes centros. A capacidade de processar a linguagem natural para extrair informação de textos jornalísticos e de boletins oficiais apresenta-se, assim, como uma oportunidade de modernizar a percepção situacional dos agravos de saúde (Jurafsky; Martin, 2025).

Vale registrar que o Piauí vem consolidando um ecossistema de Inteligência Artificial aplicada ao interesse público. Dele fazem parte iniciativas como o CIATEN, voltado à inteligência epidemiológica de agravos negligenciados, e, no âmbito do governo estadual, a plataforma SoberanIA e sua primeira aplicação, o BO Fácil. Esse contexto, detalhado no referencial teórico, demonstra maturidade técnica e institucional. Resta, porém, um nicho específico em aberto: o aproveitamento de fontes abertas, como notícias e boletins, para a extração geolocalizada de menções a arboviroses. É

essa lacuna que o presente trabalho procura caracterizar e investigar.

Diante disso, este trabalho propôs e avaliou criticamente uma metodologia para automatizar o ciclo de inteligência de dados: da coleta automatizada (*web scraping*) em fontes públicas, passando pelo reconhecimento de entidades (doenças, sintomas e localidades) por meio do Processamento de Linguagem Natural, até a disponibilização dos resultados em um painel com mapa interativo. Essa metodologia foi materializada em um protótipo funcional, tratado como prova de conceito, e não como sistema de vigilância em produção. Convém delimitar o objeto de avaliação: o que se examina neste trabalho é o comportamento do *pipeline* de Processamento de Linguagem Natural, isto é, sua capacidade de identificar e geolocalizar menções a arboviroses e os erros que comete, e não a entrega de um sistema pronto para uso. Cabe ainda a ressalva de que a coleta automatizada se mostrou instável no contexto local, frequentemente bloqueada pelas fontes; por isso, a avaliação foi conduzida sobre um corpus curado de textos verificados, e a própria instabilidade do *web scraping* constitui um dos achados do trabalho, e não um pressuposto atendido. Assim, mais do que entregar uma solução acabada, o propósito foi caracterizar os desafios locais do problema e analisar a viabilidade técnica e as limitações do uso de fontes abertas para compor um cenário de vigilância proativa no contexto piauiense.

Além desta introdução, o trabalho apresenta o problema e os objetivos da pesquisa, a justificativa do estudo, o referencial teórico que o fundamenta, os procedimentos metodológicos e a ferramenta desenvolvida, os resultados obtidos e a respectiva discussão e, por fim, as considerações finais.

1.1 PROBLEMA DE PESQUISA

O problema de pesquisa constitui o elemento central que orienta o desenvolvimento do estudo e, para ser investigável, deve ser formulado de maneira clara, precisa e operacional (Gil, 2019). No caso desta pesquisa, observa-se que a vigilância epidemiológica tradicional, ainda que essencial, opera de forma reativa e está sujeita a atrasos de notificação e à subnotificação, o que dificulta a percepção precoce de surtos de arboviroses no estado do Piauí.

Ao mesmo tempo, portais de notícias e boletins oficiais publicam informações sobre localidades afetadas de forma não estruturada. Contudo, há uma lacuna na validação empírica sobre como algoritmos de extração se comportam diante da variação linguística e estrutural dessas fontes no cenário piauiense. Isso evidencia a necessidade de investigar não apenas a viabilidade de desenvolver a ferramenta, mas sobretudo o comportamento, a precisão e os limites de um pipeline de PLN aplicado a esse domínio. Diante desse cenário, formula-se a seguinte questão de pesquisa:

Em que medida uma abordagem baseada em Processamento de Linguagem Natural e *web scraping* é capaz de avaliar, com precisão, as menções a

arboviroses e suas geolocalizações em fontes públicas online do Piauí?

Responder a essa questão implica avaliar tanto a capacidade técnica de extrair automaticamente as menções, associando cada doença ao respectivo município, quanto a confiabilidade dessa extração, mensurada por indicadores de desempenho consagrados na literatura de Recuperação da Informação, como a precisão, a revocação e o F1-score.

1.2 OBJETIVOS

Este capítulo apresenta o objetivo geral e os objetivos específicos que orientaram o desenvolvimento do trabalho. Para cada objetivo específico, indica-se a etapa em que foi alcançado.

1.2.1 Objetivo geral

Propor, implementar em caráter de prova de conceito e avaliar criticamente uma metodologia de Epidemiologia Digital para identificar e geolocalizar menções a arboviroses em fontes públicas online no estado do Piauí, com ênfase na caracterização dos desafios locais e na análise da viabilidade técnica e das limitações da abordagem.

1.2.2 Objetivos específicos

Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos, todos cumpridos ao longo deste trabalho:

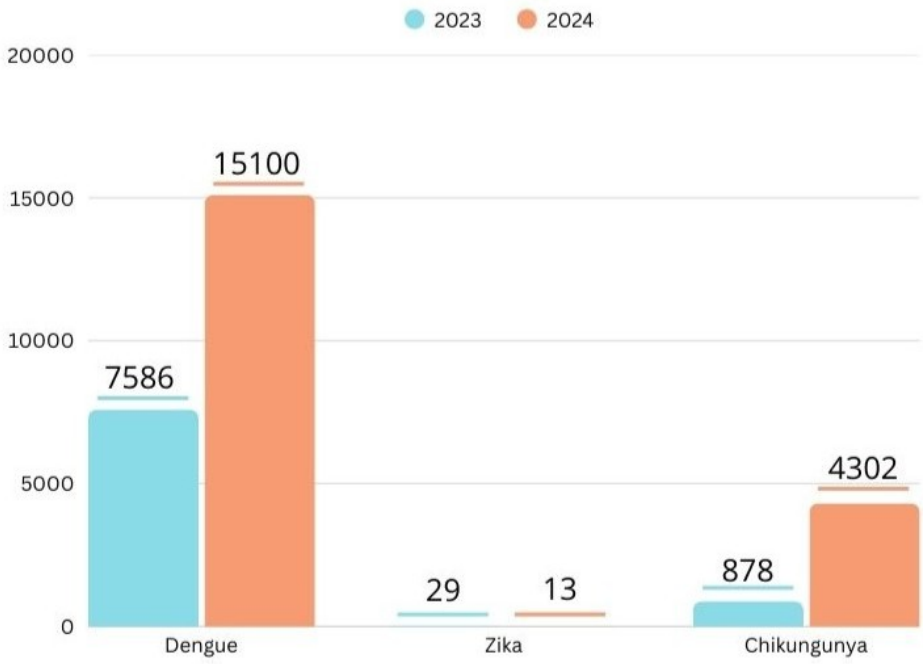
1. fundamentar o estudo por meio de revisão bibliográfica sobre Epidemiologia Digital, *web scraping* e Processamento de Linguagem Natural aplicados à saúde pública (seção 2);
2. desenvolver um módulo de coleta e armazenamento de dados para obter notícias e boletins de fontes públicas do Piauí e caracterizar a estabilidade dessa coleta no contexto local (seção 3);
3. implementar um módulo de análise com Processamento de Linguagem Natural para identificar e estruturar menções de arboviroses e suas localidades (seção 3);
4. construir um protótipo de painel web para visualizar os padrões identificados por meio de mapa e gráficos (seção 3);
5. avaliar criticamente a abordagem, mensurando seu desempenho na identificação automática de doenças e localidades e analisando a viabilidade técnica, os desafios locais e as limitações observadas no contexto piauiense (seção 4).

1.3 JUSTIFICATIVA

A vigilância epidemiológica é uma ferramenta estratégica para a saúde pública, e sua importância acentua-se diante de desafios endêmicos como o enfrentado pelo

estado do Piauí no combate às arboviroses. Embora o Brasil enfrente surtos recorrentes de dengue, zika e chikungunya, essa realidade manifesta-se de forma particularmente preocupante no cenário piauiense. Segundo os boletins epidemiológicos da Secretaria de Estado da Saúde do Piauí (SESAPI), o estado registrou em 2024 um aumento expressivo no número de casos, com destaque para a dengue, sobrecarregando os serviços de saúde, sobretudo em municípios-polo como Teresina e Floriano (Brasil, 2024). A Figura 1 ilustra a evolução dos casos de arbovirose no estado ao longo de 2024.

Figura 1 – Casos de arbovirose no Piauí em 2024



Fonte: SESAPI (2024).

A resposta a esse cenário apoia-se, fundamentalmente, no Sistema de Informação de Agravos de Notificação (SINAN). Embora essencial para a consolidação de dados robustos, esse modelo apresenta limitações que comprometem a agilidade da resposta a surtos. A primeira é o atraso temporal entre o início dos sintomas na população e a disponibilização dos dados consolidados aos gestores, decorrente do fluxo de procura por atendimento, diagnóstico e digitação da notificação (Antunes et al., 2014). A segunda é a subnotificação, fenômeno no qual casos leves ou assintomáticos não chegam a entrar nas estatísticas oficiais (Leandro et al., 2020). Em conjunto, o atraso na análise e a invisibilidade de parte dos casos criam uma janela de vulnerabilidade, na qual uma doença pode disseminar-se antes que sua real dimensão seja conhecida pelas autoridades de saúde.

Em contrapartida, a crescente digitalização da sociedade abre uma fronteira promissora para a vigilância em saúde. Diariamente, um grande volume de dados

públicos é gerado em tempo real por cidadãos e pela mídia, contendo relatos e discussões que podem funcionar como sinais da saúde coletiva (Eysenbach, 2009). Essa fonte, objeto de estudo da Epidemiologia Digital, oferece um contraponto direto às lacunas do sistema oficial: sua velocidade ajuda a combater o atraso na notificação e seu alcance contribui para mitigar a subnotificação (Gomide et al., 2011). Nesse contexto, o Processamento de Linguagem Natural (PLN) apresenta-se como a tecnologia-chave para transformar esse volume de dados não estruturados em informação acionável, identificando padrões como a correlação entre menções a doenças e localidades geográficas específicas (Vieira et al., 2025).

Apesar da vasta literatura sobre o uso de fontes digitais para a detecção de surtos epidêmicos (Charles-Smith et al., 2015), identifica-se uma lacuna quanto a soluções focadas especificamente no estado do Piauí que integrem, em uma única abordagem, a coleta automatizada em fontes abertas, o processamento de linguagem natural e a visualização geográfica. Mesmo diante de um ecossistema local já em consolidação, as iniciativas existentes operam predominantemente sobre dados estruturados e oficiais, o que mantém em aberto o aproveitamento sistemático de notícias e boletins públicos. Para contribuir com o preenchimento dessa lacuna, este trabalho investigou e caracterizou o problema, propondo uma abordagem que aplica *web scraping* e PLN a notícias e a boletins oficiais do estado para identificar correlações entre menções a arboviroses e os municípios em que ocorrem.

A relevância do estudo não se restringe ao artefato técnico. Ao avaliar criticamente a abordagem e ao documentar seus acertos, erros e limitações no contexto piauiense, o trabalho gera conhecimento metodológico e didático sobre as condições reais de aplicação da Epidemiologia Digital no estado, oferecendo subsídios para que pesquisas futuras repliquem, critiquem e aprimorem a proposta. Essa contribuição soma-se ao valor técnico de demonstrar a viabilidade da extração geolocalizada de menções a partir de fontes abertas.

A viabilidade do trabalho sustentou-se na disponibilidade de dados públicos e no uso de tecnologias de código aberto. A opção por um protótipo de escopo delimitado, concebido como prova de conceito, tornou a execução compatível com o cronograma acadêmico e permitiu aplicar, na prática, os conhecimentos de desenvolvimento de sistemas e de inteligência artificial adquiridos ao longo da formação tecnológica.

2 REFERENCIAL TEÓRICO

Apresentam-se os fundamentos da pesquisa em Design Science, que orienta a construção do artefato; a Epidemiologia Digital e suas relações com a vigilância em saúde; as técnicas de coleta automatizada de dados na web; o Processamento de Linguagem Natural e, em particular, a tarefa de Reconhecimento de Entidades Nomeadas, central para a extração de menções epidemiológicas a partir de textos.

2.1 EPIDEMIOLOGIA DIGITAL E VIGILÂNCIA EM SAÚDE

A Epidemiologia Digital, ou infovigilância, é um campo de estudo emergente que utiliza dados gerados na internet para complementar a vigilância em saúde pública. Conforme Eysenbach (2009), essa abordagem analisa informações de fontes como mídias sociais, notícias e motores de busca para rastrear doenças e monitorar a saúde coletiva. Ela se contrapõe às limitações do sistema de vigilância tradicional, como o Sistema de Informação de Agravos de Notificação (SINAN), que, embora robusto, opera de forma reativa. É nesse contexto que se situa este trabalho, ao utilizar dados digitais públicos para gerar sinais de alerta que o sistema oficial, por sua natureza, não capta em tempo real. Em escala global, a BlueDot, plataforma canadense de inteligência de surtos baseada em inteligência artificial que minera notícias, relatórios oficiais e dados de aviação comercial, ilustra esse potencial preditivo: segundo Bogoch et al. (2020), a mineração automatizada de fontes abertas sinalizou o aglomerado de pneumonia de causa desconhecida em Wuhan, na China, ainda em dezembro de 2019, cerca de uma semana antes do comunicado público da Organização Mundial da Saúde. Tratava-se dos primeiros casos do novo coronavírus (SARS-CoV-2), causador da COVID-19, cuja disseminação internacional pelas rotas aéreas a plataforma chegou a estimar com antecedência. O caso evidencia como fontes digitais abertas podem antecipar sinais de agravos, ainda que a confirmação epidemiológica permaneça a cargo dos sistemas oficiais de vigilância. O Quadro 1 sintetiza as principais diferenças entre a vigilância tradicional e a Epidemiologia Digital.

Quadro 1 – Comparativo entre a vigilância tradicional e a Epidemiologia Digital

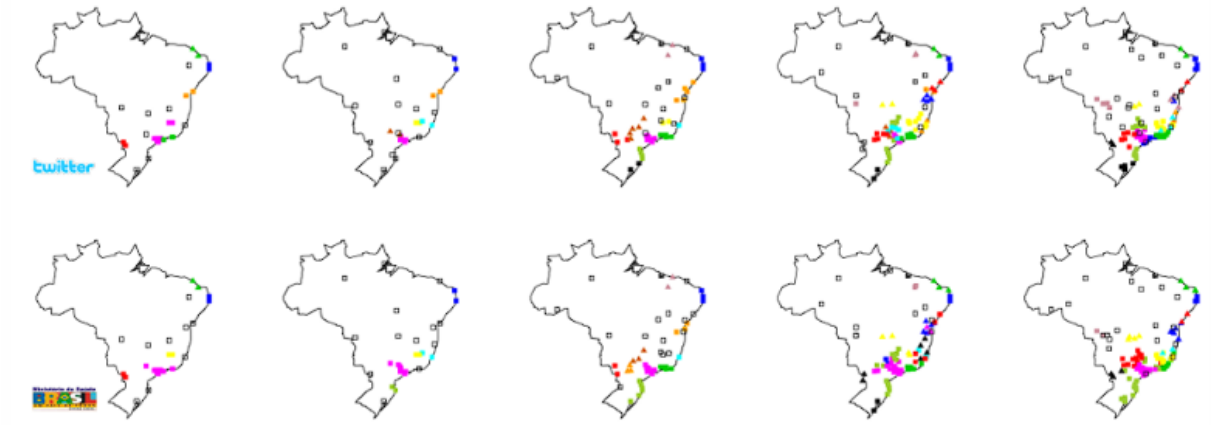
Vigilância tradicional	**Epidemiologia Digital (infovigilância)**
Baseada em notificações clínicas e laboratoriais.	Baseada em conteúdo gerado pelo usuário e em notícias.
Alta especificidade (casos confirmados).	Alta sensibilidade, porém sujeita a ruído (rumores).
Fluxo lento (semanas ou meses para consolidação).	Fluxo em tempo real ou quase real (horas).
Custo elevado (infraestrutura de saúde).	Custo baixo (reúso de dados já disponíveis na web).

Fonte: Elaborado pelo autor (2026), com base em Salathé (2018) e Eysenbach (2009).

Um traço recorrente da Epidemiologia Digital é a visualização geográfica dos sinais coletados, que traduz relatos dispersos em um panorama espacial dos agravos. Dois trabalhos ilustram bem esse ponto e antecedem, em espírito, a abordagem aqui proposta. O sistema HealthMap, descrito por Brownstein, Freifeld e Madoff (2009), coleta relatos de fontes abertas na web e os exibe geolocalizados em um mapa mundial

de surtos, em tempo quase real. Mais próximo do contexto nacional, Gomide et al. (2011) realizaram vigilância de dengue no Brasil a partir do Twitter, com análise espaço-temporal que correlaciona a atividade nas redes à incidência da doença por região. A Figura 2 reproduz a visualização espacial proposta por esses autores, que guarda semelhança direta com o mapa por município adotado neste trabalho, ainda que a partir de outra fonte de dados (mídia social, e não notícias e boletins).

Figura 2 – Evolução espaço-temporal da dengue no Brasil: menções no Twitter (linha superior) e dados oficiais de vigilância (linha inferior)



Fonte: Gomide et al. (2011).

2.2 COLETA DE DADOS: WEB SCRAPING E FONTES PÚBLICAS

A obtenção de dados textuais a partir de páginas da internet apoia-se na técnica de *web scraping*, um processo automatizado de extração de informações de páginas web (Mitchell, 2018). Aplicada a portais de notícias, ela percorre o HTML das páginas e recupera o título, a data e o corpo das matérias, etapa usualmente apoiada por bibliotecas como a BeautifulSoup. Além das fontes jornalísticas, a literatura de Epidemiologia Digital também explora o monitoramento de mídias sociais, nas quais relatos espontâneos podem antecipar sinais de surtos (Gomide et al., 2011). Essas fontes, contudo, apresentam desafios de geolocalização e de acesso que extrapolam o escopo deste trabalho, voltado a fontes jornalísticas e a boletins oficiais, nas quais a referência geográfica é explícita e rastreável.

2.3 PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

Os dados coletados encontram-se em formato de texto livre, isto é, como dados não estruturados. Para extrair deles valor analítico, recorre-se ao Processamento de Linguagem Natural (PLN), subárea da Inteligência Artificial voltada a capacitar computadores a processar e compreender a linguagem humana, transformando o texto bruto em dados estruturados e analisáveis (Jurafsky; Martin, 2025). No âmbito

deste trabalho, o PLN é a tecnologia central que permite interpretar, em escala, o conteúdo das notícias coletadas e identificar padrões relevantes, como a associação entre menções de doenças e localidades.

2.4 RECONHECIMENTO DE ENTIDADES NOMEADAS (NER)

Entre as tarefas do PLN, a central nesta pesquisa é o Reconhecimento de Entidades Nomeadas (NER), processo de extração de informação que localiza e classifica entidades em um texto segundo categorias predefinidas, como nomes de pessoas, locais e organizações (Nadeau; Sekine, 2007). Sua principal vantagem é a possibilidade de reconhecer entidades customizadas, recurso explorado neste trabalho para identificar três categorias: DOENÇA (por exemplo, dengue), SINTOMA (por exemplo, febre alta) e MUNICÍPIO (por exemplo, um município do Piauí). O Quadro 2 ilustra como o modelo anota um trecho de notícia dentro desse escopo.

Quadro 2 – Exemplo de anotação semântica com as entidades do escopo

A Secretaria de Saúde informou que o município de Picos [MUNICÍPIO] registrou novos casos confirmados de chikungunya [DOENÇA]. Os pacientes relataram fortes dores articulares [SINTOMA] e manchas na pele [SINTOMA].
Legenda: DOENÇA, SINTOMA, MUNICÍPIO.

Fonte: Elaborado pelo autor (2026), com base em Jurafsky e Martin (2025).

Além da identificação, o NER cumpre papel essencial na estruturação dos dados: após a extração, descartam-se os termos comuns da língua (artigos, preposições e verbos de ligação) e mantêm-se apenas as entidades classificadas, convertendo o texto livre em registros computáveis e passíveis de agregação estatística. O Quadro 3 detalha esse processo de filtragem, mostrando como o algoritmo percorre sequencialmente o texto do exemplo anterior e separa o conteúdo irrelevante das entidades de interesse.

Quadro 3 – Processamento sequencial do texto pelo algoritmo de NER

Segmento de texto	**Classificação atribuída**
A Secretaria de Saúde informou que o município de	O (outro / irrelevante)
Picos	MUNICÍPIO
registrou novos casos confirmados de	O (outro / irrelevante)
chikungunya	DOENÇA
Os pacientes relataram fortes	O (outro / irrelevante)
dores articulares	SINTOMA
e	O (outro / irrelevante)
manchas na pele	SINTOMA

Fonte: Elaborado pelo autor (2026), com base na taxonomia de Nadeau e Sekine (2007).

2.5 PESQUISA EM DESIGN SCIENCE

A construção de sistemas computacionais como forma de produzir conhecimento científico é o cerne da *Design Science Research* (DSR), ou pesquisa de *design*. Diferentemente do paradigma das ciências comportamentais, que busca explicar e prever fenômenos, a DSR busca resolver problemas por meio da criação e da avaliação de artefatos, como modelos, métodos e sistemas (Hevner et al., 2004). O conhecimento, nessa abordagem, emerge tanto do artefato construído quanto do processo rigoroso de sua avaliação.

Para operacionalizar esse paradigma, adota-se a *Design Science Research Methodology* (DSRM), proposta por Peffers et al. (2007), que organiza a pesquisa em seis etapas: identificação do problema e motivação; definição dos objetivos da solução; *design* e desenvolvimento do artefato; demonstração; avaliação; e comunicação. Esse encadeamento oferece um roteiro sistemático que articula a relevância prática do problema, o rigor da construção e a avaliação empírica dos resultados, conciliando a produção de um artefato útil com a geração de conhecimento verificável.

2.6 INICIATIVAS CORRELATAS NO CONTEXTO PIAUIENSE

A discussão sobre Inteligência Artificial aplicada ao interesse público no Piauí não parte de um vazio institucional. Ao contrário, o estado vem consolidando um ecossistema de pesquisa e de serviços apoiados em ciência de dados e em PLN, do qual se destacam três iniciativas. O Centro de Inteligência em Agravos Tropicais Emergentes e Negligenciados (CIATEN), vinculado à Universidade Federal do Piauí e apoiado pela Fundação de Amparo à Pesquisa do Estado do Piauí, dedica-se à inteligência epidemiológica de agravos como leishmaniose visceral, dengue, chikungunya e zika, com forte emprego de geoprocessamento, modelagem preditiva e

aprendizado de máquina sobre dados estruturados e oficiais (UFPI, 2025). No âmbito do governo estadual, a plataforma SoberanIA, lançada em 2025, constitui o primeiro modelo de linguagem treinado por um estado brasileiro, concebido para compreender a diversidade linguística e cultural nacional, incluindo sotaques e expressões regionais (Piauí, 2025b). Sua primeira aplicação prática, o BO Fácil, emprega PLN para interpretar relatos coloquiais de cidadãos, registrados por texto ou áudio, no domínio da segurança pública (Piauí, 2025a).

Esse panorama evidencia maturidade técnica e institucional, mas também delimita, por contraste, a lacuna que este trabalho procura ocupar. As iniciativas existentes operam sobre dados estruturados e oficiais (caso do CIATEN) ou sobre texto submetido diretamente pelo cidadão em outro domínio (caso do BO Fácil); nenhuma delas combina, de forma específica, a coleta automatizada em fontes abertas (notícias e boletins públicos), o Reconhecimento de Entidades Nomeadas e a geolocalização heurística de menções a arboviroses. O Quadro 4 posiciona este trabalho frente a essas iniciativas. Cabe registrar que SoberanIA e BO Fácil foram lançados em 2025, ao passo que o corpus avaliado nesta pesquisa é anterior; tais iniciativas são, portanto, contexto institucional, e não objeto de comparação empírica. Ainda assim, a existência de um modelo de linguagem sensível a sotaques regionais sinaliza um caminho promissor para mitigar, em trabalhos futuros, o ruído linguístico que afeta a abordagem aqui proposta.

Quadro 4 – Posicionamento do trabalho frente a iniciativas correlatas no Piauí

Iniciativa	**Dados de entrada**	**Técnica predominante**	**Domínio**
CIATEN	Estruturados e oficiais (clínicos)	Modelagem preditiva e geoprocessamento	Agravos e arboviroses
SoberanIA	Dados governamentais e corpora em português	Modelo de linguagem generalista	Serviços públicos
BO Fácil	Texto coloquial submetido pelo cidadão	PLN sobre relatos	Segurança pública
Este trabalho	Fontes abertas (notícias e boletins)	*Web scraping*, NER e geolocalização heurística	Arboviroses

Fonte: Elaborado pelo autor (2026), com base em UFPI (2025), Piauí (2025b) e Piauí (2025a).

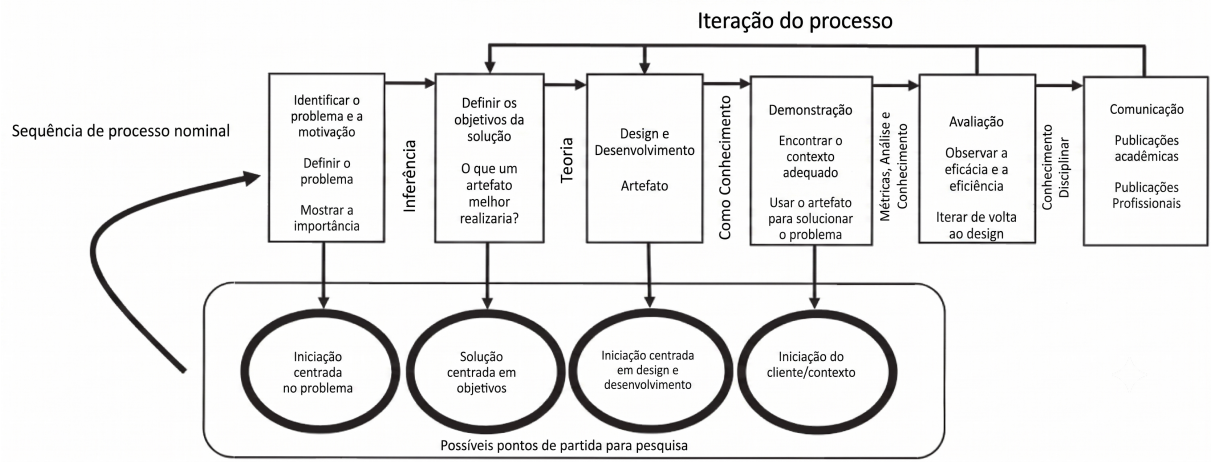
3 METODOLOGIA

Este capítulo descreve os procedimentos metodológicos adotados na construção e na avaliação do protótipo de Epidemiologia Digital desenvolvido neste trabalho. Apresenta a caracterização da pesquisa, a ferramenta desenvolvida e seus módulos, a delimitação e a construção do corpus, a construção do conjunto de referência (Padrão-Ouro), o procedimento de avaliação quantitativa e os aspectos éticos.

3.1 CARACTERIZAÇÃO DA PESQUISA

Quanto à natureza, a pesquisa classifica-se como aplicada, por produzir um artefato tecnológico, um protótipo de software concebido como instrumento de investigação, voltado a um problema prático de vigilância em saúde. Como estratégia metodológica, adota-se a *Design Science Research* (DSR), ou pesquisa de *design*, paradigma voltado à geração de conhecimento por meio da construção e da avaliação de artefatos que resolvem problemas reais (Hevner et al., 2004). Nessa perspectiva, o software não é um fim em si, mas o meio para produzir conhecimento sobre o problema e sobre as condições de aplicação da solução. O trabalho estrutura-se segundo as seis etapas da *Design Science Research Methodology* (DSRM) proposta por Peffers et al. (2007), sintetizadas no Figura 3. Quanto aos objetivos, é exploratória, ao investigar a aplicação de técnicas de PLN ao contexto específico do Piauí, e descritiva, ao registrar e analisar as etapas de desenvolvimento e o desempenho do sistema. Quanto à abordagem, é mista (quali-quantitativa): qualitativa na interpretação semântica dos textos e na análise dos erros de extração, e quantitativa na mensuração estatística do desempenho do modelo e na agregação das menções extraídas.

Figura 3 – Modelo de processo do DSRM



Fonte: Adaptado de Peffers et al. (2007).

As seis etapas do DSRM foram percorridas neste trabalho da seguinte forma. Na (1) identificação do problema e motivação, partiu-se da vigilância de arboviroses no

Piauí, sujeita a atraso de notificação e subnotificação, e da dispersão das fontes abertas. Na (2) definição dos objetivos da solução, estabeleceu-se identificar e geolocalizar automaticamente menções a arboviroses em notícias e boletins públicos. No (3) design e desenvolvimento, construiu-se o artefato EpiPiauí Monitor, com módulo de coleta, *pipeline* de PLN (NER por regras e associação por co-ocorrência) e painel com mapa. Na (4) demonstração, o artefato foi executado sobre o corpus de notícias e boletins do Piauí de 2024. Na (5) avaliação, o desempenho foi mensurado contra o Padrão-Ouro pelas métricas de precisão, revocação e F1, complementadas pela análise de erros e pela ablação da janela de associação. Por fim, a (6) comunicação materializa-se nesta monografia, que registra o artefato, os resultados, as limitações e os desafios locais caracterizados. Quanto aos pontos de partida previstos no modelo, esta pesquisa é do tipo iniciada pelo problema.

Cabe destacar que o produto resultante é tratado como instrumento metodológico de investigação, e não como sistema de vigilância em produção. O protótipo não substitui os sistemas oficiais de vigilância epidemiológica, não realiza diagnóstico e não deve constituir fonte única para decisões em saúde pública.

3.2 A FERRAMENTA EPIPIAUÍ MONITOR

Esta seção descreve a ferramenta desenvolvida, o EpiPiauí Monitor, em termos de sua arquitetura e de seus módulos. O sistema implementa um fluxo de extração, transformação e carga acoplado a um visualizador, organizado em três camadas: coleta; processamento e persistência; e visualização.

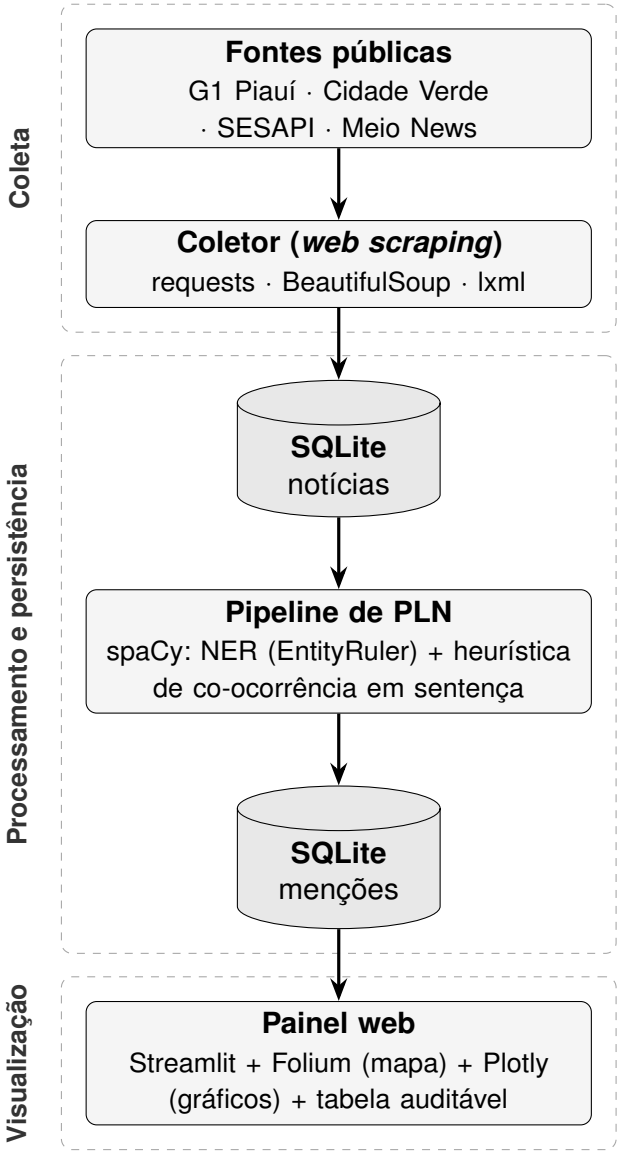
3.2.1 Visão geral e arquitetura

O desenvolvimento utilizou a linguagem Python (versão 3.11) e bibliotecas de código aberto. O reconhecimento de entidades apoiou-se na biblioteca spaCy, com o modelo *pt_core_news_lg* e mecanismo de contingência para os modelos *md*, *sm* e, em última instância, um pipeline em branco com segmentador de sentenças. A coleta e a análise de páginas empregaram as bibliotecas *requests*, *BeautifulSoup* e *lxml*; a persistência foi feita em *SQLite*; a manipulação tabular, com *pandas*; e a interface de visualização, com *Streamlit*, integrado ao *Folium* e ao *Plotly*. A suíte de testes automatizados foi escrita com *pytest*.

A arquitetura segue um pipeline clássico de extração, transformação e carga, dividido em três camadas. A camada de coleta obtém os textos das fontes; a camada de processamento e persistência aplica o Processamento de Linguagem Natural e grava os resultados em banco de dados; e a camada de visualização apresenta os dados em um painel web interativo. A visualização foi implementada por meio do Streamlit integrado ao Folium, biblioteca que encapsula o Leaflet, o que facilitou a

integração entre o processamento em Python e a interface web. A Figura 4 sintetiza essa organização e o fluxo dos dados entre as três camadas.

Figura 4 – Arquitetura e fluxo de dados do EpiPiauí Monitor



Fonte: Elaborada pelo autor (2026).

3.2.2 Módulo de coleta

O módulo de coleta de sementes parte de uma lista de URLs reais verificadas e tenta baixar o HTML original de cada matéria no momento da execução. Quando o download não é possível, por exemplo quando o domínio cidadeverde.com responde com bloqueio (HTTP 403, via Cloudflare) à coleta automatizada, o sistema recorre a um texto de reserva previamente verificado e marca o registro como reserva utilizada. Esse mecanismo tem dupla função e deve ser lido nas duas faces. Por um lado, ele preserva a rastreabilidade e assegura a reprodutibilidade da avaliação, condição metodológica

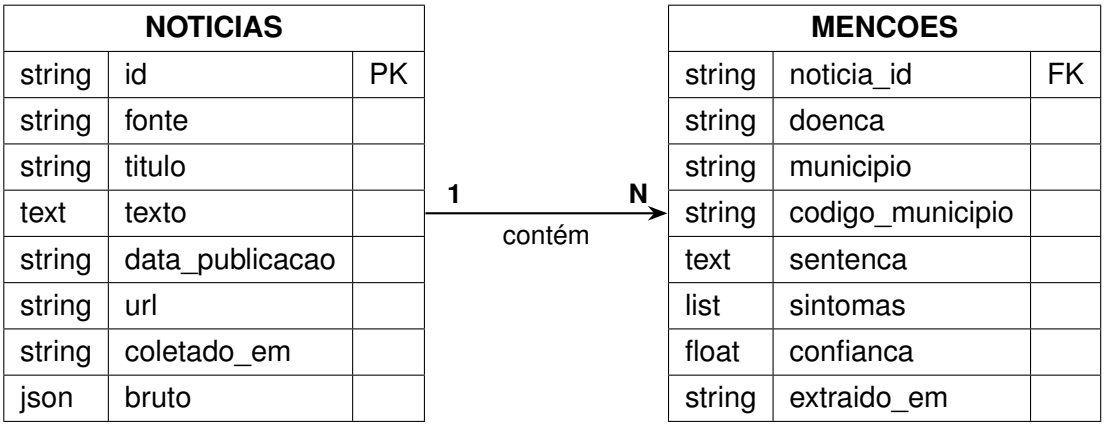
necessária. Por outro, sua própria necessidade constitui evidência empírica de uma limitação real do contexto: a instabilidade da coleta automatizada em fontes públicas do Piauí, frequentemente bloqueada ou sujeita a páginas com estrutura instável. Esse comportamento, longe de ser apenas um detalhe de implementação, é um dos desafios locais que a pesquisa se propôs a caracterizar. Um módulo de coleta ao vivo, configurado para quatro fontes (G1 Piauí, Cidade Verde, SESAPI e Meio News), foi implementado e mantido no sistema, mas não constituiu a base da avaliação, para garantir a estabilidade dos resultados.

3.2.3 Estruturação e persistência dos dados

A persistência foi implementada em um banco de dados relacional SQLite, escolhido por ser embutido, sem servidor e adequado ao escopo local da pesquisa. O formato JSON ficou reservado ao armazenamento bruto das sementes e às exportações, e o uso de um banco relacional conferiu integridade referencial, deduplicação e consultas mais eficientes.

O banco organiza-se em duas tabelas. A tabela de notícias preserva o conteúdo bruto de cada matéria (fonte, título, texto, data de publicação, URL e metadados de coleta), garantindo a rastreabilidade até o documento original. A tabela de menções registra cada menção extraída pelo PLN, vinculada à respectiva notícia por chave estrangeira e contendo a doença, o município (com o código do IBGE), a sentença de origem, os sintomas identificados, o escore de confiança e o instante de extração. Cada par (doença, município) identificado em uma sentença foi atomizado em um registro independente, de modo que uma única matéria que cite múltiplos municípios ou agravos gera múltiplas menções distintas. A duplicação foi evitada por uma restrição de unicidade, e a integridade entre menções e notícias foi assegurada por chave estrangeira com exclusão em cascata. A Figura 5 representa esse modelo de dados, com as duas tabelas e o relacionamento que as vincula.

Figura 5 – Modelo de dados do EpiPiauí Monitor



Fonte: Elaborada pelo autor (2026).

3.2.4 Pipeline de PLN e reconhecimento de entidades (NER)

O núcleo analítico foi implementado com a biblioteca spaCy. O texto de cada notícia, composto pela concatenação de título e corpo, foi submetido à segmentação em sentenças e ao reconhecimento de entidades. O NER é a tarefa de extração de informação que localiza e classifica entidades em categorias predefinidas (Nadeau; Sekine, 2007). A identificação apoiou-se em um componente *EntityRuler*, configurado com correspondência sobre a forma em caixa baixa e com sobreposição às entidades do modelo base, o que privilegia regras explícitas e auditáveis em vez de predições probabilísticas opacas (Jurafsky; Martin, 2025). Cabe delimitar o papel do modelo pré-treinado em português *pt_core_news_lg* nesse arranjo: ele atua como infraestrutura linguística de base, responsável pela tokenização e pela segmentação do texto em sentenças, e apenas subsidiariamente como reconhecedor de entidades genéricas. O reconhecimento das entidades de interesse desta pesquisa (DOENÇA, MUNICÍPIO e SINTOMA) não provém do modelo estatístico, mas das regras do *EntityRuler*, que têm precedência sobre as predições do modelo. Em outras palavras, o modelo em português fornece o processamento linguístico fundamental, ao passo que a extração propriamente dita é governada pelo vocabulário do domínio.

A opção por um reconhecimento baseado em regras, em vez de um modelo estatístico treinado ou de um grande modelo de linguagem (LLM), foi uma decisão metodológica coerente com o caráter de prova de conceito do trabalho, sustentada por quatro razões. A primeira é a auditabilidade: cada menção extraída pode ser rastreada até a regra que a produziu, o que torna acertos e erros plenamente explicáveis, requisito central em uma pesquisa voltada à caracterização do problema. A segunda é o baixo custo computacional, pois a abordagem por regras dispensa unidades de processamento gráfico e infraestrutura de treino ou de inferência de modelos de grande porte, executando em hardware modesto. A terceira é a independência de dados rotulados: a construção de um modelo supervisionado exigiria um corpus de treino anotado no nível do texto, com as menções demarcadas em cada sentença, inexistente para notícias de arboviroses no Piauí. Os dados estruturados disponíveis, como boletins e contagens oficiais, não suprem essa carência, pois fornecem resultados agregados, e não anotações posicionais no texto jornalístico. As regras, por sua vez, partem diretamente do vocabulário do domínio. A quarta é a reprodutibilidade, já que regras determinísticas produzem sempre o mesmo resultado para a mesma entrada, sem a variabilidade típica de modelos generativos. Esses atributos fazem das regras o ponto de partida adequado para estabelecer uma linha de base interpretável; a incorporação de modelos de linguagem treinados para o português, sensíveis à variação regional, constitui evolução natural para etapas futuras, quando houver corpus anotado e infraestrutura disponíveis.

Foram definidas três categorias de entidades: DOENÇA, MUNICÍPIO e SIN-

TOMA. As doenças e suas variantes textuais constam do Quadro 5. Os municípios corresponderam à lista oficial dos 224 municípios do Piauí, obtida da base de localidades do IBGE, com correspondência tolerante à acentuação. Os sintomas compreenderam um conjunto de termos clínicos associados às arboviroses, como febre, dor de cabeça, dores articulares, manchas vermelhas e exantema. Cabe uma precisão terminológica: o reconhecimento de entidades adotado é baseado em regras e listas de termos (*gazetteer*), e não em um modelo estatístico aprendido; e a geolocalização consiste em associar a menção a um município previamente cadastrado na lista do IBGE, sem inferência de coordenadas geográficas a partir do texto livre.

Quadro 5 – Doenças monitoradas e principais variantes reconhecidas

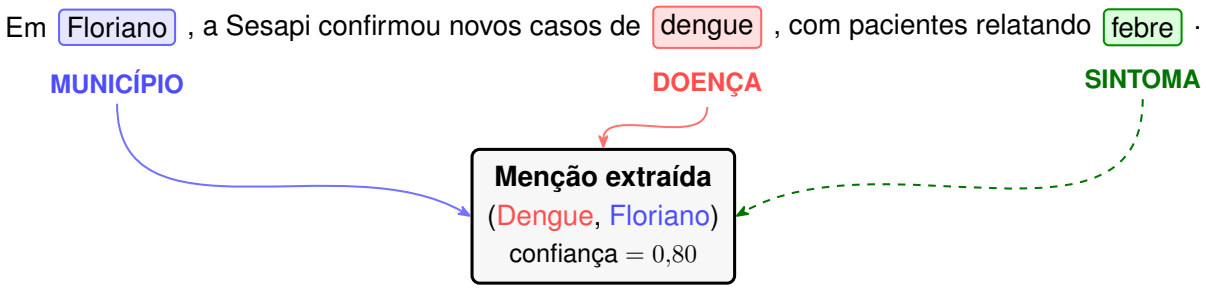
Doença	**Variantes reconhecidas**
Dengue	dengue; dengue grave; dengue com sinais de alarme; arbovirose dengue
Zika	zika; vírus zika; febre do zika
Chikungunya	chikungunya; febre chikungunya

Fonte: Elaborado pelo autor (2026). Nota: relação ilustrativa das principais variantes; o vocabulário completo consta do arquivo de configuração do domínio.

3.2.5 Heurística de co-ocorrência e escore de confiança

Para converter entidades isoladas em informação útil à vigilância, aplicou-se uma heurística de co-ocorrência em escopo de sentença. A unidade de análise foi a sentença: somente quando uma entidade DOENÇA e uma entidade MUNICÍPIO ocorrem na mesma sentença estabelece-se o vínculo, gerando uma menção. Quando uma sentença contém n doenças e m municípios, o sistema gera o produto cartesiano das $n \times m$ combinações; menções idênticas dentro de uma mesma notícia são descartadas. Essa regra é deliberadamente conservadora: privilegia a rastreabilidade e a precisão, ao custo de não capturar associações distribuídas por sentenças diferentes. A Figura 6 ilustra o funcionamento da heurística sobre uma sentença exemplo, e o Algoritmo 1 a formaliza.

Figura 6 – Exemplo de reconhecimento de entidades e associação por co-ocorrência



Fonte: Elaborado pelo autor (2026).

Algoritmo 1: Associação doença–município por co-ocorrência sentencial

Entrada: notícia n; vocabulários D (doenças), M (municípios), S (sintomas)
Saída: conjunto de menções R

$R \leftarrow \varnothing$
$tituloDoenca \leftarrow$ (algum termo de D aparece no título de n)
para cada *sentença* s *de* n **faça**
  $Doe \leftarrow$ doenças de D reconhecidas em s
  $Mun \leftarrow$ municípios de M reconhecidos em s
  $Sin \leftarrow$ sintomas de S reconhecidos em s
  se $Doe = \varnothing$ ***ou*** $Mun = \varnothing$ **então** próxima sentença
  para cada $(d, m) \in Doe \times Mun$ **faça**
    $c \leftarrow \textsc{Confiança}(Sin, tituloDoenca)$
    $R \leftarrow R \cup \{(d, m, c)\}$
retorne R `// menções repetidas na notícia são descartadas`

Fonte: Elaborado pelo autor (2026). Legenda: $\leftarrow$ atribuição de valor; $\varnothing$ conjunto vazio; $\in$ pertence a; $\times$ produto cartesiano (todas as combinações entre os dois conjuntos); $\cup$ união de conjuntos; $|X|$ número de elementos do conjunto X; $\geq$ maior ou igual.

A cada menção atribuiu-se um escore de confiança operacional, a partir de um valor-base de 0,62. Esse valor é acrescido de 0,18 quando há pelo menos um sintoma na sentença e de mais 0,08 quando há dois ou mais sintomas. Soma-se ainda 0,08 quando o título da notícia contém termo epidemiológico. O escore final é limitado a um teto de 0,96. Trata-se de uma pontuação operacional destinada a ordenar e filtrar menções, e não de uma probabilidade epidemiológica. Como os textos de reserva raramente descrevem sintomas, a maioria das menções do corpus fechado situou-se em 0,70 (valor-base acrescido do bônus de título). A validade discriminativa desse escore, isto é, sua capacidade real de separar acertos de erros, é examinada empiricamente nos resultados. O cálculo é detalhado no Algoritmo 2.

Algoritmo 2: Função CONFIANÇA: escore operacional de uma menção

Função `Confiança(`$Sin,\ tituloDoenca$`)`**:**
  $c \leftarrow 0{,}62$ `// valor-base`
  se $|Sin| \geq 1$ **então** $c \leftarrow c + 0{,}18$
  se $|Sin| \geq 2$ **então** $c \leftarrow c + 0{,}08$
  se $tituloDoenca$ **então** $c \leftarrow c + 0{,}08$
  retorne $\min(c,\ 0{,}96)$

Fonte: Elaborado pelo autor (2026).

3.2.6 Painel de visualização

A camada de visualização foi implementada com a biblioteca Streamlit, que levanta um servidor web local. O painel consulta diretamente as tabelas do banco SQLite e oferece filtros dinâmicos por doença, fonte, município, intervalo de datas e nível de confiança. A distribuição espacial das menções é apresentada em um mapa coroplético do Piauí, construído com Folium (Leaflet), em que a intensidade de cor de cada município reflete o número de menções detectadas. Complementam o painel um gráfico de distribuição por doença, uma série temporal por mês e uma tabela auditável que expõe, para cada menção, a sentença de origem e o link da notícia, assegurando a rastreabilidade exigida pela pesquisa.

Cabe delimitar o papel do painel neste trabalho. Ele foi concebido como instrumento de visualização direta e de auditoria dos dados extraídos, isto é, uma camada para inspecionar as menções e rastreá-las até a sentença de origem, e não como um produto de interface a ser entregue a usuários finais. Por essa razão, a construção da camada web restringiu-se à exibição bruta dos dados, sem preocupações de design de interação, e não foi submetida a análise de usabilidade. O objeto avaliado nesta pesquisa é o comportamento do *pipeline* de Processamento de Linguagem Natural, e não a interface, coerente com o caráter de prova de conceito do artefato.

3.2.7 Arquitetura genérica por domínio

O reconhecimento de entidades foi estruturado de forma genérica: o vocabulário do tema investigado, isto é, as doenças e seus sintomas, não está fixo no código, mas em um arquivo de configuração externo, denominado domínio de investigação. O domínio padrão corresponde às arboviroses; para investigar outro tema, basta substituir o arquivo de configuração, mantendo o município como eixo geográfico fixo. Trata-se de uma propriedade de projeto do artefato, que amplia seu potencial de reúso. Ressalva-se, contudo, que essa flexibilidade não foi validada empiricamente para outros temas: as métricas reportadas referem-se exclusivamente ao domínio de arboviroses.

3.3 DELIMITAÇÃO E CONSTRUÇÃO DO CORPUS

A coleta de dados foi delimitada a um corpus fechado e curado, composto de notícias e boletins oficiais. A opção por um corpus fechado e verificado apoiou-se em três razões metodológicas: reprodutibilidade, pois um corpus fixo permite que a avaliação seja replicada com os mesmos dados; controle de qualidade, já que os textos foram verificados manualmente antes da inclusão; e viabilidade, dado que a coleta ao vivo depende da disponibilidade das URLs e da robustez do scraping, variáveis fora do escopo desta investigação.

Optou-se por não incorporar a rede social X (Twitter) como fonte: o acesso à sua API é restrito e a maioria das postagens não dispõe de geolocalização precisa, o que comprometeria o requisito central da pesquisa, a associação entre doença e município. Por isso, a investigação concentrou-se em fontes jornalísticas e boletins oficiais, nas quais a referência geográfica é explícita e rastreável.

O recorte temporal abrangeu o período de janeiro a dezembro de 2024, marcado pela alta incidência de arboviroses no estado. Foram selecionadas matérias que continham, no título ou na URL, ao menos uma das palavras-chave dengue, zika, chikungunya, arbovirose ou *Aedes*. O corpus final reuniu 35 artigos, distribuídos entre oito fontes, conforme a Tabela 1.

Tabela 1 – Composição do corpus por fonte

Fonte	**Artigos**
Cidade Verde	18
SESAPI	6
G1 Piauí	5
Ministério da Saúde	2
Governo do Piauí / SESAPI	1
Portal O Dia	1
Portal B1	1
A10+	1
Total	**35**

Fonte: Elaborada pelo autor (2026).

3.4 CONSTRUÇÃO DO PADRÃO-OURO (GOLD STANDARD)

Diante da inexistência de bases públicas anotadas para notícias de arboviroses no Piauí, construiu-se um conjunto de referência (Padrão-Ouro) por rotulagem manual. A anotação cobriu a totalidade dos 35 artigos do corpus, no nível de relação, isto é, tomando como unidade o par (doença, município) por artigo, sem duplicatas. Essa decisão alinhou a avaliação à pergunta de pesquisa, centrada na associação entre

agravo e local, e ampliou o rigor ao cobrir todo o corpus. A anotação foi realizada por um único anotador, o autor, o que constitui uma limitação reconhecida, dada a ausência de verificação por um segundo anotador e de medida de concordância entre anotadores, como o Kappa de Cohen. Reconhece-se que a própria necessidade de critérios formais de anotação (Quadro 6), como o tratamento de menções negativas, de referências geográficas e de contextos de vacinação, evidencia que houve julgamento a resolver, e não uma tarefa isenta de interpretação. Ainda assim, três fatores atenuam a limitação: os critérios foram definidos previamente e de forma explícita, justamente para reduzir a variabilidade de julgamento; a decisão de anotação recai sobre a presença factual de um par doença-município, e não sobre juízos de valor ou de sentimento; e a anotação cobriu a totalidade do corpus, e não uma amostra. A ausência de um segundo anotador e de medida de concordância permanece, portanto, uma limitação assumida do trabalho.

A anotação resultou em 87 menções corretas. Os critérios formais adotados são apresentados no Quadro 6.

Quadro 6 – Critérios de anotação do Padrão-Ouro

Nº	**Critério**
1	Menção válida: arbovirose (Dengue, Zika ou Chikungunya) associada a município piauiense com presença confirmada ou suspeita, isto é, casos, óbitos, estado de alerta ou risco de infestação.
2	Menção negativa é inválida: artigo que informa explicitamente a ausência da doença no município não gera menção válida.
3	Referência geográfica é inválida: município citado apenas como referência espacial (por exemplo, distância) não gera menção válida.
4	Contexto de vacinação é válido: município associado a doença em contexto de vacinação conta como dado epidemiológico relevante.
5	Nível estadual é inválido: menções genéricas ao Piauí, sem município específico, não são contabilizadas.

Fonte: Elaborado pelo autor (2026).

3.5 PROCEDIMENTO DE AVALIAÇÃO

O desempenho do NER foi mensurado pela comparação entre as menções extraídas automaticamente e o Padrão-Ouro, tomando como chave a tripla (notícia, doença, município) deduplicada por artigo. A avaliação foi conduzida sobre os textos de reserva do corpus fechado, assegurando reprodutibilidade[3]. A partir da matriz de

[3] O corpus, o Padrão-Ouro e os *scripts* de avaliação estão disponíveis em <https://github.com/pedrolucas2530/E-Piau->, permitindo a reexecução integral do procedimento e a verificação das métricas reportadas.

confusão, calcularam-se três métricas consagradas na literatura de Recuperação da Informação. A partir dessa comparação, cada menção é classificada em uma matriz de confusão, da qual interessam três categorias: os verdadeiros positivos (VP), menções extraídas pelo sistema que constam do Padrão-Ouro; os falsos positivos (FP), menções extraídas que não constam do Padrão-Ouro; e os falsos negativos (FN), menções do Padrão-Ouro que o sistema não recuperou. A partir dessas contagens, calcularam-se três métricas consagradas na literatura de Recuperação da Informação.

A Precisão (P) mede, entre as menções extraídas, a proporção de corretas, penalizando os falsos positivos:

$$P = \frac{VP}{VP + FP}. \tag{1}$$

A Revocação (R), também chamada de sensibilidade, mede, entre as menções de fato existentes, a proporção efetivamente recuperada, penalizando os falsos negativos:

$$R = \frac{VP}{VP + FN}. \tag{2}$$

O F1-score (F_1) combina as duas anteriores por meio de sua média harmônica, fornecendo uma medida única e equilibrada:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} = \frac{2VP}{2VP + FP + FN}. \tag{3}$$

A escolha da média harmônica, e não da aritmética, deve-se à sua maior sensibilidade ao menor dos dois valores: o F_1 só é alto quando precisão e revocação são simultaneamente altas, penalizando sistemas que privilegiam uma métrica em detrimento da outra. As métricas foram calculadas de forma agregada e por doença.

3.6 ASPECTOS ÉTICOS

A pesquisa utilizou exclusivamente dados públicos e secundários, sem participantes humanos e sem dados sensíveis, o que dispensa a submissão ao Comitê de Ética em Pesquisa. A coleta automatizada observou as diretrizes de ética na web, incluindo a verificação do arquivo robots.txt dos portais e o respeito às restrições de acesso, e o tratamento dos dados observou a Lei Geral de Proteção de Dados (LGPD).

4 RESULTADOS E DISCUSSÃO

Esta seção apresenta os resultados obtidos com a execução do protótipo sobre o corpus fechado de 2024 e discute seus significados e limitações. São reportados a composição do corpus, o Padrão-Ouro construído, o desempenho do NER, a análise dos erros, o comportamento da visualização e a discussão integrada à luz da pergunta de pesquisa.

4.1 COMPOSIÇÃO DO CORPUS

O corpus reuniu 35 artigos publicados ao longo dos doze meses de 2024, cobrindo o ciclo epidêmico da dengue no estado, com crescimento no primeiro trimestre, pico em torno de março e abril e posterior redução, além da vigilância de chikungunya no segundo semestre. A distribuição por fonte (Tabela 1) evidencia a predominância do portal Cidade Verde, com 18 dos 35 artigos, seguido dos boletins da SESAPI, com 6, e do G1 Piauí, com 5; as demais fontes contribuíram com um a dois artigos cada.

4.2 PADRÃO-OURO CONSTRUÍDO

A anotação manual identificou 87 menções válidas. Desse total, 22 dos 35 artigos continham ao menos uma menção, enquanto 13 não apresentaram menção válida segundo os critérios definidos. A distribuição por doença (Tabela 2) revelou forte predominância da dengue, com 83 menções, seguida da chikungunya, com 4, e ausência de menções de zika associadas a município específico, pois a zika apareceu no corpus apenas em nível estadual, sem vínculo municipal.

Tabela 2 – Distribuição do Padrão-Ouro por doença

Doença	**Menções**
Dengue	83
Chikungunya	4
Zika	0
Total	**87**

Fonte: Elaborada pelo autor (2026).

4.3 DESEMPENHO DO RECONHECIMENTO DE ENTIDADES

O pipeline extraiu, após deduplicação por artigo, 39 pares (doença, município), dos quais 37 corretos e 2 incorretos, deixando de recuperar 50 menções presentes no Padrão-Ouro. O desempenho agregado e por doença é apresentado na Tabela 3.

Tabela 3 – Desempenho do NER por doença e agregado

Entidade	VP	FP	FN	Precisão	Revocação	F1
Dengue	33	2	50	94,3%	39,8%	55,9%
Chikungunya	4	0	0	100,0%	100,0%	100,0%
Zika	0	0	0	N/A	N/A	N/A
Geral	**37**	**2**	**50**	**94,9%**	**42,5%**	**58,7%**

Fonte: Elaborada pelo autor (2026). Legenda: VP, verdadeiros positivos; FP, falsos positivos; FN, falsos negativos.

Os resultados revelam um sistema de alta precisão (94,9%) e revocação moderada (42,5%), sintetizadas por um F1-score de 58,7%. A elevada precisão indica que, quando o sistema extrai uma menção, quase sempre acerta, comportamento coerente com a regra conservadora de co-ocorrência sentencial. O resultado perfeito da chikungunya (F1 de 100%) deve ser lido com cautela, pois se baseia em apenas quatro instâncias, base amostral insuficiente para generalização. Para a zika, não havendo menções municipais no corpus, não foi possível mensurar o desempenho.

Cabe ressalvar que, dado o desbalanceamento amostral, o resultado agregado reflete majoritariamente o comportamento do sistema frente à dengue: dos 37 verdadeiros positivos, 33 são de dengue e apenas 4 de chikungunya. Por isso, o F1 geral de 58,7% aproxima-se do F1 da dengue (55,9%), e a leitura por doença (Tabela 3) é mais informativa do que o número agregado tomado isoladamente. Em termos práticos, o desempenho reportado é essencialmente o desempenho do sistema sobre a dengue. Além disso, o tamanho amostral reduzido implica incerteza estatística não desprezível. Estimados por reamostragem *bootstrap* por artigo (10.000 repetições), os intervalos de confiança de 95% mostraram-se amplos: precisão de 94,9% (IC de 83,3% a 100%), revocação de 42,5% (IC de 19,7% a 68,1%) e F1 de 58,7% (IC de 32,3% a 79,7%). A amplitude, sobretudo na revocação, reforça a leitura dos números como indicativos, e não como estimativas de precisão populacional.

Verificou-se ainda a validade discriminativa do escore de confiança, comparando o escore médio dos verdadeiros positivos ao dos falsos positivos. Ambos os grupos apresentaram escore idêntico, de 0,70, sem qualquer separação entre acertos e erros. O resultado decorre da própria composição do corpus: como os textos de reserva raramente descrevem sintomas, os incrementos previstos na fórmula quase nunca se aplicam, e o escore colapsa em um valor praticamente constante. Conclui-se que, neste corpus, o escore opera como indicador de completude textual, e não como medida validada de confiabilidade epidemiológica, limitação que uma calibração futura contra a correção observada deveria tratar.

4.4 ANÁLISE DE ERROS

Os dois falsos positivos têm causas distintas e ilustram limitações conhecidas de sistemas baseados em regras, sintetizadas no Quadro 7. No primeiro caso, o sistema extraiu a co-ocorrência em uma sentença que negava a ocorrência da doença, pois não há tratamento de negação. No segundo, o município de Teresina foi citado apenas como referência espacial, e o sistema não distingue contexto de ocorrência de contexto de referência geográfica.

Quadro 7 – Falsos positivos e suas causas

Menção extraída	**Causa do erro**
(Dengue, Picos)	Negação não detectada: a sentença afirmava que Picos ainda não havia registrado casos de dengue em 2024.
(Dengue, Teresina)	Referência geográfica: Teresina foi citada como distância de referência (a cerca de 600 km de um óbito ocorrido em Bom Jesus), e não como local de ocorrência.

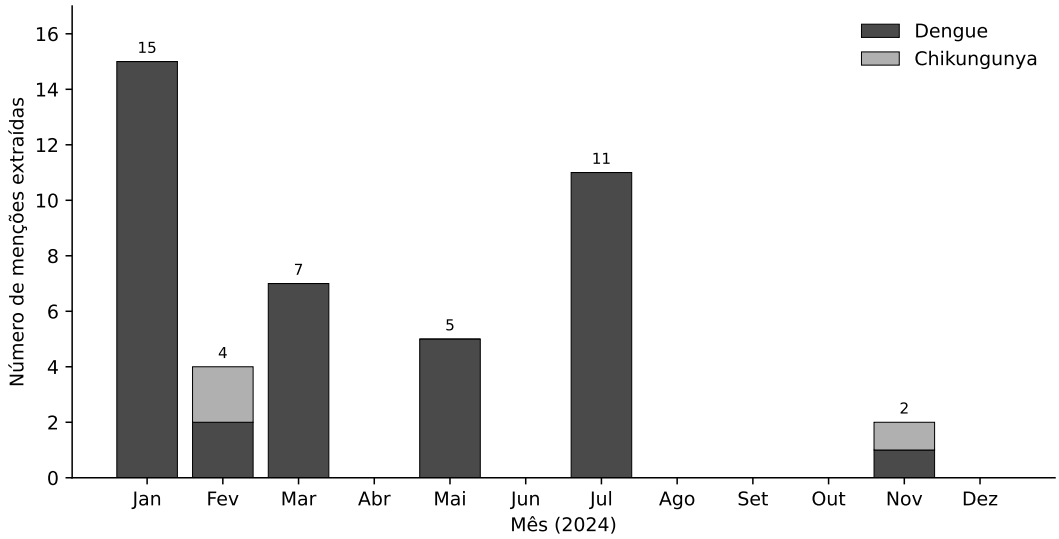
Fonte: Elaborado pelo autor (2026).

Os 50 falsos negativos têm causa quase exclusivamente estrutural: artigos que mencionam a doença em uma sentença e listam os municípios afetados em outra, padrão típico de notas de imprensa e boletins epidemiológicos. Como a heurística exige co-ocorrência na mesma sentença, essas associações não são capturadas. Exemplos recorrentes envolvem boletins que afirmam o aumento da dengue no estado em uma sentença e, na seguinte, enumeram os municípios com casos ou óbitos confirmados.

4.5 VISUALIZAÇÃO E PADRÃO TEMPORAL

Em operação normal, com tentativa de download e recurso ao texto de reserva quando necessário, o sistema consolidou 49 menções no banco, sendo 46 de dengue e 3 de chikungunya, ante 47 obtidas estritamente sobre os textos de reserva; a pequena diferença decorre de uma ou duas matérias efetivamente baixadas com conteúdo mais extenso. A Figura 7 apresenta a distribuição mensal das menções extraídas ao longo de 2024, discriminadas por doença.

Figura 7 – Distribuição mensal das menções extraídas em 2024, por doença



Fonte: Elaborada pelo autor (2026).

As menções concentram-se no primeiro semestre, com maior volume em janeiro, seguido de julho e março, e a chikungunya restrita a fevereiro e novembro, em pequeno número. Cabe interpretar esse padrão com cautela: a distribuição reflete sobretudo a cobertura jornalística e a disponibilidade de matérias no período, e não uma medida direta da incidência epidemiológica. Soma-se a isso uma limitação do corpus: parte das matérias não dispunha de data de publicação precisa, tendo sido registrada com data genérica de início de ano, o que concentra artificialmente contagens em janeiro. Assim, a série deve ser lida como a distribuição das menções no corpus, e não como uma curva de casos, o que reafirma, mais uma vez, tratar-se de uma prova de conceito.

Ressalta-se que os dois conjuntos de números respondem a granularidades e execuções distintas e, por isso, não são diretamente comparáveis. As métricas de precisão, revocação e F1 tomam como unidade o par (doença, município) deduplicado por artigo, apurado sobre os textos de reserva, o que corresponde aos 39 pares extraídos. As 47 a 49 menções, por sua vez, são contadas por sentença, sem deduplicação por artigo, e refletem o modo operacional do sistema, com tentativa de download e recurso à contingência. Assim, a diferença entre 39 e 49 não indica inconsistência nos dados, mas mudança da unidade de contagem: pares deduplicados por artigo, na avaliação, e menções por sentença, na operação.

4.6 DISCUSSÃO

Para um leitor da área da saúde, convém traduzir essas métricas ao contexto de vigilância. A precisão responde à pergunta: das menções que o sistema aponta, quantas correspondem a fatos reais? Uma precisão alta significa poucos alarmes falsos e, portanto, maior confiança do gestor no sinal produzido. A revocação responde à

pergunta complementar: de todos os casos efetivamente mencionados nas fontes, quantos o sistema conseguiu captar? Uma revocação baixa significa que muitos sinais passam despercebidos. Em vigilância epidemiológica, os dois tipos de erro têm custos assimétricos: um falso positivo gera, no pior caso, uma verificação desnecessária, ao passo que um falso negativo pode corresponder a um foco não detectado em tempo hábil. É essa assimetria que orienta a leitura dos resultados a seguir e sustenta a decisão de posicionar a ferramenta como filtro de alta precisão, complementar, e não substituto, da vigilância oficial.

Os resultados respondem à pergunta de pesquisa de forma matizada: é viável identificar e geolocalizar automaticamente menções a arboviroses em notícias do Piauí com alta precisão, mas a cobertura é limitada quando se adota uma heurística de co-ocorrência estritamente sentencial. Em termos práticos, o protótipo é mais útil como filtro de alta precisão, que minimiza falsos alarmes, do que como mecanismo de cobertura exaustiva. Esse achado atende ao objetivo de avaliar criticamente a abordagem, articulando a mensuração do desempenho com a análise da viabilidade técnica e das limitações observadas no contexto piauiense.

As principais limitações identificadas foram: a heurística de co-ocorrência sentencial, responsável por praticamente todos os falsos negativos; a ausência de tratamento de negação; a ambiguidade de referência geográfica; a impossibilidade de avaliar a zika, ausente em nível municipal no corpus; o viés de cobertura das fontes, com predominância de um único portal; e o acoplamento do sistema ao recorte do Piauí e às três arboviroses. Tais limitações delimitam o alcance das conclusões e apontam caminhos de aprimoramento, como o tratamento de negação, a expansão da janela de associação para além da sentença e a incorporação de classificação supervisionada para reduzir falsos positivos. Soma-se a elas uma limitação de validade externa: a avaliação foi conduzida sobre textos de reserva previamente verificados e selecionados, presumivelmente mais limpos e regulares do que o material capturado por coleta ao vivo em condições reais, sujeito a HTML malformado, truncamento e bloqueios parciais. Esse viés de seleção do corpus de avaliação limita a generalização das métricas para um cenário de produção. Não é possível, porém, afirmar a priori a direção do efeito sobre cada métrica, pois texto mais ruidoso tanto pode omitir associações, reduzindo a revocação, quanto criar co-ocorrências espúrias, reduzindo a precisão. Por essa razão, os valores reportados devem ser lidos como um limite superior de desempenho: uma estimativa do que o método alcança em condições favoráveis de entrada, e não do que se observaria sobre material coletado ao vivo.

A expansão da janela de associação, contudo, não é um ajuste trivial, e a opção por mantê-la na sentença foi deliberada. Para dimensionar esse efeito de forma empírica, e não apenas especulativa, reprocessou-se o mesmo corpus com a janela ampliada ao documento inteiro. Nessa configuração, doença e município passam a

ser associados sempre que ambos ocorram na mesma matéria. O resultado foi uma revocação de 100% (todos os 87 pares recuperados), uma precisão de 64,9% (47 falsos positivos) e um F1 de 78,7%.

Esse resultado deve ser lido como um contraexemplo controlado, e não como uma recomendação de uso. A revocação de 100% não significa que o método seja bom nessa configuração. Ela apenas indica que o teto de cobertura do corpus é atingível, pois, em todo par do Padrão-Ouro, a doença e o município aparecem em algum ponto do artigo. Esse teto, porém, só é alcançado ao custo de uma precisão que inviabiliza o uso: cerca de um terço das menções extraídas torna-se espúrio. Isso ocorre porque a janela de documento vincula doença e município por mera proximidade textual, sem garantia de relação semântica. Capturar corretamente essas associações distribuídas exigiria resolução de correferência, tarefa reconhecidamente complexa em Processamento de Linguagem Natural, sobretudo em português e em textos com muitas entidades.

O experimento esclarece ainda um ponto mais sutil. A escolha da janela não se resume a maximizar o F1, que é inclusive maior na janela de documento. O que está em jogo é alinhar a métrica ao objetivo do trabalho. Como a proposta se posiciona como filtro de alta precisão, que minimiza falsos alarmes, a co-ocorrência sentencial é a decisão coerente. Caso o objetivo fosse a cobertura exaustiva, a janela de documento seria preferível. Mantém-se, portanto, a janela de sentença como decisão de projeto, deixando a ampliação controlada, acompanhada de mecanismos de desambiguação, como caminho de trabalho futuro.

Mais do que falhas pontuais, parte dessas limitações configura a caracterização dos desafios locais que constituiu objetivo central deste trabalho. Três deles merecem destaque por serem próprios do contexto piauiense, e não meras contingências de implementação. O primeiro é a instabilidade da coleta automatizada, evidenciada pelos bloqueios recorrentes ao scraping e pela consequente dependência de textos de reserva. O segundo é a baixa densidade e o viés de cobertura das fontes locais, com a produção jornalística concentrada em poucos veículos, o que reduz a quantidade de associações detectáveis. O terceiro é a natureza heurística da geolocalização, que, ao depender da co-ocorrência sentencial entre doença e município, mostra-se sensível ao estilo redacional de notas oficiais e boletins regionais. Esses achados, tomados em conjunto, descrevem as condições concretas sob as quais a Epidemiologia Digital sobre fontes abertas pode operar no Piauí, e constituem contribuição da pesquisa independentemente do desempenho numérico do protótipo.

Por fim, reitera-se o caráter de prova de conceito do artefato: os resultados demonstram a viabilidade técnica da abordagem com ferramentas de baixo custo e código aberto, sem a pretensão de substituir os sistemas oficiais de vigilância ou de servir como fonte única para decisões em saúde pública.

5 CONSIDERAÇÕES FINAIS

Este trabalho propôs, implementou em caráter de prova de conceito e avaliou criticamente uma metodologia de Epidemiologia Digital, baseada em *web scraping* e Processamento de Linguagem Natural, para identificar e geolocalizar menções a arboviroses em notícias e boletins oficiais do Piauí, com ênfase na caracterização dos desafios locais. Os resultados permitem responder, de modo qualificado, à questão de pesquisa: é viável extrair e geolocalizar essas menções com alta precisão, ainda que com cobertura limitada quando se adota uma heurística de co-ocorrência restrita ao escopo da sentença. Em termos práticos, o protótipo mostrou-se mais adequado como filtro de alta precisão, que minimiza falsos alarmes, do que como mecanismo de cobertura exaustiva.

Entre as contribuições do trabalho, destacam-se a construção de um instrumento de baixo custo, baseado em tecnologias de código aberto, capaz de automatizar o ciclo de coleta, extração e visualização geográfica; a elaboração de um conjunto de referência (Padrão-Ouro) específico para arboviroses no Piauí, anotado manualmente e útil a avaliações futuras; e a análise sistemática dos erros de extração, que evidenciou as causas dos acertos e das falhas do método. Para além do artefato, uma contribuição central foi a caracterização dos desafios concretos de aplicar a Epidemiologia Digital sobre fontes abertas no estado, em especial a instabilidade da coleta automatizada, o viés e a baixa densidade das fontes locais e a sensibilidade da geolocalização heurística ao estilo redacional dos textos. Esse conhecimento situado, de natureza metodológica e didática, oferece subsídios para que trabalhos futuros repliquem e aprimorem a abordagem em contextos semelhantes.

Reconhecem-se, contudo, limitações que delimitam o alcance das conclusões. A principal é estrutural: a regra de co-ocorrência sentencial deixa de capturar associações em que a doença e os municípios aparecem em sentenças distintas, o que respondeu por praticamente todos os casos não recuperados. Somam-se a isso a ausência de tratamento de negação, a ambiguidade de referências geográficas, a impossibilidade de avaliar a zika por ausência de menções municipais no corpus e o viés de cobertura das fontes. Por essas razões, reitera-se o caráter de prova de conceito do artefato: ele demonstra a viabilidade técnica da abordagem, sem a pretensão de substituir os sistemas oficiais de vigilância ou de servir como fonte única para decisões em saúde pública.

Em consonância com o princípio da reprodutibilidade, o código-fonte, o corpus e o Padrão-Ouro utilizados neste trabalho estão disponíveis publicamente em <https://github.com/pedrolucas2530/E-Piau->, permitindo a verificação e a extensão dos resultados aqui reportados.

5.1 TRABALHOS FUTUROS

A partir das limitações identificadas, vislumbram-se as seguintes direções para trabalhos futuros:

- avaliar a usabilidade do painel junto a gestores de saúde, por meio de testes com usuários, caso a ferramenta evolua de prova de conceito para instrumento de uso efetivo;
- incorporar o tratamento de negação e a desambiguação de referências geográficas, de modo a reduzir os falsos positivos;
- ampliar a janela de associação para além da sentença, considerando o parágrafo ou o documento, com vistas a elevar a revocação;
- complementar as regras com técnicas de aprendizado supervisionado, treinadas sobre o Padrão-Ouro construído;
- expandir o corpus e a diversidade de fontes, incluindo a construção de um conjunto de referência específico para a zika;
- explorar a generalização do artefato para outros temas de investigação, mantendo o município como eixo geográfico, possibilidade já prevista em sua arquitetura por meio de domínios configuráveis;
- evoluir o painel para permitir a investigação interativa de um termo digitado pelo usuário, com a respectiva visualização no mapa;
- integrar fontes de dados oficiais para fins de comparação e validação cruzada dos resultados; e
- avaliar o emprego de modelos de linguagem em português sensíveis à variação regional, a exemplo de iniciativas locais recentes, com vistas a mitigar o ruído linguístico e a melhorar o reconhecimento de entidades em textos do estado.
- aprimorar a robustez do módulo de coleta ao vivo, tornando o *web scraping* resiliente a bloqueios e a páginas instáveis, por exemplo mediante estratégias de repetição, rotação de agentes e tratamento de HTML malformado, de modo a reduzir a dependência de textos de reserva e a viabilizar a operação contínua;

Em conjunto, essas direções podem ampliar a cobertura e a robustez da abordagem, aproximando-a de aplicações mais amplas no campo da Epidemiologia Digital.

REFERÊNCIAS

ANTUNES, M. N. et al. Monitoramento de informação em mídias sociais: o e-monitor dengue. **Transinformação**, v. 26, n. 1, p. 9–18, Jan 2014. ISSN 0103-3786. Acesso em: 30 jun. 2026. Disponível em: <https://www.scielo.br/j/tinf/a/q8bDgmrg99CwYkdpfzSggBD/>.

BOGOCH, I. I. et al. Pneumonia of unknown aetiology in Wuhan, China: potential for international spread via commercial air travel. ***Journal of Travel Medicine***, v. 27, n. 2, p. taaa008, 2020.

BRASIL. MINISTÉRIO DA SAÚDE. **Novo plano de ação prevê reduzir impactos da dengue e outras arboviroses no Piauí**. 2024. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/noticias-para-os-estados/piaui/2024/setembro/novo-plano-de-acao-preve-reduzir-impactos-da-dengue-e-outras-arboviroses-no-piaui>. Acesso em: 30 jun. 2026.

BROWNSTEIN, J. S.; FREIFELD, C. C.; MADOFF, L. C. Digital disease detection — harnessing the web for public health surveillance. ***New England Journal of Medicine***, v. 360, n. 21, p. 2153–2157, 2009. Acesso em: 30 jun. 2026. Disponível em: <https://www.nejm.org/doi/full/10.1056/NEJMp0900702>.

CHARLES-SMITH, L. E. et al. Using social media for actionable disease surveillance and outbreak management: A systematic literature review. ***PLOS ONE***, v. 10, n. 10, p. 1–20, 10 2015. Acesso em: 30 jun. 2026. Disponível em: <https://doi.org/10.1371/journal.pone.0139701>.

EYSENBACH, G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. ***Journal of Medical Internet Research***, JMIR Publications Inc., Toronto, Canada, v. 11, n. 1, p. e11, 2009. Acesso em: 30 jun. 2026. Disponível em: <https://www.jmir.org/2009/1/e11/>.

GIL, A. C. **Como elaborar projetos de pesquisa**. 6. ed. São Paulo: Atlas, 2019.

GOMIDE, J. et al. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In: ***Proceedings of the 3rd International Web Science Conference (WebSci '11)***. Nova York: ACM, 2011. p. 1–8.

HEVNER, A. R. et al. Design science in information systems research. ***MIS Quarterly***, v. 28, n. 1, p. 75–105, 2004.

JURAFSKY, D.; MARTIN, J. H. ***Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech***

Recognition with Language Models. 3. ed. [S.l.]: [s.n.], 2025. Online manuscript released August 24, 2025. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>.

LEANDRO, C. S. et al. Redução da incidência de dengue no Brasil em 2020: controle ou subnotificação de casos por COVID-19? ***Research, Society and Development***, v. 9, n. 11, p. e76891110442, 2020. Acesso em: 30 jun. 2026. Disponível em: <https://rsdjournal.org/rsd/article/view/10442>.

MAYER-SCHÖNBERGER, V.; CUKIER, K. **Big Data: Como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana**. Rio de Janeiro: Elsevier Brasil, 2014. ISBN 9788535273410.

MITCHELL, R. ***Web scraping with Python: Collecting more data from the modern web***. Sebastopol: O'Reilly Media, Inc., 2018.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. ***Lingvisticæ Investigationes***, v. 30, n. 1, p. 3–26, 2007. Acesso em: 30 jun. 2026. Disponível em: <https://www.jbe-platform.com/content/journals/10.1075/li.30.1.03nad>.

PEFFERS, K. et al. A design science research methodology for information systems research. ***Journal of Management Information Systems***, v. 24, n. 3, p. 45–77, 2007.

PIAUÍ. SECRETARIA DA SEGURANÇA PÚBLICA. **Piauí lança o BO Fácil: primeiro serviço do país que permite registrar boletim de ocorrência via WhatsApp**. 2025. Disponível em: <https://www.ssp.pi.gov.br/piaui-lanca-o-bo-facil-primeiro-servico-do-pais-que-permite-registrar-boletim-de-ocorrencia-via-whatsapp/>. Acesso em: 30 jun. 2026.

PIAUÍ. SECRETARIA DE INTELIGÊNCIA ARTIFICIAL, ECONOMIA DIGITAL, CIÊNCIA, TECNOLOGIA E INOVAÇÃO. **Piauí lança Soberanla, inteligência artificial treinada para entender sotaques, cultura e dados públicos**. 2025. Disponível em: <https://sia.pi.gov.br/piaui-lanca-soberania-inteligencia-artificial-treinada-para-entender-sotaques-cultura-e-dados-publicos/>. Acesso em: 30 jun. 2026.

SALATHÉ, M. Digital epidemiology: what is it, and where is it going? ***Life Sciences, Society and Policy***, v. 14, n. 1, p. 1, 2018. Acesso em: 30 jun. 2026. Disponível em: <https://doi.org/10.1186/s40504-017-0065-7>.

UNIVERSIDADE FEDERAL DO PIAUÍ (UFPI). CENTRO DE INTELIGÊNCIA EM AGRAVOS TROPICAIS EMERGENTES E NEGLIGENCIADOS. **CIATEN: inteligência epidemiológica aplicada a agravos tropicais e negligenciados**. 2025. Disponível em: <https://novosite.ciaten.org.br/>. Acesso em: 30 jun. 2026.

VIEIRA, G. C. et al. Processamento de linguagem natural aplicado a registros eletrônicos: monitoramento e detecção de eventos em saúde. **Ciência & Saúde Coletiva**, v. 30, n. 7, p. e18722024, 2025. Acesso em: 30 jun. 2026. Disponível em: <https://doi.org/10.1590/1413-81232025307.18722024>.