

## VERITAS: Recuperação de Contexto e Informações com Retrieval-Augmented Generation para Comunicação de Risco Climático

Julio Rafael Souza Peixoto<sup>123</sup> (Voluntário/CNPq)

João Victor Miranda Lyra<sup>123</sup> (Voluntário/CNPq)

Victor Flávio de Andrade Araujo<sup>123</sup>

<sup>1</sup>Universidade Tiradentes/Ciência da Computação/Aracaju/SE.

<sup>2</sup>Grupo de Pesquisa Interdisciplinar em Tecnologia, Computação e Sociedade (GPITCS - UNIT)

<sup>3</sup>Instituto Nacional de Ciência e Tecnologia em Neurociência Social e Afetiva (INCT-SANI)

1.03.00.00-7 – Ciência da Computação; 1.03.03.04-9; 1.07.03.06-3 – Climatologia; 1.03.03.02-2 – Engenharia de Software

### RESUMO

**Introdução:** Aracaju já sente os efeitos das mudanças climáticas, ruas que enchem com pancadas de chuva, drenagem urbana que não dá conta em pontos críticos e dias abafados que cria desconforto térmico. Para reduzir incertezas, propomos um sistema que pode, em uma única resposta, alertas e relatórios de risco, dados de chuva e temperatura e orientações de saúde locais, ajudando quem precisa decidir a agir rápido e com base em informação verificada. **Objetivos:** Construir e avaliar um pipeline capaz de recuperar e sintetizar, com precisão e transparência, dados sobre chuvas intensas, pontos críticos de alagamento, protocolos de emergência e cuidados em ondas de calor. Objetivos específicos incluem indexar documentos locais com metadados, otimizar a recuperação por similaridade semântica, calibrar a geração ao contexto de Aracaju e aferir desempenho com perguntas de uso real. **Metodologia:** Implementou-se pipeline com extração de texto por formato, segmentação em trechos de mil a mil e quinhentos caracteres com sobreposição de cento e cinquenta, enriquecidos com metadados. Embeddings foram produzidos com text-embedding-3-small e armazenados em SQLiteVec, usando similaridade por cosseno. A consulta é expandida por modelo de linguagem de baixa temperatura para sinônimos e termos correlatos, seguida de busca e recuperação dos cinco trechos mais similares. A resposta é gerada por GPT-4o-mini com prompt especializado em clima e defesa civil locais. O ranqueamento combina similaridade e posição no resultado. O acervo inclui notícias de alagamentos por bairro, relatórios oficiais, dados de precipitação e orientações médicas, além de contatos de emergência. **Resultados:** Em avaliação comparativa pontual, o pareamento entre GPT-4o-mini integrado ao nosso RAG e GPT-4o-mini sem recuperação externa evidenciou ganhos de especificidade, localidade e acionabilidade. Para a questão sobre risco de alagamento no bairro São José em cenário de chuva intensa, o nosso sistema afirmou o risco e contextualizou com histórico recente e alerta da Defesa Civil, oferecendo resposta direta e situacional. O modelo sem RAG respondeu de forma genérica, limitando-se a recomendar o acompanhamento de fontes oficiais e previsões, sem conclusão específica sobre o bairro. Na identificação de pontos a evitar próximos ao São José, o nosso sistema listou vias concretas Rua Lagarto e Avenida Filadélfia Dórea, orientando a evitá-las durante chuvas fortes. O modelo sem RAG declarou não dispor de dados locais, sugerindo consulta a canais genéricos. Em termos qualitativos, o nosso sistema apresentou maior cobertura de entidades locais, maior densidade de evidências contextuais e recomendações práticas, enquanto a versão sem recuperação manteve orientação ampla, pouco decisiva e sem referências a logradouros, histórico ou contatos de emergência. Esses achados reproduzem o padrão observado em outras perguntas do conjunto de testes, nas quais o serviço favoreceu respostas mais úteis ao usuário final no município. **Conclusões:** O RAG mostrou-se promissor para comunicação de risco climático em contexto municipal, conciliando recuperação específica e geração controlada. Persistem limitações no ranqueamento e na ausência de métricas formais de avaliação, além de cobertura documental parcial em algumas consultas. Como próximos passos, propõem-se re-ranking mais robusto, chunking sensível à semântica, métricas padronizadas de precisão e revocação e ampliação do acervo com mapas oficiais de risco e séries históricas de precipitação.

**PALAVRAS-CHAVE:** Aracaju, Clima, RAG

**Agradecimentos:** Agradecemos à Universidade Tiradentes pelo apoio institucional, aos órgãos municipais que publicam boletins de risco e à comunidade local por disponibilizar dados e relatos úteis à indexação.

## ABSTRACT

**Introduction:** Aracaju is already feeling the strain of a harsher climate streets that flood after sudden downpours, storm-drain bottlenecks at known trouble spots, and muggy days that amplify heat stress. To cut uncertainty when, we propose a system that can in one city-aware answer combining civil defense alerts and risk bulletins, live rainfall and temperature signals, and locally grounded health guidance so decision-makers can act quickly on verified information.

**Objectives:** To build and assess a RAG that accurately retrieves and synthesizes data on intense rainfall, flood hotspots, emergency protocols, and heat safety. Specific aims were to index locally relevant documents with rich metadata, optimize semantic retrieval, tailor generation to Aracaju's context, and evaluate performance with realistic questions.

**Methods:** The pipeline extracts text by format, segments documents into one to one and a half thousand character chunks with one hundred and fifty overlap, and enriches records with metadata. Embeddings are produced with text-embedding-3-small and stored in SQLiteVec, using cosine similarity. Queries are expanded by a low-temperature language model to add synonyms and related terms before search. The top five passages compose the context for answer generation with GPT-4o-mini guided by climate and civil defense prompts. Ranking combines similarity and positional score. The corpus includes local flooding news by neighborhood, official reports, precipitation figures, medical guidance for heat, and emergency contacts. **Results:** A side-by-side comparison between GPT-4o-mini with our RAG and GPT-4o-mini alone highlighted consistent gains in specificity, locality, and actionability. For the question about flood risk in the São José neighborhood during heavy rain, our system concluded there was risk and contextualized it with recent history and a civil defense alert, delivering a direct, situational answer. The model without RAG responded generically, advising to check official sources without addressing the neighborhood risk. When asked which streets near São José should be avoided, our system named concrete locations Rua Lagarto and Avenida Filadélfia Dórea, while the baseline stated it lacked local data and suggested generic channels. Qualitatively, our system increased local entity coverage, contextual evidence, and practical guidance, whereas the baseline remained broad, noncommittal, and devoid of street-level references or emergency contacts. Similar patterns were observed across other tested prompts, favoring the RAG for user utility at the municipal scale. **Conclusions:** The RAG is a promising tool for municipal climate risk communication, balancing targeted retrieval with controlled generation. Remaining gaps include simplified ranking, lack of formal evaluation metrics, and partial corpus coverage. Future work involves stronger re-ranking, semantically aware chunking, standardized metrics, and expanded sources such as official risk maps and historical precipitation series.

**KEYWORDS:** Aracaju, Climate, RAG

**ACKNOWLEDGEMENTS:** We thank Universidade Tiradentes for institutional support, municipal agencies for publishing risk bulletins, and the local community for providing reports that enabled indexing.