# Text Information Retrieval in Tetun

Gabriel de Jesus
INESC TEC
Portugal
gabriel.jesus@inesctec.pt

**Abstract**

Ensuring access to information in all languages is crucial for bridging disparities in communities' participation in the digital age and fostering a more inclusive and equitable society, particularly for speakers of low-resource languages (LRLs). However, enabling such access remains a significant challenge for many of these communities. Tetun, a dialect and lingua franca that became one of Timor-Leste's official languages when the country restored its independence in 2002, faces similar challenges. According to the 2015 census, Tetun is spoken by approximately 79% of the country's 1.18 million population. Despite its official status, Tetun remains underserved in language technology. Specifically, information retrieval-based solutions for Tetun do not exist, making it challenging to find relevant information on the internet and digital platforms through text-based search in the language. This work tackles these challenges by investigating retrieval strategies for text-based search that can enable the application of information retrieval (IR) techniques to develop search solutions for Tetun, with a specific focus on the ad-hoc text retrieval task. Given that language-specific algorithms, tools, and document collections for Tetun were previously unavailable, this work began by creating these foundational resources, which serve as contributions relevant to the IR and natural language processing (NLP) domains. These resources include a tokenizer, a language identification model, a stemmer, a stopword list, a document collection, a test collection, baselines for the ad-hoc text retrieval task, and a search log dataset. The contributions to IR for LRLs include: (1) A data collection pipeline tailored for LRLs to streamline the construction of textual data from the web; (2) A human-in-the-loop methodology for annotating, processing, and constructing a dataset well-suited for a variety of IR and NLP tasks; (3) A novel network-based approach for stopword detection; (4) Methodologies for developing a stemmer, designed for a language heavily influenced by loanwords, and the construction of a ground truth set for evaluating stemmer performance; (5) A detailed approach for constructing a test collection to evaluate the effectiveness of retrieval systems; (6) A methodology for establishing a robust baseline for the ad-hoc text retrieval task; and (7) Document contextualization and dual-parameter tuning strategies for hybrid text retrieval. The results from this work contribute to the development of technologies associated with the computational processing of Tetun, address gaps in its linguistic resources, and achieve impactful outcomes that elevate Tetun's status. These advancements open new opportunities for future research and innovation. Moreover, this work introduces promising methodologies that can be adapted to other languages facing similar challenges, thereby contributing to the broader advancement of IR for LRLs.

**Awarded by:** University of Porto, Porto, Portugal **on** 1 September 2005.
**Supervised by:** Sérgio Nunes.
**Available at:** https://hdl.handle.net/10216/169208.

,

## Selected Publications


Gabriel de Jesus and Sérgio Nunes. Exploring large language models for relevance judgments in Tetun. In C. Siro, M. Aliannejadi, H.A. Rahmani, N. Craswell, C.L.A. Clarke, G. Faggioli, B. Mitra, P. Thomas, and E. Yilmaz, editors, *Proceedings of The First Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024), co-located with the 10th International Conference on Online Publishing (SIGIR 2024)*, volume 3752, pages 19–30, Washington D.C., USA, July 2024a. URL https://ceur-ws.org/Vol-3752/.

Gabriel de Jesus and Sérgio Nunes. Labadain-30k+: A monolingual Tetun document-level audited dataset. In Maite Melero, Sakriani Sakti, and Claudia Soria, editors, *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 177–188, Torino, Italia, May 2024b. ELRA and ICCL. URL https://aclanthology.org/2024.sigul-1.22.

Gabriel de Jesus and Sérgio Sobral Nunes. Data collection pipeline for low-resource languages: A case study on constructing a Tetun text corpus. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 4368–4380. ELRA and ICCL, 2024c. URL https://aclanthology.org/2024.lrec-main.390.

Gabriel de Jesus, Siddharth A.K. Singh, Sérgio Nunes, and Andrew Yates. Zero-Shot and Hybrid Strategies for Tetun Ad-Hoc Text Retrieval. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, page 264–274, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400718618. doi: 10.1145/3731120.3744593. URL https://doi.org/10.1145/3731120.3744593.

Gabriel de Jesus and Sérgio Nunes. Insights into LLM-Based Conversational Search: A Study of Tetun-Speaking Users' Search Behavior. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, page 297–306, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400718618. doi: 10.1145/3731120.3744596. URL https://doi.org/10.1145/3731120.3744596.