

# **Labadain:** Resources and Applications for Tetun Language Technology

**Gabriel de Jesus**

Affiliated Researcher at INESC TEC

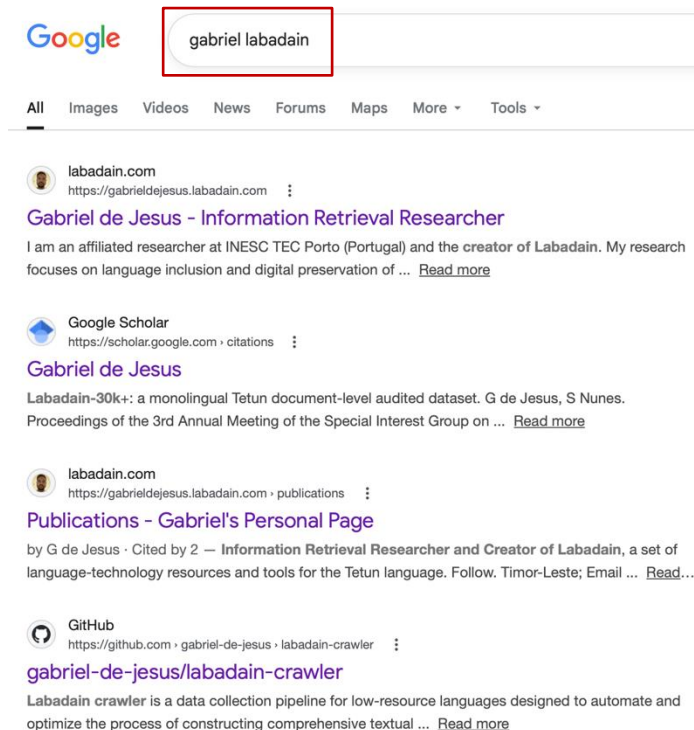
The University of Melbourne  
24 July 2026  
Melbourne, VIC, Australia



# About Me and My Research

## Text Information Retrieval in Tetun

Gabriel de Jesus



Google

gabriel labadain

All Images Videos News Forums Maps More Tools

labadain.com  
https://gabrieldejesus.labadain.com

**Gabriel de Jesus - Information Retrieval Researcher**

I am an affiliated researcher at INESC TEC Porto (Portugal) and the creator of Labadain. My research focuses on language inclusion and digital preservation of ... [Read more](#)

Google Scholar  
https://scholar.google.com/citations

**Gabriel de Jesus**

Labadain-30k+: a monolingual Tetun document-level audited dataset. G de Jesus, S Nunes. Proceedings of the 3rd Annual Meeting of the Special Interest Group on ... [Read more](#)

labadain.com  
https://gabrieldejesus.labadain.com/publications

**Publications - Gabriel's Personal Page**

by G de Jesus · Cited by 2 — **Information Retrieval Researcher and Creator of Labadain**, a set of language-technology resources and tools for the Tetun language. Follow. Timor-Leste; Email ... [Read...](#)

GitHub  
https://github.com/gabriel-de-jesus/labadain-crawler

**gabriel-de-jesus/labadain-crawler**

Labadain crawler is a data collection pipeline for low-resource languages designed to automate and optimize the process of constructing comprehensive textual ... [Read more](#)



Doctoral Program in Informatics Engineering

Supervisor: Prof. Sérgio Sobral Nunes, Ph.D.

September 2025

# Outline



Labadain  
Overview

Foundational  
Resources

Search and  
Retrieval

Conversational  
AI

Conclusion &  
Future Work

# Labadain Overview

# What is Labadain?



**Labadain** is a long-term research initiative dedicated to advancing **Tetun language technology** and enabling research, innovation, and digital inclusion

# Why Labadain?

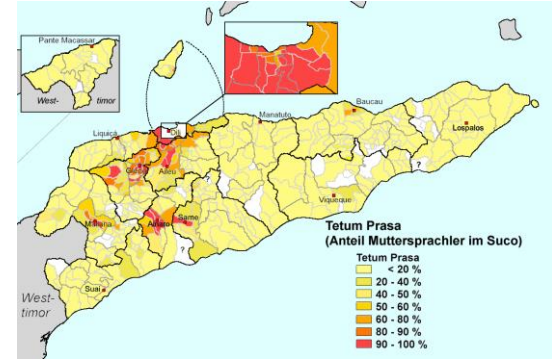


A truly **inclusive digital** transformation must speak the **languages** of its people.

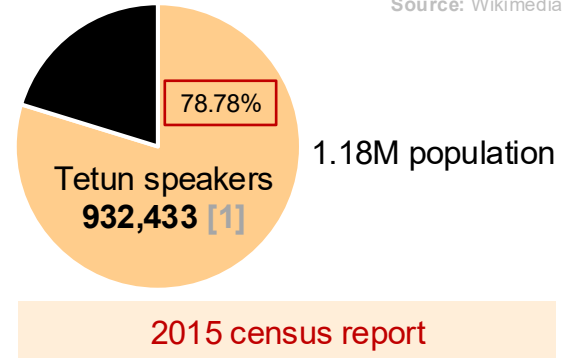


# Why Tetun?

- The most widely spoken language in **Timor-Leste**
- One of the country's **official languages**

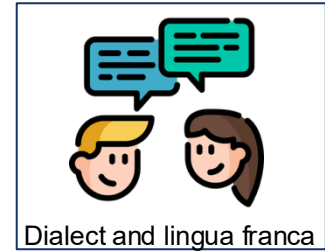
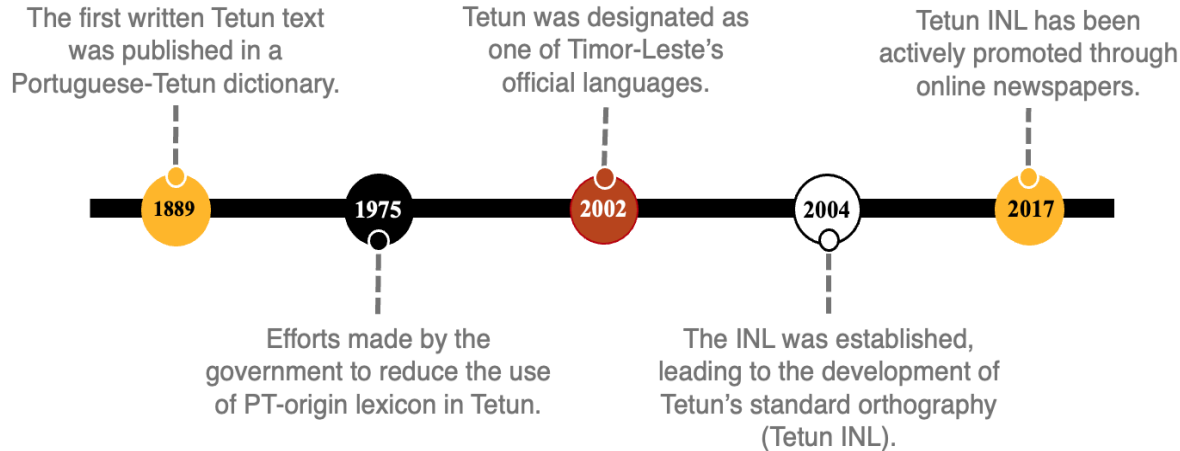


Source: Wikimedia



[1] Instituto Nacional de Estatística Timor-Leste. Timor-Leste population and housing census 2015: population distribution by administrative area – volume 2 (language). URL: <https://inet-ip.gov.tl/2023/03/09/census-2015-priority-table-population-by-language/>

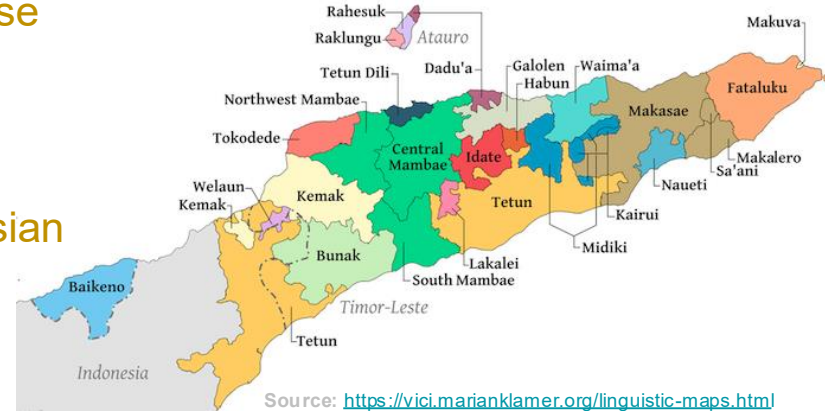
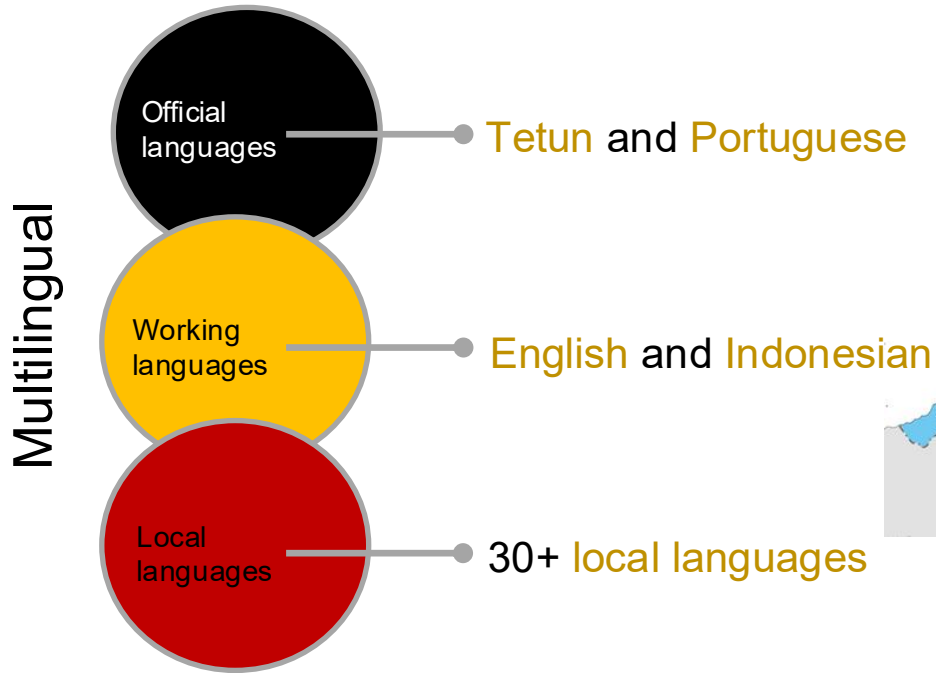
# Tetun Evolution



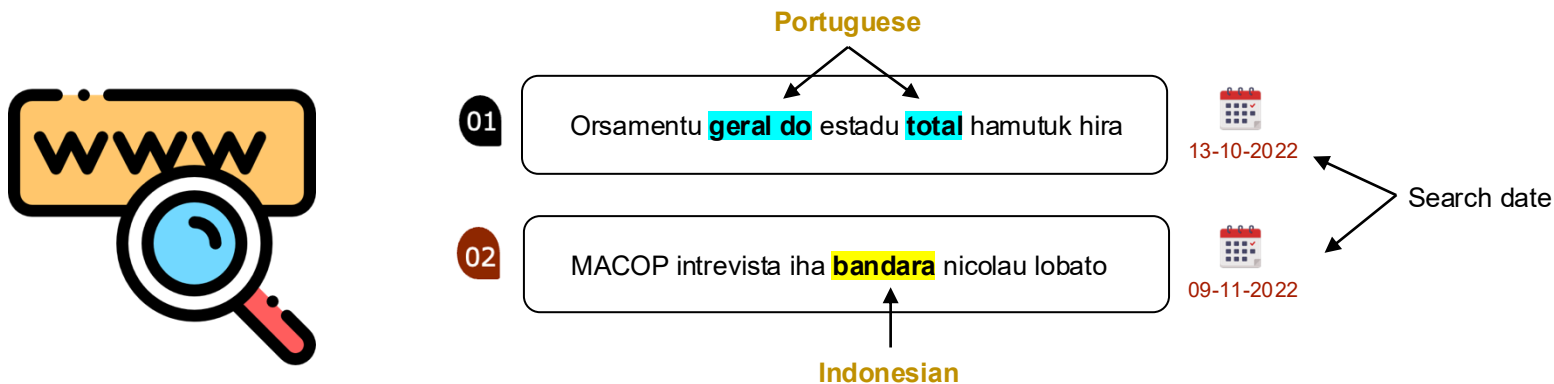
2002

Icons' source: Flaticon (<https://www.flaticon.com>)

# Linguistic Context in Timor-Leste



# Search in Multilingual Context



Search queries frequently combine Tetun with Portuguese and Indonesian words

Query source: Search logs from Timor News (<https://www.timornews.tl>)

# Research Challenges



Develop search solutions for Tetun, focusing on the ad-hoc text retrieval task.

Text processing algorithms

A text corpus

A test collection

Effective text retrieval strategies for Tetun

# Foundational Resources

# Tetun Language Technology in 2021

When I began my PhD in September 2021, Tetun had:

- No IR and NLP algorithms and tools
- No datasets
- No evaluation benchmark

Estimated Google results as of December 2022, obtained using `site:` operator:

- `.tl` top-level domain: ~4.48M indexed pages.
- `.gov.tl` government sub-domain: ~372k indexed pages.

# Labadain Crawler

## Data Collection Pipeline for Low-Resource Languages: A Case Study on Constructing a Tetun Text Corpus

Gabriel de Jesus, Sérgio Nunes

INESC TEC and Faculty of Engineering of the University of Porto (FEUP)

Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

gabriel.jesus@inesctec.pt, sergio.nunes@fe.up.pt

### Tetun tokenizer:

```
pip install tetun-tokenizer
```

```
from tetuntokenizer.tokenizer  
import TetunSimpleTokenizer
```

### Tetun LID model:

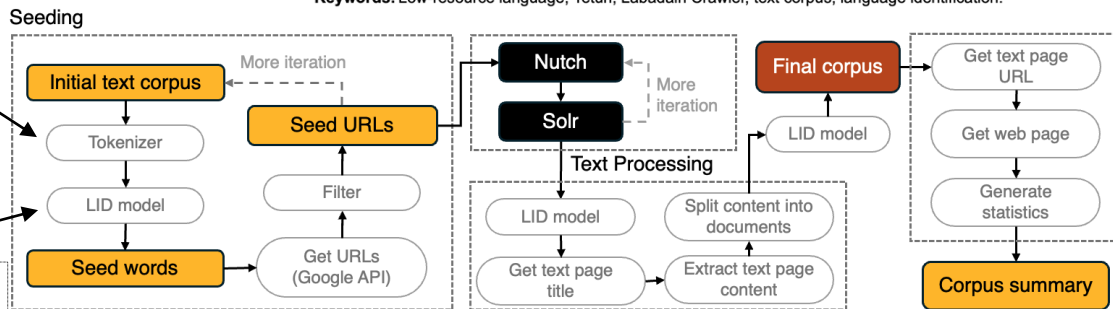
```
pip install tetun-lid
```

```
from tetunlid import lid
```



**Abstract**  
This paper proposes Labadain Crawler, a data collection pipeline tailored to automate and optimize the process of constructing textual corpora from the web, with a specific target to low-resource languages. The system is built on top of Nutch, an open-source web crawler and data extraction framework, and incorporates language processing components such as a tokenizer and a language identification model. The pipeline efficacy is demonstrated through successful testing with Tetun, one of Timor-Leste's official languages, resulting in the construction of a high-quality Tetun text corpus comprising 321.7k sentences extracted from over 22k web pages. The contributions of this paper include the development of a Tetun tokenizer, a Tetun language identification model, and a Tetun text corpus, marking an important milestone in Tetun text information retrieval.

**Keywords:** Low-resource language, Tetun, Labadain Crawler, text corpus, language identification.



Paper: <https://aclanthology.org/2024.lrec-main.390/>

Source code: <https://github.com/gabriel-de-jesus/labadain-crawler>

# Labadain Crawler

## Results from Tetun

After running approx. 46 hours, the crawler:

- Collected 22,392 text pages (containing ~9.4M tokens)
- Identified 589 PDF files and 13 PowerPoint presentations

## Distribution of text pages

By source:

| Data source category     | #text pages | Proportion |
|--------------------------|-------------|------------|
| Online newspapers        | 16,509      | 73.73%     |
| Gov. institutions        | 3,222       | 14.39%     |
| Non-gov. institutions    | 1,965       | 8.78%      |
| Educational institutions | 388         | 1.73%      |
| Blogs and Forums         | 117         | 0.52%      |
| Wikipedia                | 105         | 0.47%      |
| Personal Pages           | 57          | 0.25%      |
| Banks and courts         | 29          | 0.13%      |

By TLDs:

| Domain | #text pages | Proportion |
|--------|-------------|------------|
| .tl    | 10,741      | 47.97%     |
| .com   | 9,859       | 44.03%     |
| .org   | 1,000       | 4.47%      |
| Others | 792         | 3.54%      |

# Labadain-30k+ Dataset

## Labadain-30k+: A Monolingual Tetun Document-Level Audited Dataset

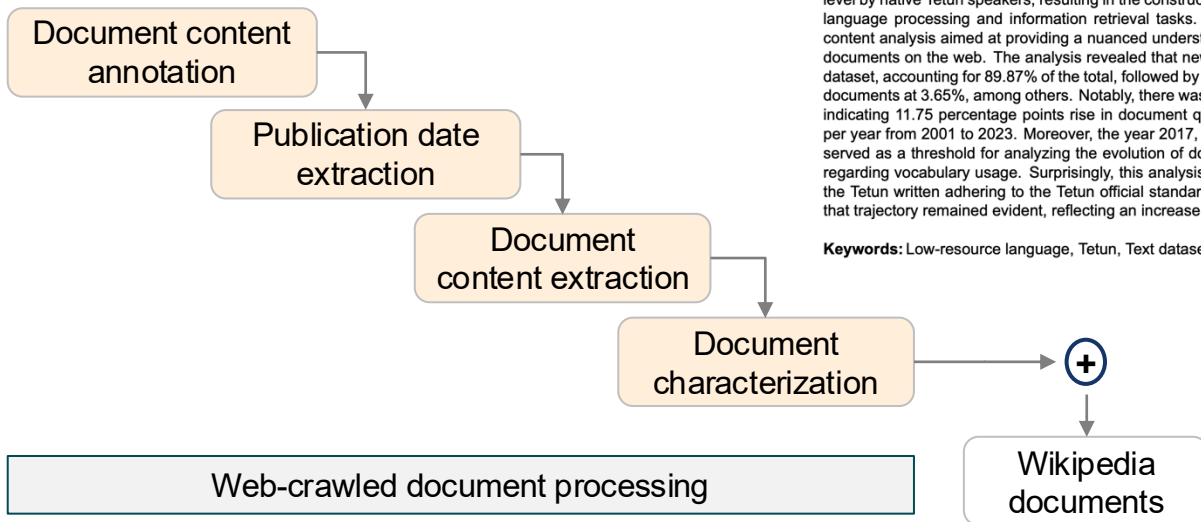
Gabriel de Jesus, Sérgio Nunes

INESC TEC and Faculty of Engineering of the University of Porto (FEUP)

Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

gabriel.jesus@inesctec.pt, sergio.nunes@fe.up.pt

## Dataset construction process



### Abstract

This paper introduces Labadain-30k+, a monolingual dataset comprising 33.6k documents in Tetun, a low-resource language spoken in Timor-Leste. The dataset was acquired through web crawling and augmented with Wikipedia documents released by Wikimedia. Both sets of documents underwent thorough manual audits at the document level by native Tetun speakers, resulting in the construction of a Tetun text dataset well-suited for a variety of natural language processing and information retrieval tasks. This dataset was employed to conduct a comprehensive content analysis aimed at providing a nuanced understanding of document composition and the evolution of Tetun documents on the web. The analysis revealed that news articles constitute the predominant documents within the dataset, accounting for 89.87% of the total, followed by Wikipedia documents at 4.34%, and legal and governmental documents at 3.65%, among others. Notably, there was a substantial increase in the number of documents in 2020, indicating 11.75 percentage points rise in document quantity, compared to an average of 4.76 percentage points per year from 2001 to 2023. Moreover, the year 2017, marked by the increased popularity of online news in Tetun, served as a threshold for analyzing the evolution of document writing on the web pre- and post-2017, specifically regarding vocabulary usage. Surprisingly, this analysis showed a significant increase of 6.12 percentage points in the Tetun written adhering to the Tetun official standard. Additionally, the persistence of Portuguese loanwords in that trajectory remained evident, reflecting an increase of 5.09 percentage points.

**Keywords:** Low-resource language, Tetun, Text dataset, Corpus content analysis.

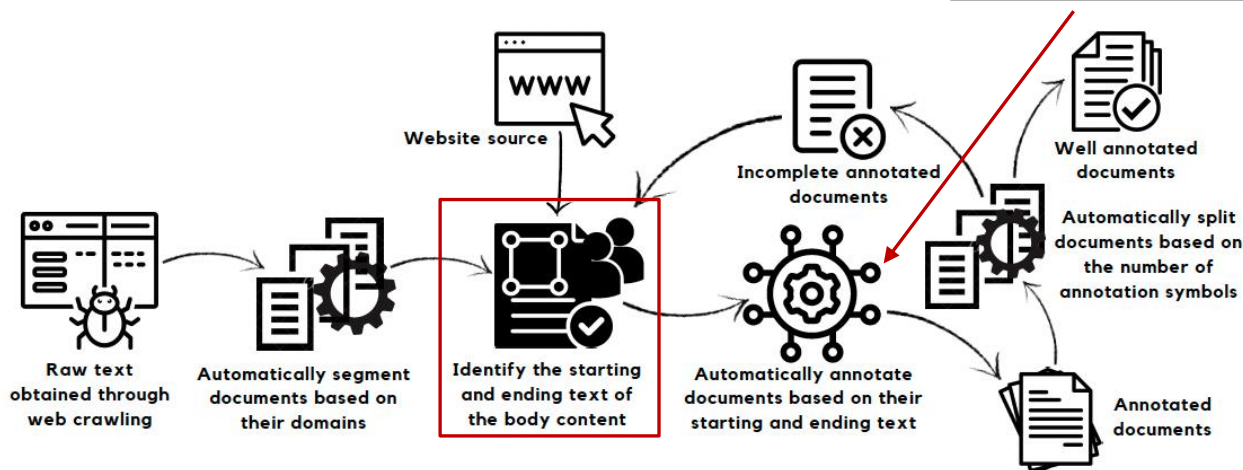
# Labadain-30k+ Dataset

## Document content annotation and processing

### Algorithm 1 Content Annotation Algorithm.

**Require:** *start\_text*, *end\_text*, *documents*, *output\_file*

```
1: for all document in documents do
2:   get title and url from document
3:   write title and url to output_file           ▷ The “annotated documents” file in Figure 4.3.
4:   get body_content from document
5:   annotation_t_counter ← 0           ▷ To control the occurrence of < t > to a maximum of two.
6:   for all text_line in body_content do
7:     get text_line_lower by lowercasing text_line and removing extra spaces
8:     if text_line_lower starts with start_text and annotation_t_counter equals 0 then
9:       write annotation string < t >, a newline, text_line, and a newline to output_file
10:      Increment annotation_t_counter by 1
11:    else if text_line_lower ends with end_text and annotation_t_counter equals 1 then
12:      write text_line, a newline, annotation string < t >, and a newline to output_file
13:      Increment annotation_t_counter by 1
14:    else
15:      write text_line and a newline to output_file
16:    end if
17:  end for
18:  write an additional newline to output_file ▷ To separate each document by two newlines.
19: end for
```



# Labadain-30k+ Dataset

## Results

### Labadain-30k+ dataset summary

|                                 |            |
|---------------------------------|------------|
| Total documents in the dataset  | 33,550     |
| Total paragraphs in the content | 334,875    |
| Total sentences in the content  | 414,370    |
| Total tokens in the corpus      | 12,300,237 |
| Vocabulary in the corpus        | 162,466    |

### Dataset statistics

|                    | Min | Max    | Avg    |
|--------------------|-----|--------|--------|
| #Paragraphs        | 1   | 1,109  | 9.98   |
| #Sentences         | 1   | 936    | 12.35  |
| #Tokens (titles)   | 1   | 29     | 9.15   |
| #Tokens (contents) | 2   | 27,166 | 357.48 |

### Dataset distribution by data sources

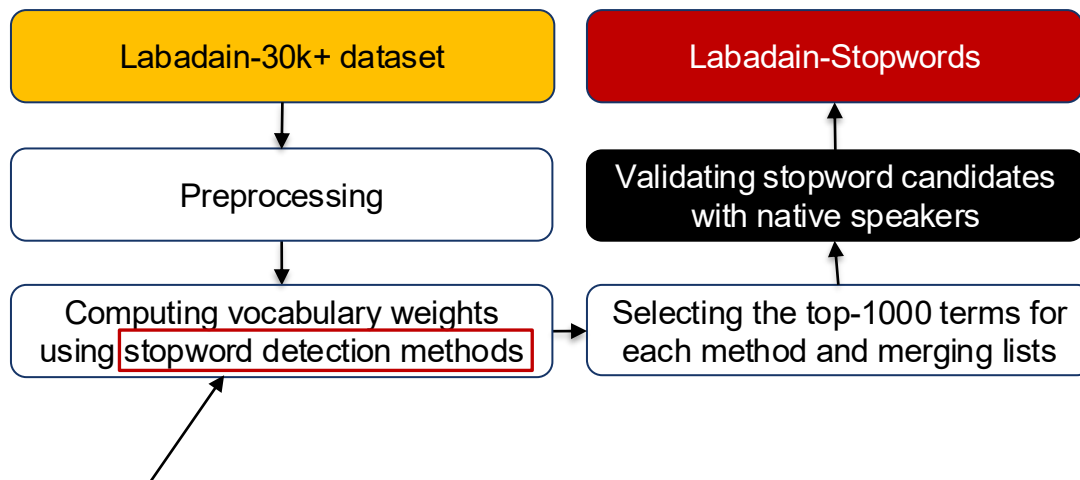
| Source           | #docs | Proportion |
|------------------|-------|------------|
| tatoli.tl        | 9,122 | 27.19%     |
| timorpost.com    | 4,687 | 13.97%     |
| naunil.com       | 3,501 | 10.43%     |
| tempotimor.com   | 2,760 | 8.23%      |
| old.timornews.tl | 2,642 | 7.87%      |

### Dataset distribution by TLDs

| TLD    | #docs  | Proportion |
|--------|--------|------------|
| .com   | 15,034 | 44.81%     |
| .tl    | 14,174 | 42.25%     |
| .org   | 2,629  | 7.84%      |
| .co    | 678    | 2.02%      |
| .pt    | 608    | 1.81%      |
| others | 427    | 1.27%      |

# Labadain Stopwords

**160** Tetun stopwords



Term-weighting methods: TF, IDF, TF-IDF  
Network properties: **In-degree, out-degree, degree**

## Network-based Approach for Stopwords Detection

Felermimo D. M. A. Ali<sup>1,3</sup>, Gabriel de Jesus<sup>2</sup>

Henrique Lopes Cardoso<sup>2</sup>, Sérgio Nunes<sup>2</sup>, Rui Sousa-Silva<sup>3</sup>

<sup>1</sup>Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC)

<sup>2</sup>Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência (INESC TEC)

<sup>3</sup>Centro de Linguística da Universidade do Porto (CLUP)

{up202100778, hlc, sergio.nunes}@fe.up.pt

gabriel.jesus@inesctec.pt, rssilva@letras.up.pt

### Abstract

Stopword lists, an essential resource for natural language processing and information retrieval, are often unavailable for low-resource languages. Creating these lists is time-consuming

systems (Croft et al., 2009).

Many existing methods often either rely on a predefined list of stopwords or are computed using traditional unsupervised methods, such as term frequency (TF) (Baeza-Yates and Ribeiro-Neto, 2011;

Paper: <https://arxiv.org/abs/2412.11758>

Dataset: <https://doi.org/10.25747/PG2V-KX70>

# Labadain Stemmer

## Tetun linguistic overview

### Orthography

Tetun uses **Latin** script and has **5 vowels** and **21 consonants**.

### Examples of basic phrases

| Tetun                | English            |
|----------------------|--------------------|
| Dadeer di'ak!        | Good morning!      |
| Di'ak ka la'e?       | How are you?       |
| Ita-nia naran saida? | What is your name? |

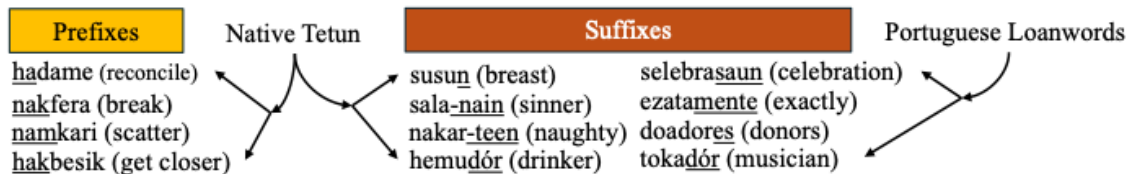
### Portuguese loanwords

“Ha’u **agradese** **tebes** **tanba** halo **investigasaun** **ida-ne’ebé** **krusiál**.”

Labels: **Verb** (agradese), **Noun** (tebes), **Adjective** (krusiál)

Up to 40% are **Portuguese loanwords** [2–4]

### Affixes



[2] de Jesus and Nunes (2024). Labadain-30k+: A Monolingual Tetun Document-Level Audited Dataset. Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages, 177-188, LREC-COLING, 2024

[3] van-Klinken et al. (2019). Language Contact and Gender in Tetun Dili: What Happens When Austronesian Meets Romance? University of Hawai'i Press 58, 59–91, 2019.

[4] Greksakova (2018). Tetun in Timor-Leste: The role of language contact in its development. PhD thesis, Universidade de Coimbra, Portugal, 2018.

# Labadain Stemmer

## Stemmer variants

01

**Light** stemmer: Removes **suffixes** from **Portuguese-derived** words.

02

**Moderate** stemmer: Removes **suffixes** from both Portuguese-derived words and **native Tetun** words.

03

**Heavy** stemmer: Removes **suffixes** from Portuguese-derived words and native Tetun, and **prefixes** from **native Tetun** words.

Paper: <https://arxiv.org/abs/2412.11758>

Algorithms: <https://github.com/gabriel-de-jesus/labadain-stemmer>

# Labadain Stemmer

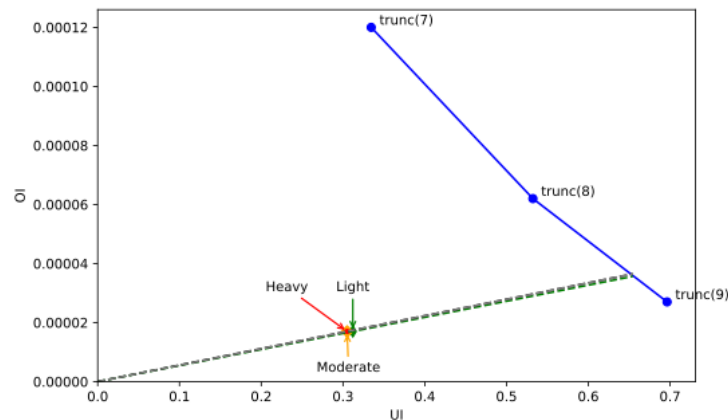
## Evaluation and results

Evaluation using the Paice metric [5]

### Creation of conceptual groups

"ajente": ["ajénsia", "ajénsias"]  
"akompañ": ["akompañ", "akompañamentu", "akompañante"]  
"akontes": ["akontese", "akontesimentu", "akontesimentus"]  
"hatete": ["hatete", "hateten"]  
"kbiit": ["kbiit-laek", "kbiit-na'in", "kbiit"]  
"komunik": ["komunikadu", "komunikasaun", "komunikativa"]  
"otél": ["otél"]

Plot of ERRT value



### Stemming algorithm performance

|          | UI       | OI       | SW       | ERRT     |
|----------|----------|----------|----------|----------|
| Light    | 0.312062 | 0.000017 | 0.000056 | 0.481367 |
| Moderate | 0.305049 | 0.000017 | 0.000057 | 0.472802 |
| Heavy    | 0.305049 | 0.000017 | 0.000057 | 0.472802 |

[5] Paice (1994). An evaluation method for stemming algorithms. Proceedings of the 17th ACM SIGIR, Dublin, Ireland, July 3-6, 1994.

# Labadain Avaliadór

## A Tetun **test collection** for ad-hoc text retrieval

### Query formulation

- **Five (5)** native Tetun-speaking **students** formulated queries using established guidelines.
- Queries were derived from **Timor News** search logs and **Google Search Console** for Timor News.

### Example of query revision

| Original query log                   | Reformulated query                                      |
|--------------------------------------|---------------------------------------------------------|
| Problema lixu (Waste problem)        | Problema lixu <b>iha Dili</b> (Waste problem in Dili)   |
| Soe bebe (Baby abandonment)          | <b>Kazu</b> soe <b>bebé</b> (Case of baby abandonment)  |
| Konsumu tabaku (Tobacco consumption) | <b>Dadus</b> konsumu tabaku (Tobacco consumption data)  |
| Kilat ilegal (Illegal firearms)      | <b>Problema</b> kilat ilegal (Illegal firearms problem) |



New term(s) added



Term revised

# Labadain **Avaliadór**

## A Tetun **test collection** for ad-hoc retrieval

A total of **61 topics** were created

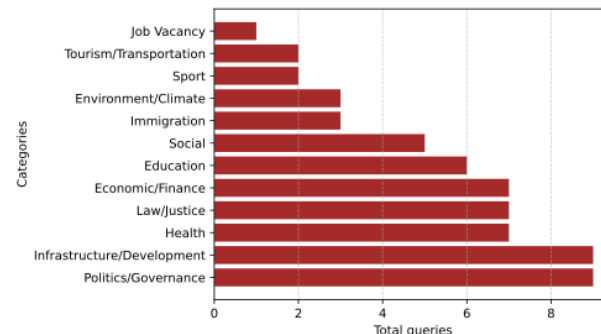
Example of a topic formatted according to TREC guidelines

|                                                                                                                                                                                                                                                                                                                      | English version                                                                                                                                                                                                                                                                                    |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>&lt;top&gt;</code>                                                                                                                                                                                                                                                                                             | <code>&lt;top&gt;</code>                                                                                                                                                                                                                                                                           |
| <code>&lt;num&gt; <b>Number:</b> TTI-00001</code>                                                                                                                                                                                                                                                                    | <code>&lt;num&gt; Number: TTI-00001</code>                                                                                                                                                                                                                                                         |
| <code>&lt;title&gt; Topic: Prevensaun moras HIV-SIDA</code>                                                                                                                                                                                                                                                          | <code>&lt;title&gt; Topic: Prevention of HIV-AIDS</code>                                                                                                                                                                                                                                           |
| <code>&lt;desc&gt; <b>Description:</b><br/>Informasaun kona-ba mekanizmu prevensaun moras HIV-SIDA, inklui moras HIV-SIDA ninia kauza.</code>                                                                                                                                                                        | <code>&lt;desc&gt; Description:<br/>Information about HIV-AIDS prevention mechanisms, including the causes of HIV-AIDS.</code>                                                                                                                                                                     |
| <code>&lt;narr&gt; <b>Narrative:</b><br/>Dokumentu relevante bainhira kontein informasaun kona-ba esforsu no mekanizmu sira hodi prevene HIV-SIDA. Informasaun kona-ba kauza hosi HIV-SIDA mós relevante. Informasaun kona-ba seksu ne'ebé la detalhe kona-ba ninia ligasaun ho moras HIV-SIDA, la relevante.</code> | <code>&lt;narr&gt; Narrative:<br/>Relevant documents are those that contain information about efforts and mechanisms to prevent HIV-AIDS. Information about the causes of HIV-AIDS is also relevant. Information about sex that does not detail its connection to HIV-AIDS is not relevant.</code> |
| <code>&lt;/top&gt;</code>                                                                                                                                                                                                                                                                                            | <code>&lt;/top&gt;</code>                                                                                                                                                                                                                                                                          |

### Summary of queries

| Description             | Value |
|-------------------------|-------|
| Queries                 | 61    |
| Three-word queries      | 37    |
| Four-word queries       | 22    |
| Five-word queries       | 2     |
| Average words per query | 3.43  |

### Distribution of queries over categories



# Labadain Avaliadór

## A Tetun test collection for ad-hoc retrieval

### Relevance judgments

### Web interface used by human assessors

Query-Documents Assessment Platform <sup>1</sup>

This platform is designed to streamline the workflow of query-documents relevance Assessment.

List of Queries

Sensu uma-kain 2022

Evaluate

Progresu dezenvolvimentu Infraestrutur

Evaluate

Preparasaun ba expo Dubai

Evaluate

Agresan fizika hosi PNTL

Evaluate

Panorama eleisaun PR 2022

Evaluate

Inundasaun iha Dili

Evaluate

Query-Documents Relevance Assessment <sup>2</sup>

Step 1 of 2: Query Description

Topic

Sensu uma-kain 2022

Information need

Informasaun kona-ba sensu uma-kain 2022.  
Dokumentu sira ne'ebé fô sai informasaun kona-ba implementasaun no dadus estatistika ba sensu uma-kain 2022.  
relevante. Karik dokumentu kontein informasaun kona-ba sensu seluk, la relevante.

Relevant documents

Next

Query-Documents Relevance Assessment <sup>3</sup>

Step 2 of 2: Evaluation and Submission

Uma-kain 40% partisipa ona sensu populasau 2022 iha Bobonaro

Tatoli II

BOBONARO, 13 setembru 2022 — Diretor Estatistika munisipiu Bobonaro, Marti.

Relevance score: La relevante Relevante naton Relevante Relevante tebes

Uma-kain 21.6% iha teritóriu partisipa ona sensu populasau 2022

Tatoli II

DILI, 10 setembru 2022 — Governu liuhosi Direasaun Jerál Estatistika (DGE, s.

Relevance score: La relevante Relevante naton Relevante Relevante tebes

PM Taur no família partisipa ona sensu populasau no uma-kain 20.

Tatoli II

DILI, 30 setembru 2022 (TATOLI) —Primeiru-Ministru (PM), Taur Matan Ruak, .

Relevance score: La relevante Relevante naton Relevante Relevante tebes

Abitante miliaun 1 resin partisipa ona sensu uma-kain 2022

Tatoli II

DILI, 07 outubru 2022 —Direasaun Jerál Estatistika (DGE, sigla portugues) ho.

Relevance score: La relevante Relevante naton Relevante Relevante tebes

Ministériu Finansas Lansa Rezultadu Prelimináriu Sensu Populasau.

timorpost.com

Dili — Governu Timor-Leste liuhusi Ministériu Finansas (MF), Rui Augusto G.

Four graded levels of relevance [8]:

- 0 – Non-relevant
- 1 – Marginally relevant
- 2 – Relevant
- 3 – Highly relevant

### Summary of the final test collection (Labadain-Avaliadór)

| Description                                            | Value |
|--------------------------------------------------------|-------|
| Total number of topics                                 | 59    |
| Total number of <i>qrels</i>                           | 5,900 |
| Minimum number of relevant documents per query*        | 11    |
| Maximum number of relevant documents per query         | 99    |
| Average number of relevant documents per query         | 36.76 |
| The standard deviation of relevant documents per query | 20.89 |

[8] Sormunen (2002). Liberal relevance criteria of TREC: counting on negligible documents? Proceedings of the 25<sup>th</sup> ACM SIGIR, Tampere, Finland, August 11-15, 2002.

## LLMs for Relevance Judgments

### Using the Labadain-Avaliadór collection

#### Few-shot prompting with four examples

Prompt 6.3.1: Example of the system prompt.

You are an expert assessor and you are tasked with assessing the relevance between the input query and its corresponding document, assigning a score from 0 to 3. A score of 0 indicates irrelevant; 1 is marginally relevant; 2 is relevant; and 3 is highly relevant.

Example:

**query:** “Kursu mestradu no pós-graduaun UNTL”

**document:** “Kursu Desportu UNTL sei realiza graduaun dahuluk tinan ne’e”

**reason:** “The query is about postgraduate and master’s courses at UNTL, whereas the document focuses on a sports course. Despite both courses in the query and document being offered at UNTL, the sports course in the document is not specifically designed for postgraduate or master’s levels. Thus, the document is only marginally relevant”.

**score:** 1

The query and document to be evaluated are the following:

**query:** {query}

**document:** {document}

Your response must be in JSON format with the first field is “reason”, explaining your reasoning, and the second field is “score”.

#### Abstract

The Cranfield paradigm has served as a foundational approach for developing test collections, with relevance judgments typically conducted by human assessors. However, the emergence of large language models (LLMs) has introduced new possibilities for automating these tasks. This paper explores the feasibility of using LLMs to automate relevance assessments, particularly within the context of low-resource languages. In our study, LLMs are employed to automate relevance judgment tasks, by providing a series of query-document pairs in Tetun as the input text. The models are tasked with assigning relevance scores to each pair, where these scores are then compared to those from human annotators to evaluate the inter-annotator agreement levels. Our investigation reveals results that align closely with those reported in studies of high-resource languages.

#### Keywords

Large language models, Relevance judgments, Low-resource languages, Tetun

### LLMs evaluated and Cohen’s $k$ agreement with human assessors

|       | LLaMA3 70B    | Claude3 Haiku | GPT-3.5-turbo-0125 |
|-------|---------------|---------------|--------------------|
| Human | <b>0.2634</b> | 0.2450        | 0.2462             |

### Comparison with the previous studies

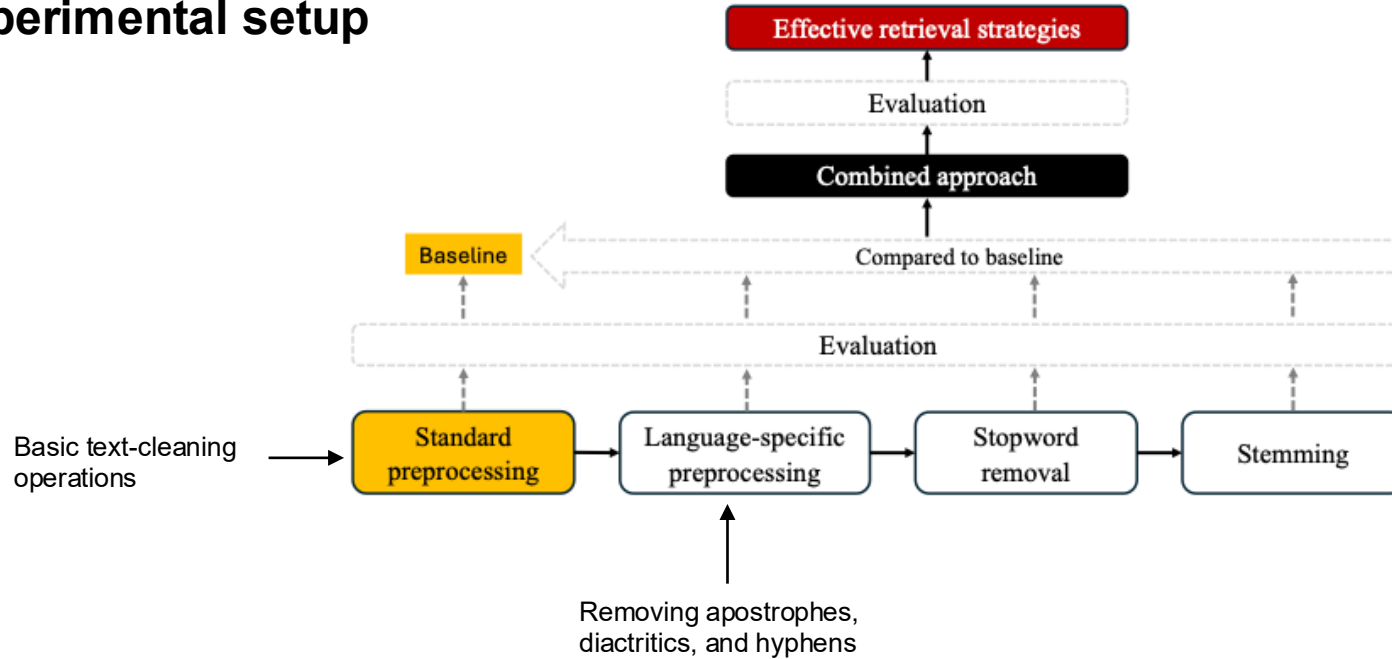
|                 | LLM        | Cohen’s $k$ Score |
|-----------------|------------|-------------------|
| Bueno et al.    | GPT-4      | 0.31              |
| Thomas et al.   | GPT-4      | 0.26 – 0.64       |
| Faggioli et al. | GPT 3.5    | 0.26              |
| <b>Ours</b>     | LLaMA3 70B | 0.26              |

Paper: <https://ceur-ws.org/Vol-3752/paper2.pdf>

# Search and Retrieval

# Labadain Baselines

## Experimental setup



# Labadain Baselines

## Evaluation and results

### Short-text retrieval: combined approach

DFR BM25 achieved the best performance at cutoff up to ten (k=10) across Precision, MAP, and NDCG.

Hiemstra LM performed best at higher cutoff (k >= 20) across Precision, MAP, and NDCG.

DFR BM25 improved at NDCG@10 by up to 30.35% relative to baseline.

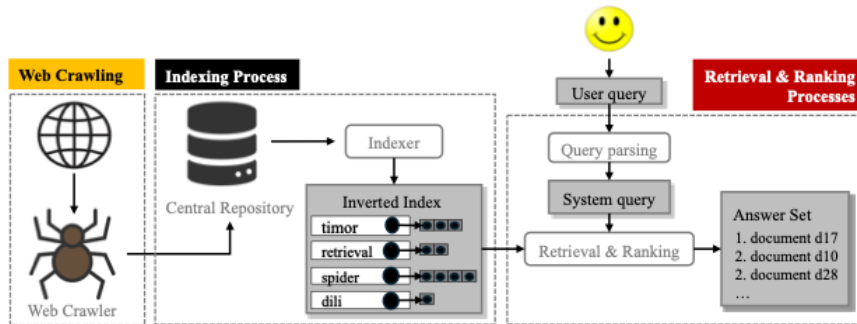
Table 20 Effectiveness of Combined Text Preprocessing Techniques in Short-Text Retrieval. Values highlighted with a green background indicate the best score at the respective metric and cutoff. Statistical significance compared to the baseline is indicated by <sup>†</sup> ( $p < 0.05$ ) and <sup>‡</sup> ( $p < 0.01$ ), based on a paired t-test over per-query scores.

| Retrieval Strategies                                                                              | Model        | Precision at Cutoff |                     |                     | MAP at Cutoff       |                     |                     | NDCG at Cutoff      |                     |                     | MAP                 | NDCG                |
|---------------------------------------------------------------------------------------------------|--------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
|                                                                                                   |              | @5                  | @10                 | @20                 | @5                  | @10                 | @20                 | @5                  | @10                 | @20                 |                     |                     |
| Baseline                                                                                          | BM25         | 0.8169              | 0.7763              | 0.6602              | 0.1444              | 0.2568              | 0.3903              | 0.6801              | 0.6668              | 0.6454              | 0.5925              | 0.7408              |
|                                                                                                   | DFR BM25     | 0.8169              | 0.7763              | 0.6619              | 0.1440              | 0.2563              | 0.3901              | 0.6811              | 0.6666              | 0.6454              | 0.5926              | 0.7407              |
|                                                                                                   | TF-IDF       | 0.8136              | 0.7746              | 0.6458              | 0.1432              | 0.2546              | 0.3825              | 0.6739              | 0.6640              | 0.6380              | 0.5802              | 0.7364              |
|                                                                                                   | Dirichlet LM | 0.7898              | 0.7525              | 0.6398              | 0.1299              | 0.2361              | 0.3671              | 0.6359              | 0.6356              | 0.6174              | 0.5780              | 0.7208              |
|                                                                                                   | Hiemstra LM  | 0.8136              | 0.7695              | 0.6669              | 0.1428              | 0.2521              | 0.3928              | 0.6670              | 0.6588              | 0.6465              | 0.6090              | 0.7435              |
| Remove apostrophes                                                                                | BM25         | 0.8237              | 0.7763              | 0.6644              | 0.1453              | 0.2572              | 0.3930              | 0.6866              | 0.6685              | 0.6499              | 0.5938              | 0.7419              |
|                                                                                                   | DFR BM25     | 0.8237              | 0.7763              | 0.6661              | 0.1450              | 0.2568              | 0.3929              | 0.6878              | 0.6684              | 0.6515              | 0.5942              | 0.7420              |
|                                                                                                   | TF-IDF       | 0.8203              | 0.7746              | 0.6500              | 0.1443              | 0.2552              | 0.3854              | 0.6808              | 0.6660              | 0.6428              | 0.5818              | 0.7377              |
|                                                                                                   | Dirichlet LM | 0.7898              | 0.7542              | 0.6432              | 0.1301              | 0.2365              | 0.3686              | 0.6380              | 0.6376              | 0.6206              | 0.5794              | 0.7219              |
|                                                                                                   | Hiemstra LM  | 0.8169              | 0.7712              | 0.6712              | 0.1429              | 0.2529              | 0.3953              | 0.6725              | 0.6609              | 0.6507              | 0.6102              | 0.7443              |
| Remove hyphens                                                                                    | BM25         | 0.8271              | 0.7881              | 0.6856              | 0.1459              | 0.2616              | 0.4069              | 0.7143              | 0.7014              | 0.6871              | 0.6498              | 0.8130              |
|                                                                                                   | DFR BM25     | 0.8271              | 0.7881              | 0.6856              | 0.1459              | 0.2616              | 0.4070              | 0.7138              | 0.7016              | 0.6873              | 0.6506              | 0.8135              |
|                                                                                                   | TF-IDF       | 0.8271              | 0.7814              | 0.6805              | 0.1457              | 0.2573              | 0.4028              | 0.7118              | 0.6979              | 0.6845              | 0.6402              | 0.8077              |
|                                                                                                   | Dirichlet LM | 0.7898              | 0.7576              | 0.6797              | 0.1322              | 0.2420              | 0.3860              | 0.6578              | 0.6615              | 0.6652              | 0.6679              | 0.8039              |
|                                                                                                   | Hiemstra LM  | 0.8339              | 0.7881              | 0.6898              | 0.1472              | 0.2635              | 0.4143              | 0.7142              | 0.6980              | 0.6941              | 0.6841              | 0.8239              |
| Remove stopwords                                                                                  | BM25         | 0.8102              | 0.7729              | 0.6695              | 0.1438              | 0.2547              | 0.3976              | 0.6693              | 0.6600              | 0.6488              | 0.6030              | 0.7443              |
|                                                                                                   | DFR BM25     | 0.8102              | 0.7729              | 0.6712              | 0.1439              | 0.2549              | 0.3984              | 0.6695              | 0.6602              | 0.6503              | 0.6049              | 0.7451              |
|                                                                                                   | TF-IDF       | 0.8102              | 0.7729              | 0.6686              | 0.1439              | 0.2549              | 0.3975              | 0.6691              | 0.6600              | 0.6484              | 0.6018              | 0.7438              |
|                                                                                                   | Dirichlet LM | 0.8034              | 0.7593              | 0.6653              | 0.1317              | 0.2379              | 0.3803              | 0.6329              | 0.6315              | 0.6299              | 0.5936              | 0.7255              |
|                                                                                                   | Hiemstra LM  | 0.8203              | 0.7678              | 0.6864              | 0.1437              | 0.2521              | 0.4036              | 0.6702              | 0.6587              | 0.6577              | 0.6189              | 0.7483              |
| Remove apostrophes and hyphens                                                                    | BM25         | 0.8881 <sup>†</sup> | 0.8373              | 0.7153              | 0.1553              | 0.2796              | 0.4304              | 0.7500              | 0.7347              | 0.7133              | 0.6648              | 0.8213              |
|                                                                                                   | DFR BM25     | 0.8881 <sup>†</sup> | 0.8390 <sup>†</sup> | 0.7169              | 0.1553              | 0.2804 <sup>†</sup> | 0.4313              | 0.7512 <sup>†</sup> | 0.7356 <sup>†</sup> | 0.7149              | 0.6664              | 0.8219              |
|                                                                                                   | TF-IDF       | 0.8780              | 0.8322              | 0.7119              | 0.1543              | 0.2759              | 0.4273              | 0.7401              | 0.7288              | 0.7086              | 0.6553              | 0.8149              |
|                                                                                                   | Dirichlet LM | 0.8407              | 0.8034              | 0.7068              | 0.1390              | 0.2561              | 0.4099              | 0.6834              | 0.6920              | 0.6829              | 0.6713              | 0.8018              |
|                                                                                                   | Hiemstra LM  | 0.8780              | 0.8305              | 0.7263              | 0.1524              | 0.2743              | 0.4339              | 0.7379              | 0.7245              | 0.7147              | 0.6955              | 0.8282              |
| Remove hyphens and stopwords                                                                      | BM25         | 0.8814              | 0.8237              | 0.7237              | 0.1576              | 0.2720              | 0.4356              | 0.7394              | 0.7221              | 0.7130              | 0.6752              | 0.8220              |
|                                                                                                   | DFR BM25     | 0.8847              | 0.8254              | 0.7237              | 0.1585              | 0.2729              | 0.4366              | 0.7416              | 0.7228              | 0.7139              | 0.6764              | 0.8224              |
|                                                                                                   | TF-IDF       | 0.8847              | 0.8237              | 0.7229              | 0.1585              | 0.2722              | 0.4355              | 0.7416              | 0.7220              | 0.7126              | 0.6715              | 0.8202              |
|                                                                                                   | Dirichlet LM | 0.8508              | 0.8102              | 0.7220              | 0.1409              | 0.2549              | 0.4036              | 0.6925              | 0.6925              | 0.6833              | 0.6683              | 0.8065              |
|                                                                                                   | Hiemstra LM  | 0.8746              | 0.8305              | 0.7305 <sup>†</sup> | 0.1524              | 0.2720              | 0.4372 <sup>†</sup> | 0.7294              | 0.7211              | 0.7152 <sup>†</sup> | 0.7040 <sup>†</sup> | 0.8288              |
| Remove hyphens, apostrophes, and stopwords                                                        | BM25         | 0.8814              | 0.8220              | 0.7212              | 0.1580              | 0.2725              | 0.4342              | 0.7395              | 0.7211              | 0.7123              | 0.6743              | 0.8223              |
|                                                                                                   | DFR BM25     | 0.8847              | 0.8237              | 0.7212              | 0.1589 <sup>†</sup> | 0.2735              | 0.4353              | 0.7418              | 0.7218              | 0.7131              | 0.6756              | 0.8228              |
|                                                                                                   | TF-IDF       | 0.8847              | 0.8220              | 0.7203              | 0.1589 <sup>†</sup> | 0.2727              | 0.4341              | 0.7417              | 0.7210              | 0.7118              | 0.6705              | 0.8205              |
|                                                                                                   | Dirichlet LM | 0.8508              | 0.8068              | 0.7144              | 0.1418              | 0.2546              | 0.4142              | 0.6949              | 0.6913              | 0.6872              | 0.6831              | 0.8053              |
|                                                                                                   | Hiemstra LM  | 0.8746              | 0.8254              | 0.7263              | 0.1528              | 0.2711              | 0.4353              | 0.7287              | 0.7190              | 0.7134              | 0.7029              | 0.8289 <sup>†</sup> |
| Remove hyphens and stopwords, and apply light stemmer                                             | BM25         | 0.8678              | 0.8169              | 0.7110              | 0.1540              | 0.2698              | 0.4267              | 0.7367              | 0.7203              | 0.7053              | 0.6602              | 0.8154              |
|                                                                                                   | DFR BM25     | 0.8712              | 0.8153              | 0.7110              | 0.1549              | 0.2695              | 0.4268              | 0.7389              | 0.7200              | 0.7058              | 0.6608              | 0.8157              |
|                                                                                                   | TF-IDF       | 0.8712              | 0.8169              | 0.7102              | 0.1549              | 0.2701              | 0.4267              | 0.7389              | 0.7201              | 0.7048              | 0.6584              | 0.8145              |
|                                                                                                   | Dirichlet LM | 0.8339              | 0.7932              | 0.7034              | 0.1390              | 0.2503              | 0.4056              | 0.6936              | 0.6860              | 0.6622              | 0.6202              | 0.8023              |
|                                                                                                   | Hiemstra LM  | 0.8576              | 0.8203              | 0.7229              | 0.1496              | 0.2678              | 0.4289              | 0.7256              | 0.7171              | 0.7083              | 0.6831              | 0.8208              |
| Average performance gains of the best model compared to the baseline                              |              | 8.72%               | 8.08%               | 9.54%               | 10.66%              | 9.40%               | 11.30%              | 10.29%              | 30.35%              | 10.63%              | 15.60%              | 11.49%              |
| Average performance gains of the best model compared to individual text preprocessing techniques  |              | 7.60%               | 7.39%               | 6.16%               | 9.75%               | 8.19%               | 6.93%               | 7.23%               | 7.45%               | 6.09%               | 8.33%               | 7.58%               |
| Average performance gains of the best model compared to others within the same retrieval strategy |              | 2.65%               | 1.62%               | 1.03%               | 5.54%               | 3.41%               | 1.43%               | 3.35%               | 2.22%               | 1.06%               | 3.94%               | 1.38%               |

# Labadain Search

## Architecture and prototype

The research resources were integrated into a working Tetun search prototype



<https://search.labadain.com>



Eleisaun prezidensiál 2022

Buka

Labadain

Eleisaun prezidensiál 2022

Buka

Total dokumentu ne'ebé Labadain hetan: 1064

[CNE Viqueque halo evaluasaun ba eleisaun prezidensiál 2022](#)

[Laballist](#)

VIQUEQUE, 27 maiu 2022 –Komisaun Nasionál Eleisaun (CNE, sigla português) hamutu...

Data Publikasaun: 27/05/2022

[Ekipa husi UE Sei Observa Eleisaun Prezidensiál 2022](#)

[www.naunil.com](#)

Dili, Visi-Ministru Interior Antonio Armindo, Informa Ekipa husi Uniaun Europei...

Data Publikasaun: 27/10/2021

[CNE-STAE Rona Perspektiva Eleisaun Prezidensiál 2022](#)

[timorpost.com](#)

BAUKAU (Timor Post)- Komisaun Nasionál Eleisaun no Sekretariadu Tékniku Adminis...

# Zero-Shot and Hybrid Retrieval

## Beyond Lexical Retrieval

We investigated dense retrieval under **zero-shot** and **hybrid** strategies in **out-of-distribution** scenarios.

## Hypothesis

**Augmenting** document **titles** with LLM-generated **summaries** may improve zero-shot dense retrieval.

### Zero-Shot and Hybrid Strategies for Tetun Ad-Hoc Text Retrieval

Gabriel de Jesus  
INESC TEC / University of Porto  
Porto, Portugal  
gabriel.jesus@inesctec.pt

Sérgio Nunes  
INESC TEC / University of Porto  
Porto, Portugal  
sergio.nunes@fe.up.pt

Siddharth AK Singh  
IRLab / University of Amsterdam  
Amsterdam, Netherlands  
s.a.k.singh@uva.nl

Andrew Yates  
HLTCOE / Johns Hopkins University  
Baltimore, Maryland, United States  
andrew.yates@jhu.edu

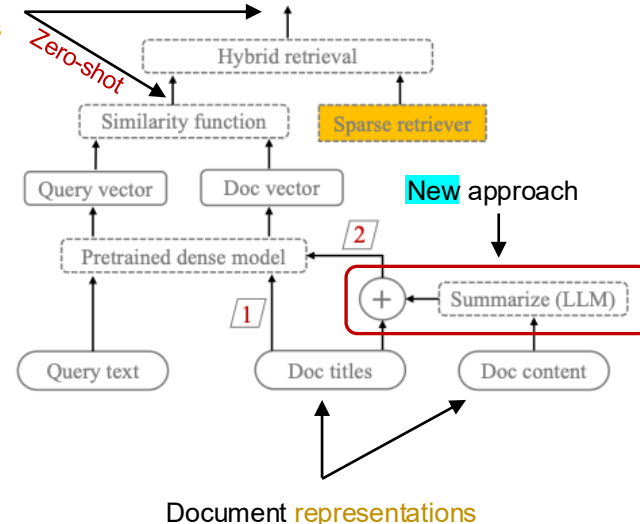
#### Abstract

Dense retrieval models are generally trained using supervised learning approaches for representation learning, which require a labeled dataset (i.e., query-document pairs). However, training such models from scratch is not feasible for most languages, particularly under-resourced ones, due to data scarcity and computational constraints. As an alternative, pretrained dense retrieval models can

ACM Reference Format:  
Gabriel de Jesus, Siddharth AK Singh, Sérgio Nunes, and Andrew Yates. 2025. 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 '25)*, July 18, 2025, Padua, Italy. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3731120.374599>

## Settings

Retrieval  
strategies



# Zero-Shot and Hybrid Retrieval

## Document summarization

### Configurations

- **Approach:** one-shot prompting, with one example per LLM call.
- **LLM:** Claude 3 Haiku.
- **Prompt engineering:** prompt variants were iteratively refined through qualitative assessment.

Prompt 7.4.1: Details of the LLM Prompt.

**User prompt:**

Provide a concise contextual description that summarizes the following Tetun document. Do not include any additional text or generate hallucinated content. Do not respond in bullet points, and write exactly one paragraph.

{document}

**Follow this example:**

Input document in Tetun:

{document\_example}

Expected summary:

{summary\_example}

**System prompt:**

You are an expert in Tetun and text understanding. Assume the document is entirely in Tetun and respond using accurate Tetun grammar. Keep your response in Tetun, including correct spelling, punctuation, and other linguistic aspects. Ignore incomplete sentences.

# Zero-Shot and Hybrid Retrieval

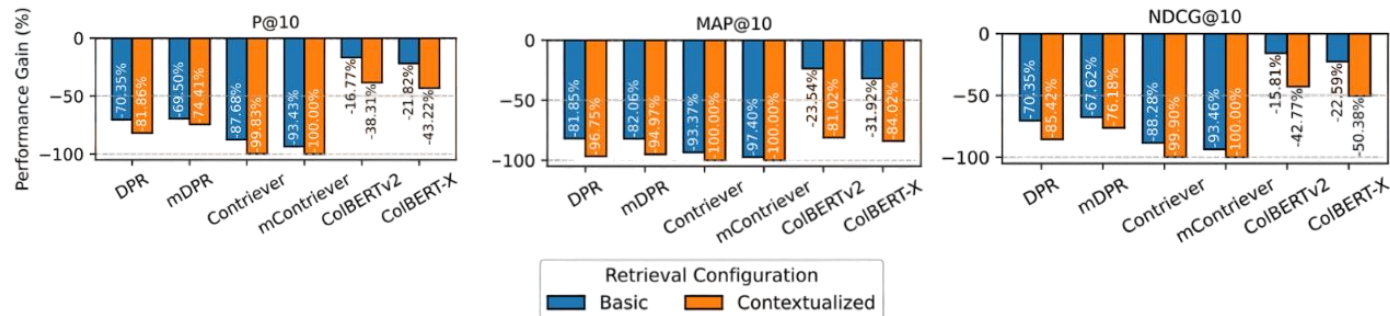
## Experiment and results

### Baselines

| Model       | P@10          | MAP@10        | NDCG@10       |
|-------------|---------------|---------------|---------------|
| BM25        | 0.8373        | 0.2796        | 0.7347        |
| DFR BM25    | <b>0.8390</b> | <b>0.2804</b> | <b>0.7356</b> |
| Hiemstra LM | 0.8305        | 0.2743        | 0.7245        |

Model performance was compared with the DFR BM25 baseline.

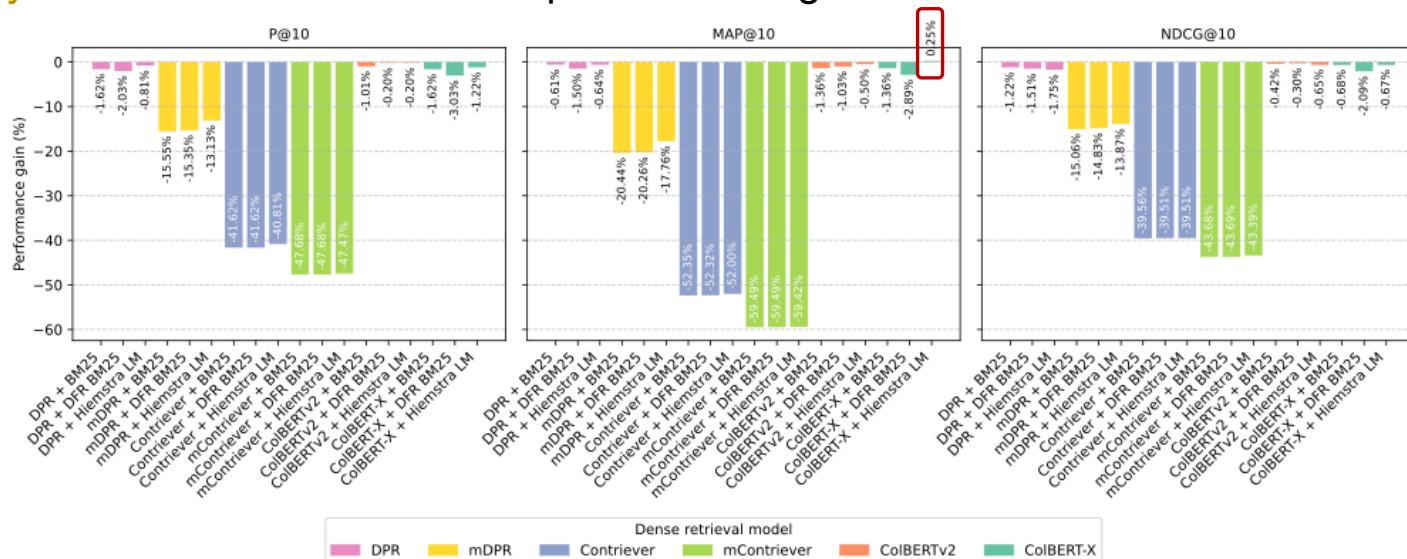
### Zero-shot retrieval: relative performance gain over baseline



# Zero-Shot and Hybrid Retrieval

## Experiment and results

Hybrid basic retrieval: relative performance gain over baseline



# Zero-Shot and Hybrid Retrieval

## Experiment and results

Hybrid contextualized retrieval: dual-parameter tuning

ColBERTv2 + Hiemstra LM with dual parameters significantly outperformed the baseline, improving MAP@10 by up to +4.24% at ( $p < 0.05$ ).

ColBERTv2 + sparse using a single parameter improved MAP@10 by up to +4.03% when the parameter was assigned to the dense-retrieval score.

|                                                                                                                 | P@10                               | MAP@10                             | NDCG@10                            |
|-----------------------------------------------------------------------------------------------------------------|------------------------------------|------------------------------------|------------------------------------|
| <b>Baseline</b>                                                                                                 |                                    |                                    |                                    |
| DFR BM25                                                                                                        | 0.8390                             | 0.2804                             | 0.7356                             |
| <b>Contextualized hybrid retrieval (<math>\alpha = 1.1</math> and <math>\beta = 0.8</math>, top 2k results)</b> |                                    |                                    |                                    |
| DPR + BM25                                                                                                      | 0.6864 (-18.19%)                   | 0.2079 (-25.86%)                   | 0.6161 (-16.25%)                   |
| DPR + DFR BM25                                                                                                  | 0.6864 (-18.19%)                   | 0.2083 (-25.71%)                   | 0.6169 (-16.14%)                   |
| DPR + Hiemstra LM                                                                                               | 0.6847 (-18.39%)                   | 0.2061 (-26.50%)                   | 0.6102 (-17.05%)                   |
| mDPR + BM25                                                                                                     | 0.7068 (-15.76%) <sup>†</sup>      | 0.2226 (-20.61%) <sup>†</sup>      | 0.6236 (-15.23%) <sup>†</sup>      |
| mDPR + DFR BM25                                                                                                 | 0.7102 (-15.35%) <sup>†</sup>      | 0.2234 (-20.33%) <sup>†</sup>      | 0.6262 (-14.87%) <sup>†</sup>      |
| mDPR + Hiemstra LM                                                                                              | 0.7288 (-13.13%) <sup>†</sup>      | 0.2305 (-17.80%) <sup>†</sup>      | 0.6334 (-13.89%) <sup>†</sup>      |
| Contriever + BM25                                                                                               | 0.3169 (-62.23%) <sup>†</sup>      | 0.0791 (-71.79%) <sup>†</sup>      | 0.2938 (-60.06%) <sup>†</sup>      |
| Contriever + DFR BM25                                                                                           | 0.3169 (-62.23%) <sup>†</sup>      | 0.0791 (-71.79%) <sup>†</sup>      | 0.2947 (-59.94%) <sup>†</sup>      |
| Contriever + Hiemstra LM                                                                                        | 0.3220 (-61.62%) <sup>†</sup>      | 0.0796 (-71.61%) <sup>†</sup>      | 0.2985 (-59.42%) <sup>†</sup>      |
| mContriever + BM25                                                                                              | 0.1814 (-78.38%)                   | 0.0408 (-85.45%)                   | 0.1936 (-73.68%)                   |
| mContriever + DFR BM25                                                                                          | 0.1814 (-78.38%)                   | 0.0409 (-85.41%)                   | 0.1938 (-73.65%)                   |
| mContriever + Hiemstra LM                                                                                       | 0.1847 (-77.99%)                   | 0.0421 (-84.99%)                   | 0.1967 (-73.26%)                   |
| ColBERTv2 + BM25                                                                                                | <b>0.8441 (+0.61%)<sup>†</sup></b> | <b>0.2862 (+2.07%)<sup>†</sup></b> | <b>0.7453 (+1.32%)<sup>†</sup></b> |
| ColBERTv2 + DFR BM25                                                                                            | <b>0.8458 (+0.81%)<sup>†</sup></b> | <b>0.2858 (+1.93%)<sup>†</sup></b> | <b>0.7480 (+1.69%)<sup>†</sup></b> |
| ColBERTv2 + Hiemstra LM                                                                                         | <b>0.8559 (+2.01%)<sup>†</sup></b> | <b>0.2923 (+4.24%)<sup>†</sup></b> | <b>0.7536 (+2.45%)<sup>†</sup></b> |
| ColBERT-X + BM25                                                                                                | 0.8305 (-1.01%)                    | 0.2765 (-1.39%)                    | 0.7335 (-0.29%)                    |
| ColBERT-X + DFR BM25                                                                                            | 0.8322 (-0.81%)                    | 0.2779 (-0.89%)                    | 0.7308 (-0.65%)                    |
| ColBERT-X + Hiemstra LM                                                                                         | 0.8356 (-0.41%)                    | <b>0.2807 (+0.11%)</b>             | 0.7303 (-0.72%)                    |
| <b>Contextualized hybrid retrieval (single parameter, <math>\alpha = 1.1</math>, top 2k results)</b>            |                                    |                                    |                                    |
| ColBERTv2 + BM25                                                                                                | <b>0.8508 (+1.41%)<sup>†</sup></b> | <b>0.2888 (+3.00%)<sup>†</sup></b> | <b>0.7493 (+1.86%)<sup>†</sup></b> |
| ColBERTv2 + DFR BM25                                                                                            | 0.8373 (-0.20%) <sup>†</sup>       | <b>0.2831 (+0.96%)<sup>†</sup></b> | <b>0.7438 (+1.11%)<sup>†</sup></b> |
| ColBERTv2 + Hiemstra LM                                                                                         | <b>0.8475 (+1.01%)<sup>†</sup></b> | <b>0.2891 (+3.10%)<sup>†</sup></b> | <b>0.7504 (+2.01%)<sup>†</sup></b> |
| ColBERT-X + BM25                                                                                                | 0.8322 (-0.81%)                    | 0.2771 (-1.18%)                    | 0.7334 (-0.30%)                    |
| ColBERT-X + DFR BM25                                                                                            | 0.8339 (-0.61%)                    | 0.2771 (-1.18%)                    | 0.7325 (-0.42%)                    |
| ColBERT-X + Hiemstra LM                                                                                         | 0.8356 (-0.41%)                    | 0.2790 (-0.50%)                    | 0.7304 (-0.71%)                    |
| <b>Contextualized hybrid retrieval (single parameter, <math>\beta = 0.8</math>, top 2k results)</b>             |                                    |                                    |                                    |
| ColBERTv2 + BM25                                                                                                | <b>0.8458 (+0.81%)<sup>†</sup></b> | <b>0.2866 (+2.21%)<sup>†</sup></b> | <b>0.7467 (+1.51%)<sup>†</sup></b> |
| ColBERTv2 + DFR BM25                                                                                            | <b>0.8441 (+0.61%)<sup>†</sup></b> | <b>0.2856 (+1.85%)<sup>†</sup></b> | <b>0.7476 (+1.63%)<sup>†</sup></b> |
| ColBERTv2 + Hiemstra LM                                                                                         | <b>0.8542 (+1.81%)<sup>†</sup></b> | <b>0.2917 (+4.03%)<sup>†</sup></b> | <b>0.7530 (+2.37%)<sup>†</sup></b> |
| ColBERT-X + BM25                                                                                                | 0.8322 (-0.81%)                    | 0.2775 (-1.03%)                    | 0.7344 (-0.16%)                    |
| ColBERT-X + DFR BM25                                                                                            | 0.8339 (-0.61%)                    | 0.2774 (-1.07%)                    | 0.7327 (-0.39%)                    |
| ColBERT-X + Hiemstra LM                                                                                         | 0.8373 (-0.20%)                    | 0.2804 (+0.00%)                    | 0.7305 (-0.69%)                    |
| <b>Reciprocal Rank Fusion</b>                                                                                   |                                    |                                    |                                    |
| ColBERTv2 + BM25                                                                                                | 0.7949 (-5.26%) <sup>†</sup>       | 0.2641 (-5.81%) <sup>†</sup>       | 0.7164 (-2.61%) <sup>†</sup>       |
| ColBERTv2 + DFR BM25                                                                                            | 0.7966 (-5.05%) <sup>†</sup>       | 0.2642 (-5.78%) <sup>†</sup>       | 0.7166 (-2.58%) <sup>†</sup>       |
| ColBERTv2 + Hiemstra LM                                                                                         | 0.8220 (-2.03%) <sup>†</sup>       | 0.2738 (-2.35%) <sup>†</sup>       | 0.7316 (-0.54%) <sup>†</sup>       |
| ColBERT-X + BM25                                                                                                | 0.7898 (-5.86%)                    | 0.2575 (-8.17%)                    | 0.6965 (-5.32%)                    |
| ColBERT-X + DFR BM25                                                                                            | 0.7898 (-5.86%)                    | 0.2577 (-8.10%)                    | 0.6969 (-5.26%)                    |
| ColBERT-X + Hiemstra LM                                                                                         | 0.8034 (-4.24%)                    | 0.2604 (-7.13%)                    | 0.6985 (-5.04%)                    |

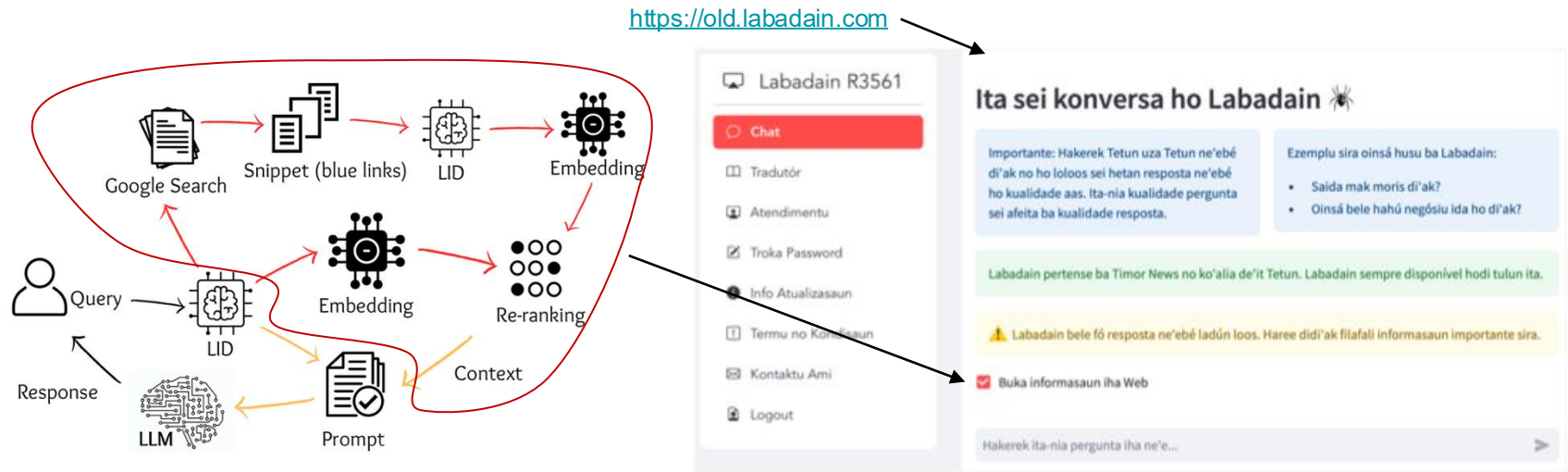
Paper: <https://dl.acm.org/doi/abs/10.1145/3731120.3744593>

Dataset: <https://doi.org/10.25747/RFZX-M945>

# Conversational AI

# Labadain-R361

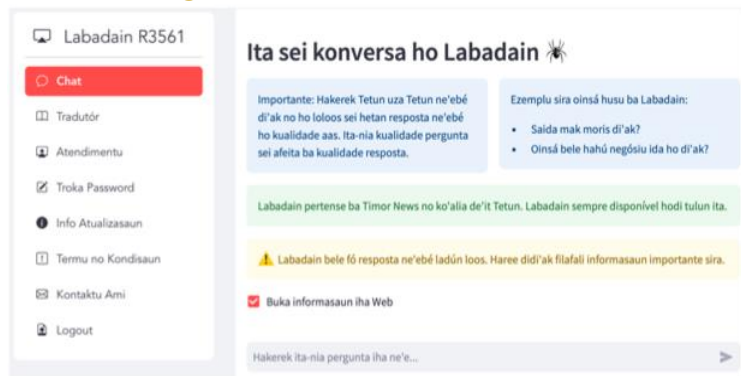
## Architecture and Prototype



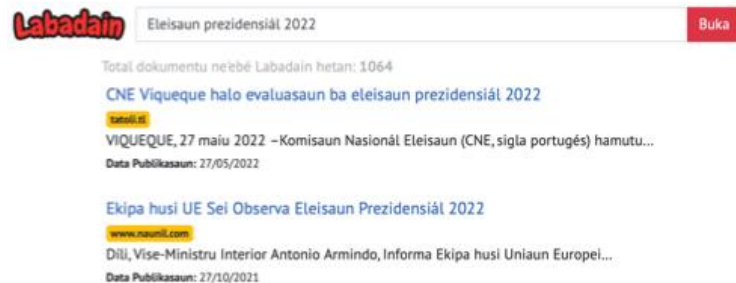
# User Search Behavior Analysis

## Data sources

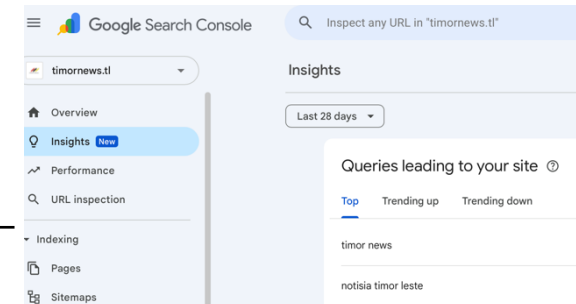
Labadain **Chat** ← [www.labadain.com](http://www.labadain.com)



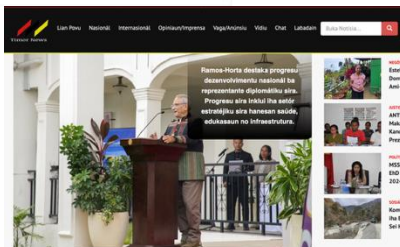
Labadain **Search** ← [www.labadain.tl](http://www.labadain.tl)



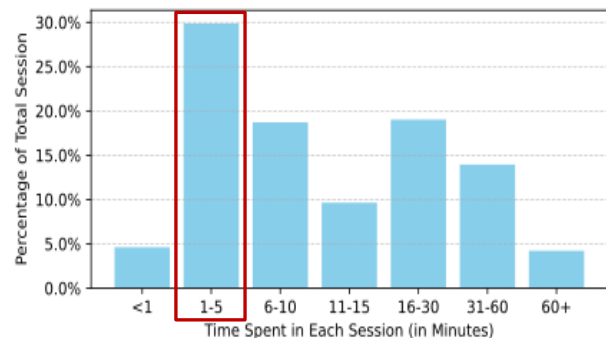
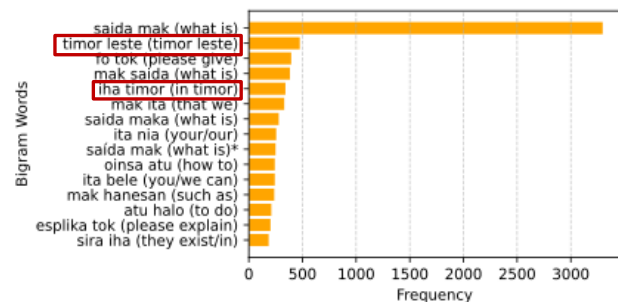
Google Search Console from **Timor News**



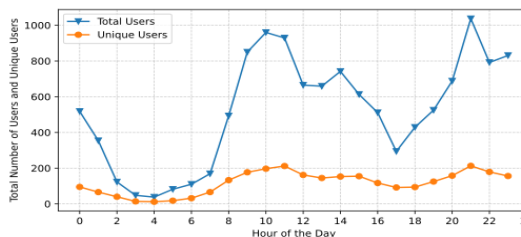
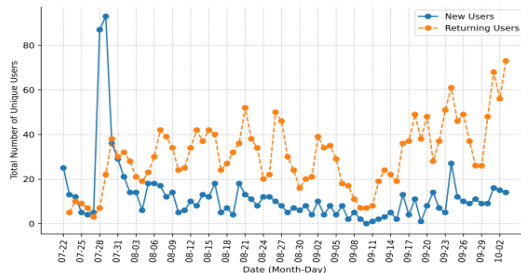
[www.timornews.tl](http://www.timornews.tl)



# User Search Behavior Analysis



Paper: <https://dl.acm.org/doi/abs/10.1145/3731120.3744596>  
 Dataset: <https://doi.org/10.25747/X2DK-5Y06>



## Insights into LLM-Based Conversational Search: A Study of Tetun-Speaking Users' Search Behavior

Gabriel de Jesus  
INESC TEC / University of Porto  
Porto, Portugal  
gabriel.jesus@inesctec.pt

Sérgio Nunes  
INESC TEC / University of Porto  
Porto, Portugal  
sergio.nunes@fe.up.pt

### Abstract

Advancements in large language model (LLM)-based conversational assistants have transformed search experiences into more natural and context-aware dialogues that resemble human conversation. However, limited access to interaction log data hinders a deeper understanding of their real-world usage. To address this gap, we analyzed 16,952 prompt logs from 904 unique users of Labadain Chat, an LLM-based conversational assistant designed for Tetun speakers, to uncover patterns in user search behavior, engagement, and intent. Our findings show that most users (29.87%) spent between one and five minutes per session, with an average of 43 unique daily users. The majority (93.97%) submitted multiple prompts per session, with an average session duration of 16.9 minutes. Most users (95.22%) were based in Timor-Leste, with education and science (28.75%)

### 1 Introduction

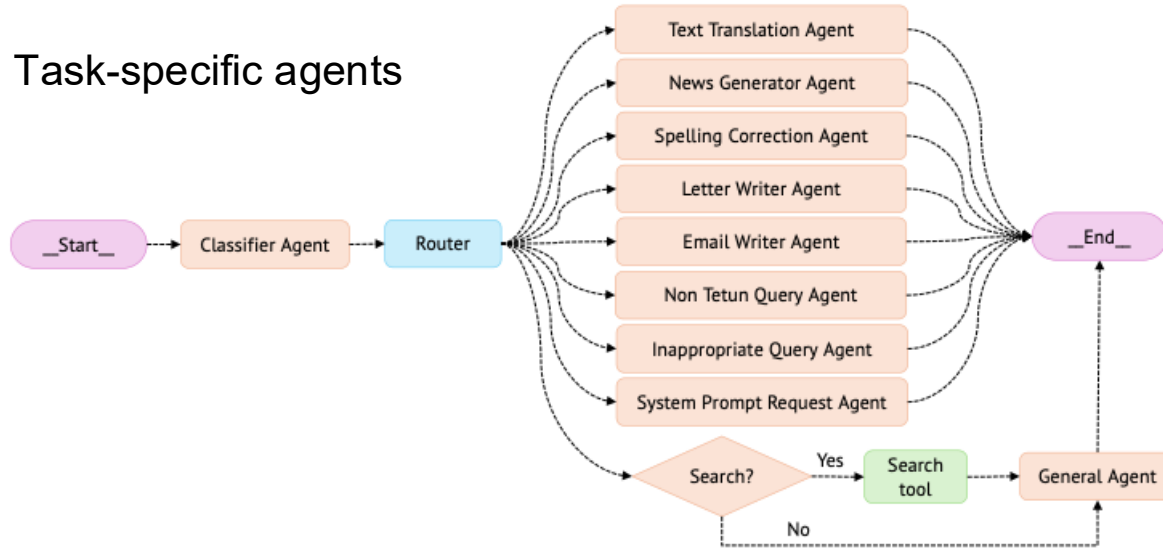
Conversational search refers to user-friendly dialogues between humans and machines, whether spoken or written, aimed at satisfying information needs [3, 4, 21, 29]. With the advent of LLM-based conversational assistants, such as OpenAI's ChatGPT,<sup>1</sup> Google's Gemini,<sup>2</sup> Anthropic's Claude,<sup>3</sup> and similar AI tools, such dialogues have become more natural and context-aware, closely resembling human conversation. These systems' ability to process unstructured language, handle complex queries, and support multi-turn interactions has made search more seamless and intuitive.

Furthermore, these developments demand a deeper understanding of user behavior, providing deeper insights into search behavior, engagement, and intent. These insights are crucial for refining LLM-based conversational systems to provide more personalized

| Topic                   | L. Chat       | L. Search    | Timor News    |
|-------------------------|---------------|--------------|---------------|
| Education/Science       | <b>28.75%</b> | 16.75%       | 11.25%        |
| Health                  | <b>28.00%</b> | 16.50%       | 12.25%        |
| Social/Culture          | <b>13.75%</b> | 11.75%       | 3.25%         |
| Research/Academia       | <b>6.75%</b>  | 4.00%        | 3.75%         |
| Politic/Government      | 2.50%         | <b>7.50%</b> | 6.00%         |
| Job vacancy             | 0.00%         | 3.50%        | <b>7.25%</b>  |
| Economy/Finance         | 1.00%         | <b>6.75%</b> | 2.75%         |
| Agriculture/Environment | <b>6.50%</b>  | 2.00%        | 1.75%         |
| Law/Justice             | 0.75%         | 3.75%        | 1.75%         |
| Language/Translation    | 2.00%         | 4.25%        | 1.50%         |
| Computer science        | 3.75%         | 3.00%        | 1.25%         |
| Personal                | 0.50%         | 2.50%        | 3.25%         |
| Porn                    | 0.00%         | 0.25%        | 2.00%         |
| Sport                   | 0.75%         | 1.00%        | 0.50%         |
| Holiday/Travel          | 0.25%         | 0.50%        | 0.75%         |
| Entertainment           | 0.00%         | 0.50%        | 0.75%         |
| Other                   | 4.75%         | 15.75%       | <b>40.00%</b> |

# Labadain LIX-R361

## Task-specific agents



<https://www.labadain.com>

Paper: <https://doi.org/10.1145/3805712.3808372>

# Labadain LIX-R361

## Prompt Engineering

- 1 Agents are instructed to act as Tetun linguists and all responses must be in Tetun, except for translation tasks.
- 2 Each agent receives explicit instructions, examples of Tetun official, and a curated lexical list.

## Model Assignment

- 1 **Gemini 2.5 Pro**: classifier, translation, news, spelling, email, and letter agents.
- 2 **Gemini 2.5 Flash**: search decision and general agents.
- 3 **Claude 3 Haiku**: exception-handling agents.

### General-agent memory

Only the most recent turn is retained (k=1) to control operational cost.

#### Prompt 4.1: Details of the General Agent Prompt.

You are Labadain, a Tetun linguist expert and helpful assistant. Your task is to assist the user with their request in Tetun, providing clear and accurate information. Assume the question is in Tetun and respond using accurate Tetun grammar.

The output must follow these examples of accurate official Tetun writing:  
{tetun\_official\_example}

To improve Tetun accuracy, follow the rules for converting Portuguese-derived characters into Tetun below:  
{pt\_loanwords\_tetun\_rules}

Incorporate the provided vocabulary pairs to improve your Tetun orthography:  
{vocabulary\_enrich\_list}

## Linguistic Resources

- 1 **Official Tetun examples**: Guide standard spelling, diacritics, and hyphenation.
- 2 **Portuguese-to-Tetun rules**: INL transformation rules support the correct use of Portuguese loanwords in Tetun.
- 3 **Curated lexical recourses**: Provides words that LLMs often fail to infer through generalization.

### Examples

- 1 Tetun la'ós de'it lian, ne'e ami-nia identidade (Tetun is more than just a language, it's our identity)
- 2 nh → ñ · lh → ll · ch → x · ão → aun
- 3 rice / arroz / etu · milk / leite / susubeen

# Labadain LIX-R361

## USER STUDY DESIGN

8,000+  
Query-log entries

60  
Curated queries

3  
Native Tetun-speaking annotators

### Query Categories

20 fact-checking

20 multi-turn

20 ambiguous

- 1 Responses from Labadain Chat (LBC), Labadain Old (LBO), and ChatGPT (CGPT) were pre-compiled and anonymized.
- 2 Two annotators reviewed each query.
- 3 Measures: binary task success and 1–5 user satisfaction.

## EVALUATION AND RESULTS

Labadain Chat outperformed both Labadain Old and ChatGPT across all query types.

### Overall task success



Cohen's  $k=0.67$

### Overall user satisfaction



Cohen's weighted  $k=0.75$

### Performance by query type

| Query Type   | Task Success Rate |     |      | User Satisfaction |      |      |
|--------------|-------------------|-----|------|-------------------|------|------|
|              | LBC               | LBO | CGPT | LBC               | LBO  | CGPT |
| Fact-seeking | 95%               | 73% | 48%  | 4.60              | 3.35 | 2.60 |
| Multi-turn   | 88%               | 65% | 58%  | 4.05              | 2.75 | 2.80 |
| Ambiguous    | 90%               | 58% | 53%  | 4.25              | 3.10 | 3.10 |

# **Conclusion and Future Work**

# Conclusion

**Labadain** is a set of language technology resources and applications created for the **Tetun language**



## Datasets

- ✓ Labadain-30k+
- ✓ Labadain-Avaliadór
- ✓ Labadain-Stopwords
- ✓ LabadainLog-17k+
- ✓ Labadain-ZSRuns
- ✓ LabadainLog-60



## Software

- ✓ Labadain Crawler



## Algorithms and tools

- ✓ Labadain-Stemmer
- ✓ Tetun Tokenizer
- ✓ Tetun LID



## Prototypes

- ✓ Labadain Search
- ✓ Labadain Chat

# Ongoing Research

## Labadain-HealthAI

*A conversational AI agent for healthcare in Tetun*



**Duration**  
12 months



**Funded by**  
TAF Timor-Leste



To develop a safe and culturally appropriate general-purpose conversational AI system that enable access to healthcare information in the Tetun language.



### Core resources:

- ✓ Medical terminology in Tetun
- ✓ Tetun healthcare text corpus

### Expected **outcomes**

- ✓ A functional conversational AI
- ✓ Healthcare datasets in Tetun
- ✓ Improve access to health information in Tetun

# Future Directions

- 01 **Advance research** in Tetun IR and NLP (resources, tools, and methods).
- 02 **Build partnerships** with universities and local communities.
- 03 Encourage **open collaboration** and **data sharing**.



**Thank** You

# **Labadain:** Resources and Applications for Tetun Language Technology

**Gabriel de Jesus**

Affiliated Researcher at INESC TEC

The University of Melbourne  
24 July 2026  
Melbourne, VIC, Australia

