Labadain: The Foundation of Tetun Language Technology

Gabriel de Jesus, PhD
Affiliated Researcher with **INESC TEC**
Founder and Director of **Timor News (LIX)**
Creator of **Labadain**

FEST-UNTL
November 20, 2025
FEST Auditorium Room
Hera, Dili, Timor-Leste

Outline

Introduction

- What is Language Technology?
- Who is the Language Tech. for?
- Why Tetun?
- The Role of Tetun

Labadain in Language Technology

- Overview
- Labadain Main Components
- Labadain Component Details

Research, Innovation, and Impact

- Research and Innovation
- Social Impact
- Academia Impact

Challenges, Opportunities, and Future Directions

- Challenges and Open Issues
- Opportunities
- Future Directions

Introduction

What is Language Technology?

Human language technology refers to systems that understand, process, search, and interact with human language.

Which one?

Inclusion

Research Areas:

- Natural language processing
- Computational Linguistics
- Speech Technologies
- Information Retrieval and Search
- ...

Who is the Language Technology for?

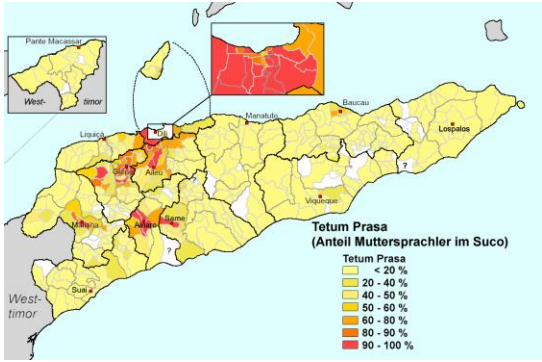
A truly inclusive digital transformation must speak the language of its people.

TETUN

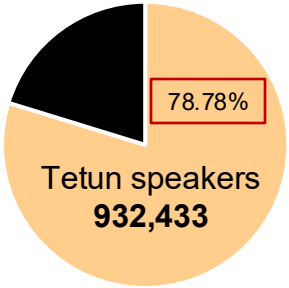
Why Tetun?

01 The most spoken language in Timor-Leste

02 One of the country's official languages



Source: Wikimedia



1.18M population

2015 census report

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

The Role of Tetun

Connects people, culture, and technology

Empowers cultural identity and local innovation

- National language of inclusion
- Bridge government and citizens
- Voice of cultural identity
- Foundation for AI and innovation
- Key to digital sovereignty

Labadain in Language Technology

Overview

Google Gemini

What is Labadain in Language Technology?

AI Mode | All | Images | Videos | Short videos | News | Forums | More | Tools

AI Overview

In language technology, **Labadain** is the name given to a set of resources developed for the **Tetun language** (a low-resource language spoken in Timor-Leste), primarily for use in natural language processing (NLP) and information retrieval (IR) research.

Key Labadain resources include:

- **Labadain-30k+**: A monolingual dataset comprising over 33,000 Tetun documents (mostly news articles and Wikipedia documents) which underwent manual auditing by native speakers. It is a key corpus for developing language models for Tetun.
- **Labadain-Stopwords**: A list of 160 common Tetun "stopwords" (function words like articles and prepositions) developed using frequency- and network-based approaches.
- **Labadain-Stemmer**: A stemming algorithm for Tetun, designed to reduce words to their base or root forms for consistent processing in NLP tasks.
- **Labadain-Avaliadór**: A Tetun test collection for ad-hoc text retrieval tasks, including topics, documents, and human-assigned relevance judgments.
- **Labadain Crawler**: A data collection pipeline used to acquire the text data from the web.

These resources were created to establish a foundation for Tetun NLP and IR, a field that previously had limited existing solutions.

Perplexity AI

What is Labadain in Language Technology?

Answer | Images

github: gabriel-de-jesus/labadain-crawler - GitHub
ACL Anthology: Labadain-30k+: A Monolingual Tetun...
aclanthology: [PDF] Labadain-30k+: A Monolingual Tetun...
cnr: [PDF] Labadain-30k+: A Monolingual Tetun...

Assistant steps >

Labadain refers to a set of language technology resources and tools created specifically for Tetun, a low-resource language spoken in Timor-Leste. It encompasses datasets, crawling pipelines, search engines, language processing components, and evaluation resources, supporting research and application development in natural language processing (NLP) for Tetun. aclanthology +2

Key Components of Labadain

- **Labadain Dataset:** The Labadain-30k+ dataset consists of over 33,000 audited Tetun documents collected through specialized web crawling and augmented with Wikipedia articles. It is notable for its comprehensive manual audit by native speakers and serves as a foundational resource for Tetun NLP and information retrieval research. aclanthology +1
- **Labadain Crawler:** The crawler automates and optimizes the collection of Tetun language data from the web, targeting the construction of extensive textual corpora for low-resource language applications. github
- **Search Engine (Labadain.tl):** Labadain also refers to a dedicated search engine for the Tetun language, providing web search capabilities tailored to Tetun-speaking users in Timor-Leste. linkedin
- **Language Processing Tools:** Labadain resources include the Labadain-Stopwords list, the Labadain-Stemmer (with variants designed for Tetun and Portuguese loanwords), and the Labadain-Avaliadór test collection for document relevance judgment and retrieval evaluation. arxiv +1
- **Conversational AI:** In 2024, Labadain Chat4 was launched as a large language model (LLM)-based conversational assistant for Tetun, further supporting digital transformation and

Labadain **Main Components**

Labadain is a set of language-technology resources and tools created for **Tetun.**

Datasets

Software

Algorithms and tools

Prototypes

Labadain Component Details

Datasets
- Labadain-30k+
- Labadain-Avaliadór
- Labadain-Stopwords
- LabadainLog-17k+
- Labadain-ZSRuns

Software ← Labadain Crawler

Algorithms and tools
- Labadain-Stemmer
- Tetun Tokenizer
- Tetun LID

Prototypes
- Labadain Search
- Labadain Chat

Research, Innovation, and Impact

“**Labadain** isn’t just a product – it represents the effort to advance research and innovation in Tetun language Technology”

Research and Innovation

Tetun tokenizer:

```
pip install tetun-tokenizer
```

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

Tetun LID model:

```
pip install tetun-lid
```

```
from tetunlid import lid
```

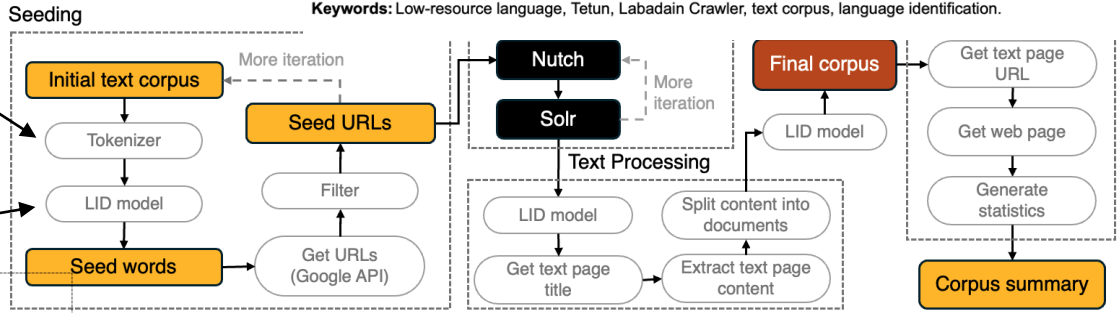
Data Collection Pipeline for Low-Resource Languages: A Case Study on Constructing a Tetun Text Corpus

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

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

Research and Innovation

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

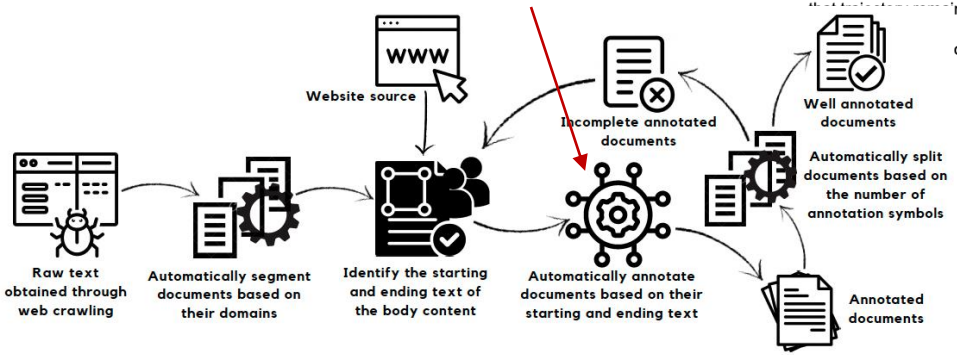
Algorithm 1 Content Annotation Algorithm.

```
Require: start_text, end_text, documents, output_file
for all document in documents do
    get title and url from document
    write title and url to output_file          ▷ The "annotated documents" file in Figure 4.3.
    get body_content from document
    annotation_t_counter ← 0    ▷ To control the occurrence of < t > to a maximum of two.
    for all text_line in body_content do
        get text_line_lower by lowercasing text_line and removing extra spaces
        if text_line_lower starts with start_text and annotation_t_counter equals 0 then
            write annotation string < t >, a newline, text_line, and a newline to output_file
            Increment annotation_t_counter by 1
        else if text_line_lower ends with end_text and annotation_t_counter equals 1 then
            write text_line, a newline, annotation string < t >, and a newline to output_file
            Increment annotation_t_counter by 1
        else
            write text_line and a newline to output_file
        end if
    end for
    write an additional newline to output_file ▷ To separate each document by two newlines.
end for
```

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 language became evident, reflecting an increase of 5.09 percentage points.

ource language, Tetun, Text dataset, Corpus content analysis.



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

Paper: https://aclanthology.org/2024.sigul-1.22/
Dataset: https://doi.org/10.25747/ydwr-n696

Research and Innovation

Labadain-Avaliadór

A Tetun test collection for ad-hoc text retrieval

> A test collection is a standardized dataset used to evaluate and compare retrieval effectiveness of search systems.

Establishing a Foundation for Tetun Ad-hoc Text Retrieval: Stemming, Indexing, Retrieval, and Ranking

GABRIEL DE JESUS, Institute for Systems and Computer Engineering, Tech. and Science (INESC TEC), Portugal
SÉRGIO NUNES, INESC TEC and Faculty of Engineering, University of Porto (FEUP), Portugal

Searching for information on the internet and digital platforms requires effective retrieval solutions. However, such solutions are not yet available for Tetun, making it difficult to find relevant documents for search queries in this language. To address this gap, we investigate Tetun text retrieval with a focus on the ad-hoc retrieval task. The study begins with the development of essential language resources—including a list of stopwords, a stemmer, and a test collection—that serve as a foundation for Tetun text retrieval. Various strategies are evaluated using document titles and content. The results show that retrieving document titles, after removing hyphens and apostrophes but without applying stemming, improves performance compared to the baseline. Efficiency increases by 31.37%, while effectiveness achieves an average relative gains of +9.40% in MAP@10 and +30.35% in NDCG@10 with DFR BM25. Beyond the top-10 cutoff point, Hiemstra LM demonstrates strong performance across multiple retrieval strategies and evaluation metrics. The contributions of this work include the development of *Labadain-Stopwords* (a list of 160 Tetun stopwords), *Labadain-Stemmer* (a Tetun stemmer with three variants), and *Labadain-Avaliadór* (a Tetun test collection comprising 59 topics, 33,550 documents, and 5,900 *qrels*). These resources are publicly available to support future research in Tetun information retrieval.

CCS Concepts: • **Information Retrieval** → **Text Retrieval**; • **Low-Resource Languages** → *Tetun*.

Additional Key Words and Phrases: Stopwords, Stemming, Test Collection, Ad-hoc Retrieval

Query

Document

Relevance

Timor News search logs

Labadain-30k+ dataset

Tetun-speaking assessors

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

Research and Innovation

Labadain-Avaliadór

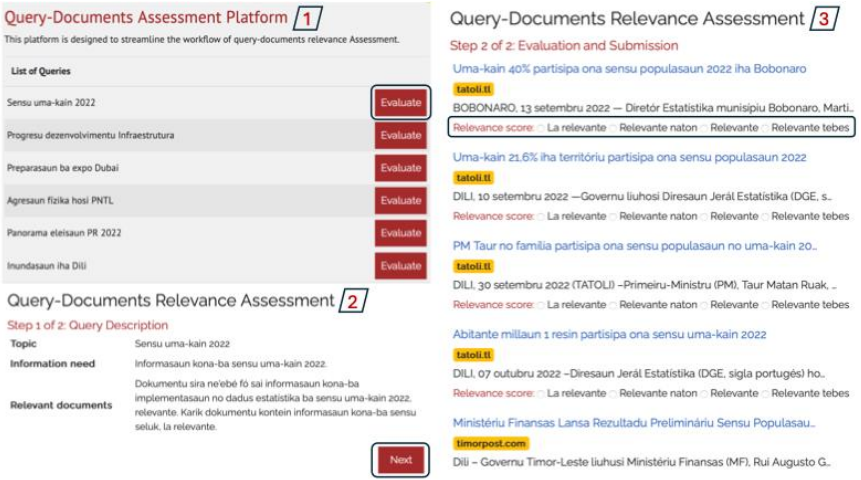
A Tetun test collection for ad-hoc text retrieval

Establishing a Foundation for Tetun Ad-hoc Text Retrieval: Stemming, Indexing, Retrieval, and Ranking

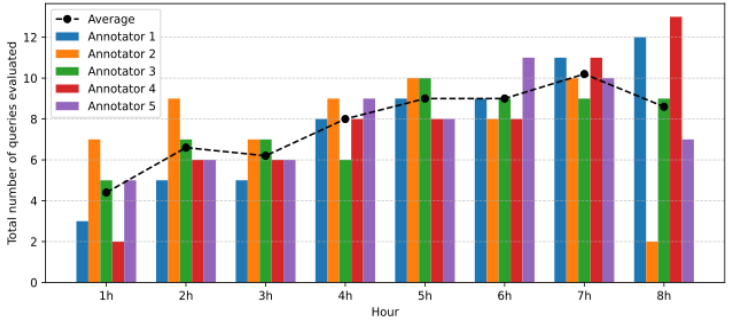
GABRIEL DE JESUS, Institute for Systems and Computer Engineering, Tech. and Science (INESC TEC), Portugal
SÉRGIO NUNES, INESC TEC and Faculty of Engineering, University of Porto (FEUP), Portugal

Searching for information on the internet and digital platforms requires effective retrieval solutions. However, such solutions are not yet available for Tetun, making it difficult to find relevant documents for search queries in this language. To address this gap, we investigate Tetun text retrieval with a focus on the ad-hoc retrieval task. The study begins with the development of essential language resources—including a list of stopwords, a stemmer, and a test collection—that serve as a foundation for Tetun text retrieval. Various

Web interface used by human assessors



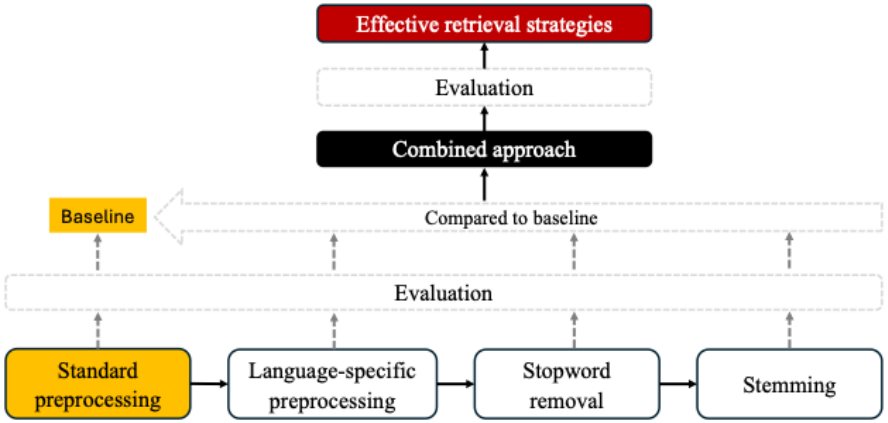
Total of queries evaluated by each annotator per hour



Description	Value
Total number of topics	59
Total number of *qrels*	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

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

Research and Innovation



Establishing a Foundation for Tetun Ad-hoc Text Retrieval: Stemming, Indexing, Retrieval, and Ranking

GABRIEL DE JESUS, Institute for Systems and Computer Engineering, Tech. and Science (INESC TEC), Portugal
SÉRGIO NUNES, INESC TEC and Faculty of Engineering, University of Porto (FEUP), Portugal

Searching for information on the internet and digital platforms requires effective retrieval solutions. However, such solutions are not yet available for Tetun, making it difficult to find relevant documents for search queries in this language. To address this gap, we investigate Tetun text retrieval with a focus on the ad-hoc retrieval task. The study begins with the development of essential language resources—including a list of stopwords, a stemmer, and a test collection—that serve as a foundation for Tetun text retrieval. Various strategies are evaluated using document titles and content. The results show that retrieving document titles, after removing hyphens and apostrophes but without applying stemming, improves performance compared to the baseline. Efficiency increases by 31.37%, while effectiveness achieves an average relative gains of +9.40% in MAP@10 and +30.35% in NDCG@10 with DFR BM25. Beyond the top-10 cutoff point, Hiemstra LM demonstrates strong performance across multiple retrieval strategies and evaluation metrics. The contributions of this work include the development of *Labadain-Stopwords* (a list of 160 Tetun stopwords), *Labadain-Stemmer* (a Tetun

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	0.6468	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	0.6914	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 apos-trophes and hyphens	BM25	0.8881	0.8373	0.7153	0.1553	0.2796	0.4304	0.7500	0.7347	0.7133	0.6648	0.8213
	DFR BM25	0.8881	0.8390	0.7169	0.1553	0.2804	0.4313	0.7512	0.7356	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 hy-	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

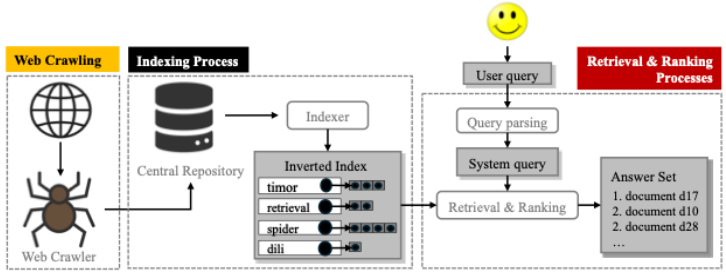
Paper: https://arxiv.org/abs/2412.11758

Research and Innovation

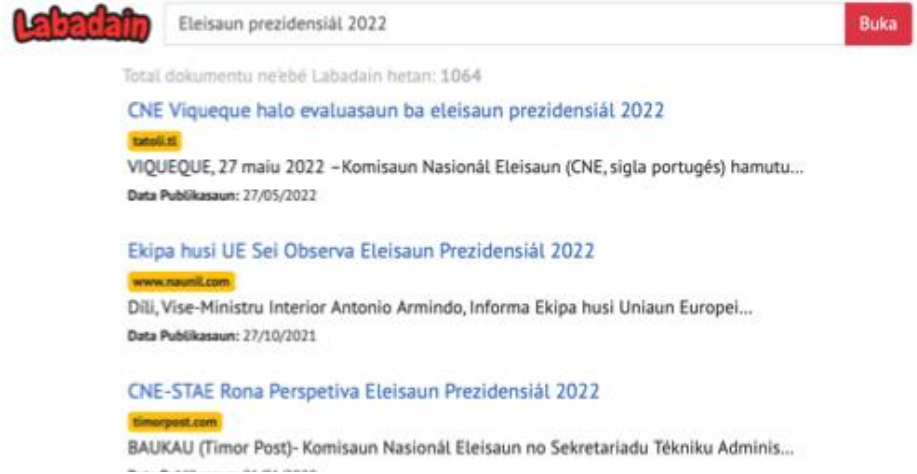
Labadain Search

www.labadain.tl

General Architecture



Labadain Crawler

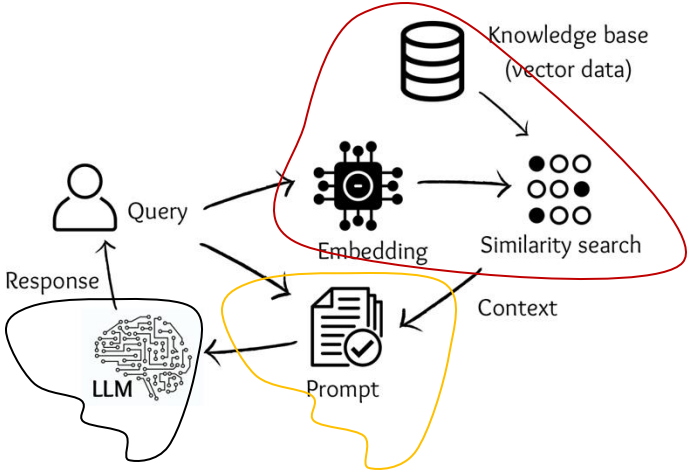


Research and Innovation

Labadain Chat (RAG Prototype)

RAG is an architecture that combines information retrieval (IR) with generative large language models (LLMs) to produce more accurate and up-to-date responses.

R - Retrieval	Find relevant documents
A - Augmented	Add it to the prompt context
G - Generation	Produce better responses

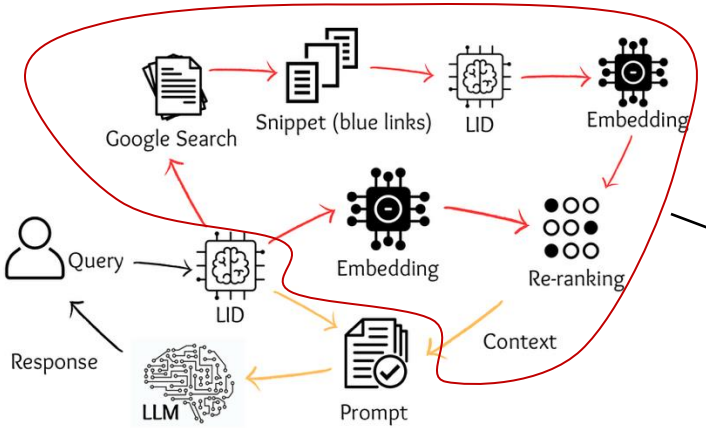


General Architecture of RAG

Research and Innovation

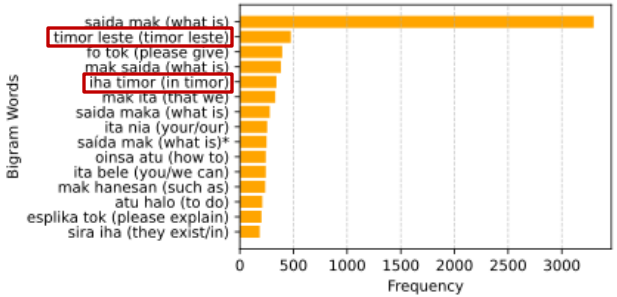
Labadain Chat (RAG Prototype)

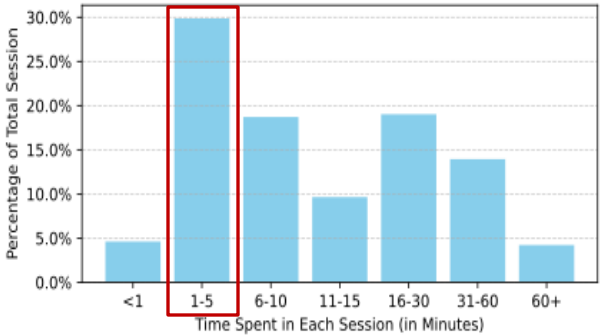
Architecture

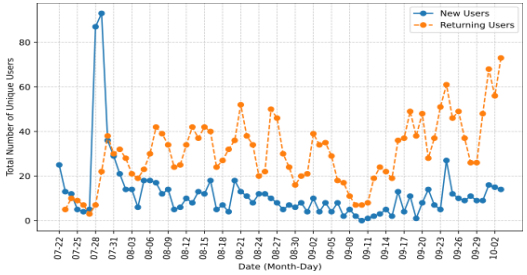


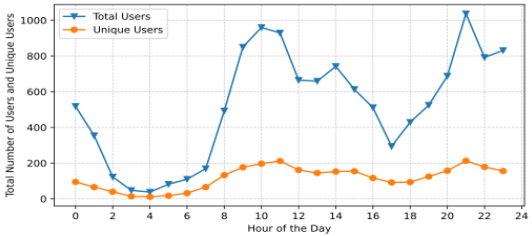
www.labadain.com

Social Impact









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,[1] Google's Gemini,[2] Anthropic's Claude,[3] 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	**28.75%**	16.75%	11.25%
Health	**28.00%**	16.50%	12.25%
Social/Culture	**13.75%**	11.75%	3.25%
Research/Academia	**6.75%**	4.00%	3.75%
Politic/Government	2.50%	**7.50%**	6.00%
Job vacancy	0.00%	3.50%	**7.25%**
Economy/Finance	1.00%	**6.75%**	2.75%
Agriculture/Environment	**6.50%**	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%	**40.00%**

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

Academia Impact

Tetun Datasets

Labadain-Stopwords: A Curated List of 160 Tetun Stopwords

Labadain-Stopwords is a curated list of 160 Tetun stopwords, compiled from the Labadain-30k+ dataset and validated by native speakers. It is well-suited for various Tetun...

TXT

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

Labadain-30k+ is a monolingual Tetun dataset containing 33,550 documents spanning from June 2001 to September 2023, excluding the years 2004 and 2005, for which no documents are...

TXT PYTHON

Labadain-Avaliadór : A Test Collection for Tetun Ad-hoc Text Retrieval

The Labadain-Avaliadór dataset is a test collection developed for the ad-hoc retrieval task. It comprises 59 topics, 33,550 documents, and 5,900 query-document relevance...

CSV TREC ZIP

LabadainLog-17k+: Search Logs from Tetun-Speaking Users Across Chat, Web,...

1. Overview LabadainLog-17k+ is a dataset of interaction logs in Tetun, collected from three different platforms: Labadain Chat (16,952 prompts): An LLM-powered conversational...

CSV TXT ZIP

Labadain-ZSRunS: Sparse and Zero-Shot Dense Retrieval Runs with...

1. Overview Labadain-ZSRunS is a dataset consisting of run files produced by classical sparse and zero-shot dense retrieval models, resulted from the experiments on Tetun ad-hoc...

ZIP

Tools, algorithm, and software

tetun-lid

Tetun Language Identification Model

tetun-tokenizer

Tetun tokenizer

labadain-stemmer Public

A rule-based stemming algorithm for Tetun

Python 1

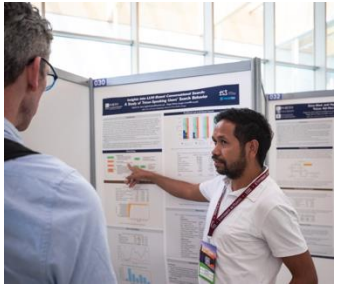
labadain-crawler Public

A data collection pipeline for low-resource languages

Python 1

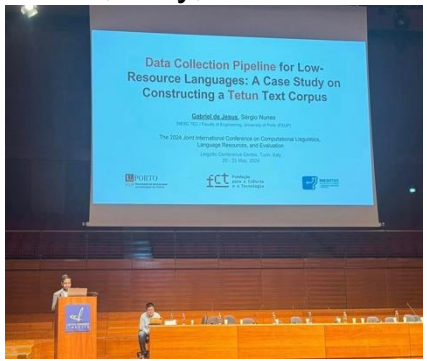
Academia **Impact**

Padova, Italy, 2025



Washington, USA, 2024

Turin, Italy, 2024



Gabriel de Jesus

FOLLOW

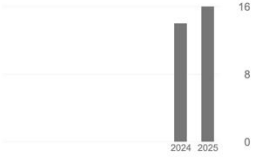
GET MY OWN PROFILE

INESC TEC - Institute for Systems and Computer Engineering, Technology and Science
Verified email at inesctec.pt
Information Retrieval Search Natural Language Processing Low-resource languages

TITLE	CITED BY	YEAR
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 Under …	7	2024
Exploring Large Language Models for Relevance Judgments in Tetun G de Jesus, S Nunes Proceedings of the First Workshop on Large Language Models for Evaluation in …	6	2024
Data collection pipeline for low-resource languages: A case study on constructing a tetun text corpus G de Jesus, SS Nunes Proceedings of the 2024 Joint International Conference on Computational …	6	2024
Text Information Retrieval in Tetun G de Jesus European Conference on Information Retrieval, 429-435	6	2023
Text Information Retrieval in Tetun: A Preliminary Study G de Jesus Proceedings of the 10th edition of the PhD Symposium on FDIA, July 20, 2022 …	3	2024
Establishing a Foundation for Tetun Ad-Hoc Text Retrieval: Stemming, Indexing, Retrieval, and Ranking G de Jesus, S Nunes arXiv preprint arXiv:2412.11758	1	2024
Network-based Approach for Stopwords Detection FDMA Ali, G de Jesus, HL Cardoso, S Nunes, R Sousa-Silva Proceedings of the 16th International Conference on Computational Processing …	1	2024
Cross-Lingual Information Retrieval in Tetun for Ad-Hoc Search A Araújo, G de Jesus, S Nunes EPIA Conference on Artificial Intelligence, 262-275		2025
Insights into LLM-Based Conversational Search: A Study of Tetun-Speaking Users' Search Behavior G Jesus, S Nunes Proceedings of the 2025 International ACM SIGIR Conference on Innovative …		2025
Zero-Shot and Hybrid Strategies for Tetun Ad-Hoc Text Retrieval G de Jesus, SAK Singh, S Nunes, A Yates Proceedings of the 2025 International ACM SIGIR Conference on Innovative …		2025

Cited by

	All	Since 2020
Citations	30	30
h-index	4	4
i10-index	0	0



Public access VIEW ALL

1 article 6 articles
not available available

Based on funding mandates

Co-authors

Sérgio Nunes
INESC TEC and Faculty of Engi…

Andrew Yates
Johns Hopkins University, Huma…

Challenges, Opportunities, and Future Directions

Challenges and Open Issues

Tetun text corpora remain limited.

Limited AI experts in Timor-Leste.

Few institutions focus on language technology research and development.

Insufficient funding for Tetun language innovation.

Opportunities

01 Advance research in Tetun NLP, IR, Generative AI, and related fields.

02 Build partnerships with universities, industry, and local communities.

03 Contribute to open Tetun datasets, resources, and AI tools.

Future Directions

Expand Tetun linguistic resources.

Expand techniques and tools for Tetun language processing.

Establish a dedicated AI R&D laboratory.

Encourage open collaboration and data sharing.

“**Labadain** is the foundation for inclusive digital future”

Thank You

Labadain: The Foundation of Tetun Language Technology

Gabriel de Jesus, PhD
Affiliated Researcher with **INESC TEC**
Founder and Director of **Timor News (LIX)**
Creator of **Labadain**

FEST-UNTL
November 20, 2025
FEST Auditorium Room
Hera, Dili, Timor-Leste