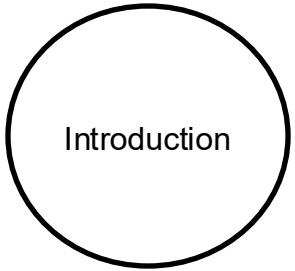
AI for Tetun: Building Timor-Leste's Inclusive Digital Future

Empowering Language, Culture, and Innovation

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

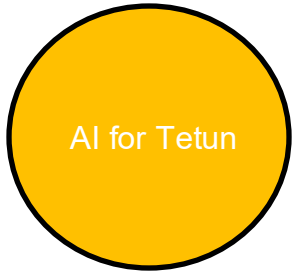
TLNOG2
November 21, 2025
F-FDTL Auditorium Room
Dili, Timor-Leste

Outline



- National Digital Plan 2022–2032
- Who is the Digital Future for?
- Why Tetun?
- The Role of Tetun
- Linguistic Context of Timor-Leste
- Multilingual Impact on Search
- Overview of the Talk

- Tetun Dataset Construction
- Labadain-30k+ Dataset
- Labadain Stopwords
- Labadain Stemmer
- Labadain Avaliadór
- Classical Retrieval Baselines
- Labadain Search
- Zero-Shot and Hybrid Retrieval



- RAG Overview
- Labadain Chat
- Learning from Labadain Users
- Why AI Matter for Tetun
- Challenges and Open Issues

- Foundation for Tetun IR, NLP, and AI
- Future Directions

Introduction

National Digital Plan 2022-2032

Vision

01 Achieve a digital government accessible from national to the *suco* (village) level.

02 Create an inclusive digital economy that connects Timor-Leste to global market and opportunities.

Strategic Digital Pillars

- Governance
- Economy
- Education
- Agriculture
- Health

Ensure efficiency and effectiveness

Core components of strategic plan and ICT development framework.

National Digital Plan 2022-2032

Current Implementation Status

Feasibility study and digital-readiness assessments across municipalities

ICT infrastructure expansion

Development of foundational platforms and e-government systems

Digital skills development

Promote awareness of e-government initiatives

Government of Timor-Leste. Timor Digital 2032: Municipal Digital Needs Assessment Report, 2025. URL: https://timordigital.gov.tl/en/documents
Government of Timor-Leste. Timor Digital 2032: Digital and E-Government Awareness Program Evaluation Report, 2025. URL: https://timordigital.gov.tl/en/documents
Government of Timor-Leste. Timor Digital 2032: Digital Readiness Baseline Study Report, 2023. URL: https://timordigital.gov.tl/en/documents

Who is the Digital Future for?

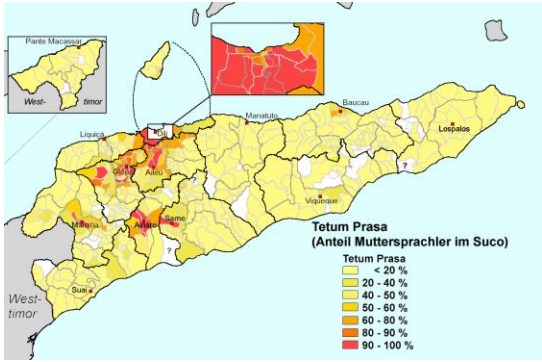
A truly inclusive digital transformation must speak the language of its people.

TETUN

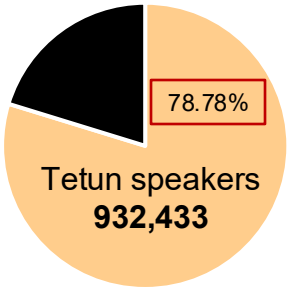
Why Tetun?

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

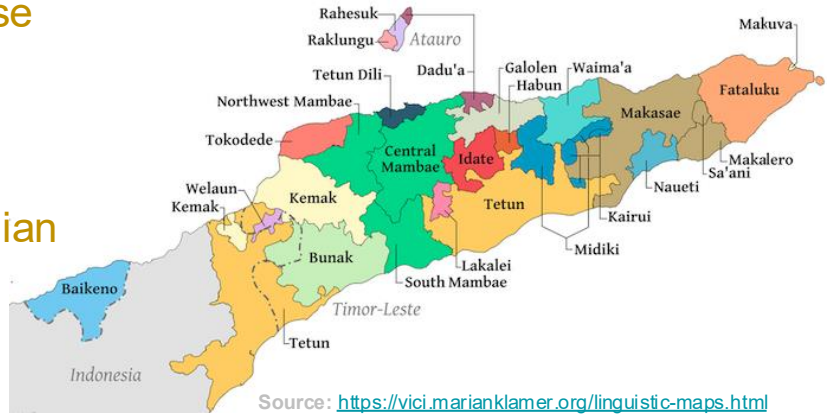
Empowers cultural identity
and local innovation

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

Linguistic Context of Timor-Leste

Multilingual

- Official languages → Tetun and Portuguese
- Working languages → English and Indonesian
- Local languages → 30+ dialects



Source: https://vici.marianklamer.org/linguistic-maps.html

Vasconcelos et al. (2011). Constituição anotada da República Democrática de Timor-Leste. URL http://hdl.handle.net/10400.22/4008

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/

Multilingual Impact on Search

Portuguese

01 Orsamentu **geral do** estadu **total** hamutuk hira — 13-10-2022

02 MACOP intrevista iha **bandara** nicolau lobato — 09-11-2022

Search date

Indonesian

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

This Talk

Presents:

01 Foundational resources for Tetun information retrieval (IR), natural language processing (NLP), and artificial intelligence (AI).

02 AI for language inclusion: Adapting Retrieval-Augmented Generation (RAG) to Tetun.

03 Prototypes showcasing search engine and conversational search for Tetun.

Text processing algorithms

A text corpus

A test collection

Baselines for text ad-hoc retrieval

Labadain

Search (www.labadain.tl)

Chat

Translate

Reception

www.labadain.com

Foundation and Application

Context

Early Stage of My PhD in 2021

Techniques and algorithms for Tetun IR and NLP did not exist.

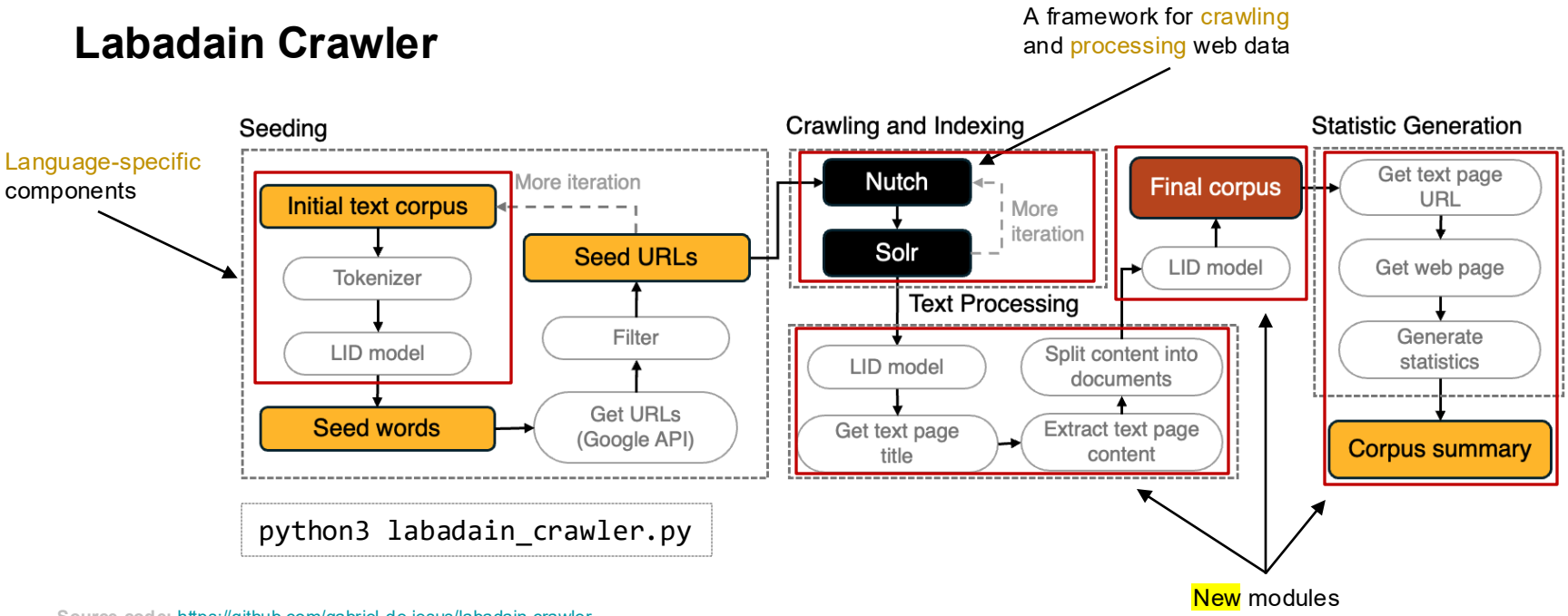
Tetun text corpora for IR and NLP tasks were not available.

→

Finding Tetun documents on the internet was challenging

Tetun Dataset Construction

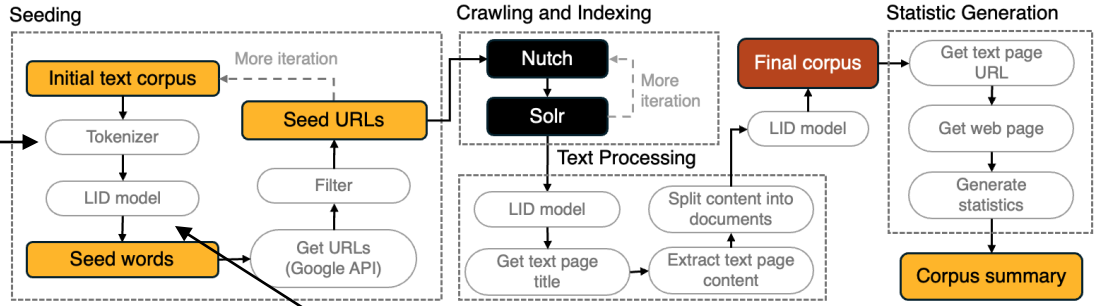
Labadain Crawler



Source code: https://github.com/gabriel-de-jesus/labadain-crawler
Paper: https://aclanthology.org/2024.lrec-main.390/

Tetun Dataset Construction

Tetun Tokenizer and LID Model



Tetun tokenizer:

```
pip install tetun-tokenizer
```

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

["Ha'u", "hela", "iha", "Dili"]

Tetun LID model:

```
pip install tetun-lid
```

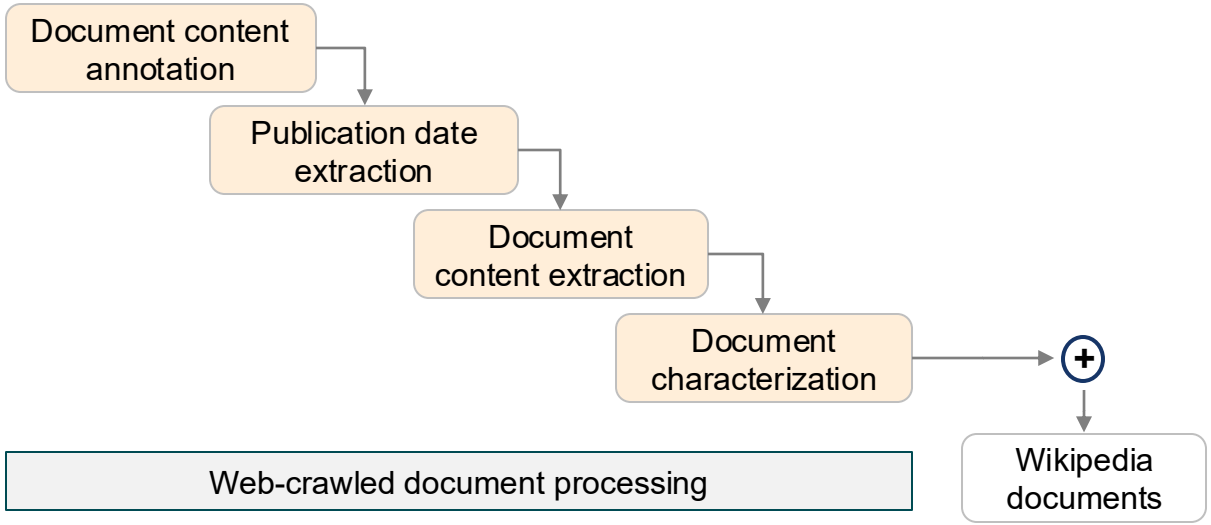
```
from tetunlid import lid
```

["Ha'u hela iha Dili"] → Tetun

Tetun tokenizer: https://pypi.org/project/tetun-tokenizer/
Tetun LID model: https://pypi.org/project/tetun-lid/
Paper: https://aclanthology.org/2024.lrec-main.390/

Tetun Dataset Construction

Dataset construction flow



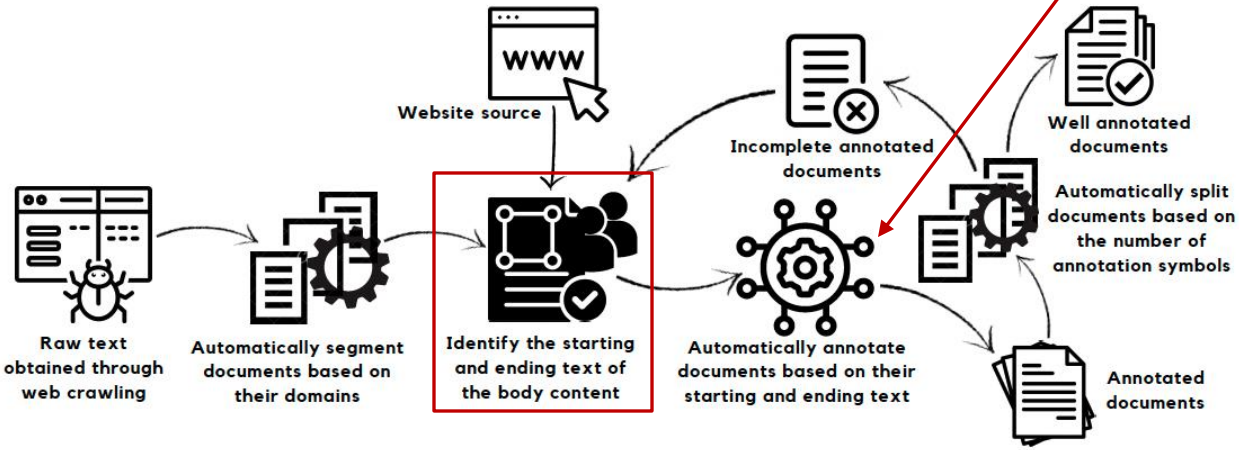
Tetun Dataset Construction

Document content annotation and processing

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



Tetun Dataset Construction

Labadain-30k+ Dataset

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 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 category

Category	#docs	Proportion
News articles	30,150	89.87%
Wikipedia documents	1,455	4.34%
Legal/gov. documents	1,223	3.65%
Technical documents	211	0.63%
Blogs and Forums	145	0.43%
Ads/announcements	124	0.37%
Research papers	83	0.25%
Personal pages	74	0.22%
Institutional information	53	0.16%
Correspondence letters	32	0.1%

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%

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

Labadain-30k+ Dataset

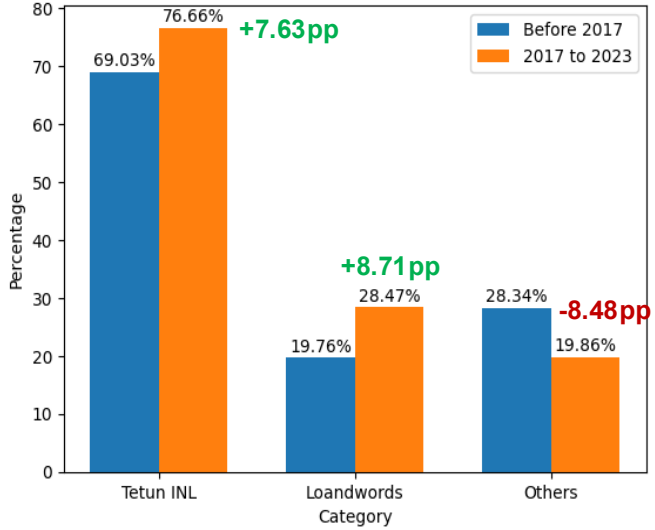
Content Analysis

Document quantity evolution on the web

Year	#docs	Proportion	Difference
2010	300	0.89%	↑0.72 pp+
2011	174	0.52%	↓0.37 pp
2012	190	0.57%	↑0.05 pp
2013	199	0.59%	↑0.02 pp
2014	252	0.75%	↓0.16 pp
2015	290	0.86%	↑0.11 pp
2016	451	1.34%	↑0.48 pp
2017	818	2.44%	↑1.10 pp
2018	1,164	3.47%	↑1.03 pp
2019	1,810	5.39%	↑1.92 pp
2020	**5,749**	**17.14%**	**↑11.75 pp**
2021	6,317	18.83%	↑1.69 pp
2022	8,500	25.34%	↑6.51 pp
2023	7,229	21.55%	↓3.79 pp

Frequent words on the dataset

The use of Tetun INL and PT loanwords in news articles

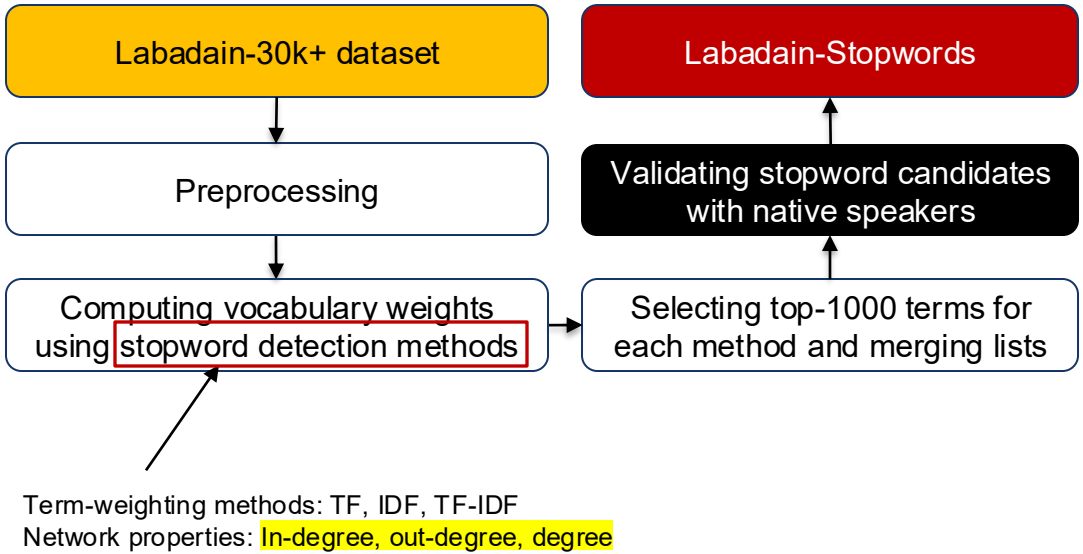


Paper: https://aclanthology.org/2024.sigul-1.22/

Labadain **Stopwords**

Constructing Labadain-Stopword List

Stopwords are function words (e.g., articles, prepositions, and conjunctions)



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

Labadain **Stemmer**

Stemming is a text-processing technique that reduces words to their base form. For example, [car, cars, car's, cars'] are all reduce to "car".

Tetun Linguistic Overview

Orthography

Tetun uses Latin script, consisting of 5 vowels and 21 consonants.

Portuguese loanwords

"Ha'u agradese tebes tanba halo investigasaun ida-ne'ebé krusiál."

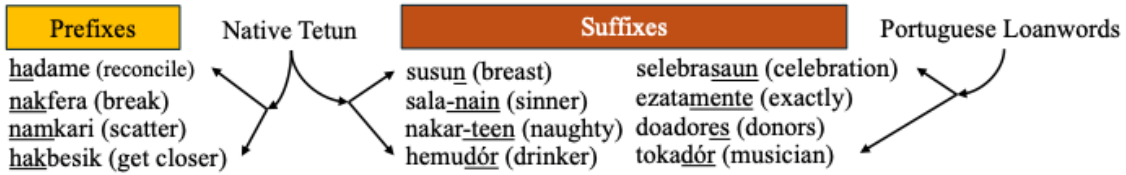
Verb: agradese
Noun: investigasaun
Adjective: krusiál

Up to 40% are Portuguese loanwords

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?

Affixes



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

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.

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

Labadain **Stemmer**

Tetun Linguistic Overview

Rules from INL for transforming Portuguese-derive words in Tetun

Rule (Portuguese → Tetun)	Effect on suffix	Example (Portuguese → Tetun)
ão → aun	as**aun**	comemoração → komemoras**aun** (celebration)
ss, c (before e, i), ç (before a, o, u) → s	s**aun**	discussão → diskus**aun** (discussion)
qu, c (before a, o, u) → k	**ik** + amente	automaticamente → automatikamente (automatically)
g (before e, i) → j	lo**ji**a	tecnolo**gi**a → teknolo**ji**a (technology)
s (between vowels) → z	o**z**a	poderosa → poderoza (powerful)
[illegible]	[illegible]	[illegible]
ánsia	ignorância → ignor**ánsia** (ignorance)	â → á
u	infermeir**o** → infermeir**u** (nurse)	o → u

Portuguese-derived suffixes

Suffix	Variable	Description
eza, ezas, iku, ika, ikus, ikas, izmu, izmus, ável, ível, ista, istas, ozu, oza, ozus, ozas, amentu, amentus, imentu, imentus, adora, adór, asaun, adoras, adores, asoens, ante, antes, ánsia, atória, atóriu, atórias, atórius, amentál	*general_suf*	A list contains general suffixes
lojia, lojias	*lojia_suf*	A list contains *loj* and *lojia* suffixes
usaun, usoens	*usaun_suf*	A list contains *usaun* and *usoens* suffixes
énsia, énsias	*ensia_suf*	A list contains *énsia* and *énsias* suffixes
amente	*amente_suf*	A string with *amente* value
iv (appears before the *amente* suffix)	*iv_suf*	A string with *iv* value
at (takes precedence over the *iv*, *iva*, *ivu*, *ivas*, or *ivus* suffixes)	*at_suf*	A string with *at* value
oz, ik, ad (presents before the *amente* suffix)	*ozikad_suf*	A list contains *oz*, *ik*, and *ad* suffixes
mente	*mente_suf*	A string with *mente* value
ante, avel, ivel (appears before the *mente* suffix)	*ante_suf*	A list contains *ante*, *avel*, and *ivel* suffixes
idade, idades	*idade_suf*	A list contains *idade* and *idades* suffixes
abil, is, iv (takes precedence over the *idade* or *idades* suffixes)	*abil_suf*	A list contains *abil*, *is*, and *iv* suffixes
iva, ivu, ivas, ivus	*iva_suf*	A list contains *iva*, *ivu*, *ivas* and *ivus* suffixes
ada, adu, adas, adus, ida, idu, idas, idus, ária, áriu, árias, árius	*verb_suf*	A list contains *verb* suffixes
a, e, i, u, us, as	*residual_suf*	A list contains *residual* suffixes

National Institute of Linguistics (2004). The standard orthography of the Tetun language: 115 years in the making, INL, August 2004.

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.

03 Heavy stemmer: removes suffixes from Portuguese-derived words and native Tetun, and prefixes from native Tetun.

Labadain **Stemmer**

Stemming algorithm (adapted from PT stemmer in Snowball)

Light Stemmer

- **Step 1:** Check if the word has four or more characters
  If not, return the original word
- **Step 2:** Apply general suffix removal
  If a suffix is removed, return the stemmed word
- **Step 3:** Try verb suffix removal
  If a suffix is removed, return the stemmed word
- **Step 4:** Try residual suffix removal
  If a suffix is removed, return the stemmed word
- **Step 5:** If no sufixes are removed, return the original word

```
Require: ensia_suf ← list of ensia suffixes
Require: amente_suf ← amente suffix
Require: iv_suf ← iv suffix
Require: at_suf ← at suffix
Require: ozikad_suf ← list of ozikad suffixes
Require: mente_suf ← mente suffix
Require: ante_suf ← list of ante suffixes
Require: idade_suf ← list of idade suffixes
Require: abil_suf ← list of abil suffixes
Require: iva_suf ← list of iva suffixes
Require: verb_suf ← list of verb suffixes
Require: residual_suf ← list of residual suffixes
for all word in word_list do                          ▷ Step 1: Word length validation
    if length(word) < 4 then
        Return word
    else                                              ▷ Step 2: Standard suffix removal
        if word ends with any suffix in general_suf sorted by length descending then
            if Position of suffix in word is in R2 then
                Delete suffix from word
                stem ← word without suffix
                Return stem
            end if
        else if word ends with any suffix in lojia_suf then
            if Position of suffix in word is in R2 then
                Replace suffix in word with loj
                stem ← word without suffix concatenates with loj
                Return stem
```

Pre-print paper: https://arxiv.org/abs/2412.11758
Labadain stemmer source code: https://github.com/gabriel-de-jesus/labadain-stemmer

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.

Query

Timor News logs

Document

Labadain-30k+ dataset

Relevance

Tetun-speaking assessors

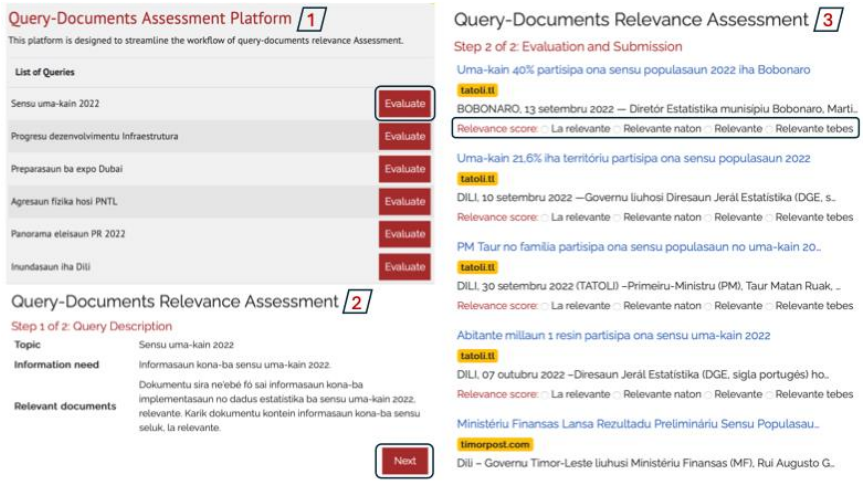
Labadain **Avaliadór**

A Tetun test collection for ad-hoc text retrieval

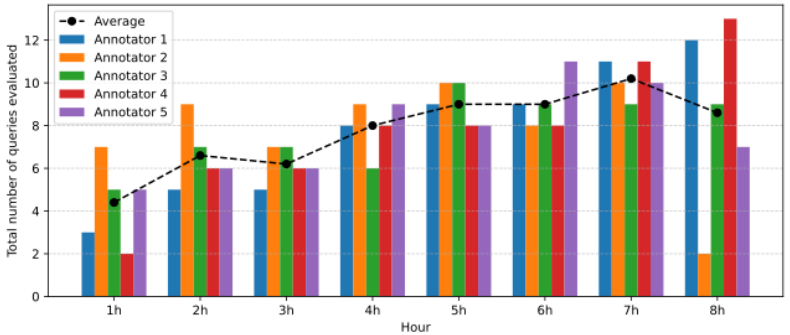
Relevance judgments

Four graded levels of relevance:
- 0 – Non-relevant
- 1 – Marginally relevant
- 2 – Relevant
- 3 – Highly relevant

Web interface used by human assessors



Total of queries evaluated by each annotator per hour



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

Labadain **Avaliadór**

A Tetun test collection for ad-hoc retrieval

Relevance judgments

Cohen's k score for inter-annotator agreement

	HA_1	HA_2	HA_3	HA_4	HA_5
Human annotator 1 (HA_1)	—	0.4344	0.4999	0.4434	0.4380
Human annotator 2 (HA_2)	0.4344	—	0.3745	0.3310	0.3500
Human annotator 3 (HA_3)	0.4999	0.3745	—	0.4646	0.4199
Human annotator 4 (HA_4)	0.4434	0.3310	0.4646	—	0.4063
Human annotator 5 (HA_5)	0.4380	0.3500	0.4199	0.4063	—
Average kappa score			**0.4236**		

Summary of the final test collection (Labadain-Avaliadór)

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

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

Dataset: https://doi.org/10.25747/2K6S-E518

Exploring LLMs for Relevance Judgments

An experiment with Tetun

Using the same test 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-graduasaun UNTL"
document: "Kursu Desportu UNTL sei realiza graduasaun 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".

LLMs evaluated and Cohen's k agreement with humans

	LLaMA3 70B	Claude3 Haiku	GPT-3.5-turbo-0125
Human	**0.2634**	0.2450	0.2462

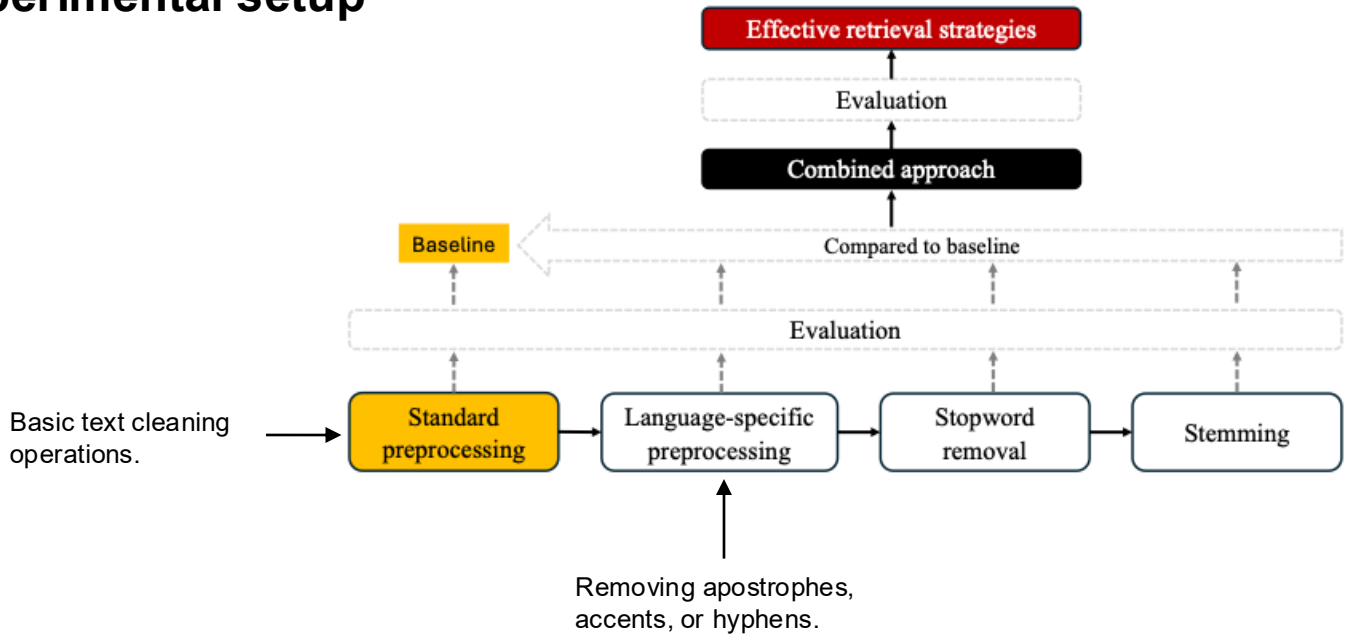
Comparison with the literature

	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
Ours	LLaMA3 70B	0.26

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

Classical Retrieval **Baselines**

Experimental setup



Classical Retrieval **Baselines**

Evaluation and results

Index compression factor (ICF) for each preprocessing setting

Description	Baseline	No Apostrophes		No Accents		No Hyphens		No Stopwords		Light		Moderate		Heavy	
		#	%	#	%	#	%	#	%	#	%	#	%	#	%
Title	25,412	25,258	0.61	23,568	7.25	17,596	30.76	25,256	0.61	23,416	7.86	23,377	8.01	23,283	8.38
Content	163,203	162,012	0.73	150,596	7.72	146,657	10.14	163,148	0.33	144,240	11.62	143,698	11.95	143,329	12.18

Writing variations and language-specific nature.

The highest ICF relative to the baseline.

As expected, the heavy stemmer achieved a higher ICF compared to the light and moderate variants.

Frakes and Fox (2003). Strength and similarity of affix removal stemming algorithms, vol 37, pages 26–30, SIGIR Forum, ACM, New York, USA, April 2003.

Classical Baselines

Evaluation and results

Short-text retrieval (document titles)

Individual language-specific preprocessing and stemming

Red values indicate scores lower than the baseline.

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 accents	BM25	0.7831	0.7542	0.6424	0.1370	0.2445	0.3735	0.6564	0.6512	0.6308	0.5744	0.7329
	DFR BM25	0.7831	0.7542	0.6441	0.1366	0.2440	0.3734	0.6566	0.6509	0.6317	0.5747	0.7328
	TF-IDF	0.7864	0.7508	0.6297	0.1364	0.2417	0.3666	0.6555	0.6477	0.6237	0.5631	0.7287
	Dirichlet LM	0.7390	0.7237	0.6331	0.1159	0.2173	0.3503	0.5945	0.6081	0.6036	0.5588	0.7104
	Hiemstra LM	0.7763	0.7441	0.6458	0.1341	0.2388	0.3732	0.6446	0.6406	0.6305	0.5880	0.7352
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
Light stemming	BM25	0.8000	0.7678	0.6500	0.1381	0.2464	0.3758	0.6693	0.6605	0.6355	0.5826	0.7364
	DFR BM25	0.8000	0.7661	0.6492	0.1376	0.2456	0.3748	0.6679	0.6588	0.6345	0.5824	0.7358
	TF-IDF	0.7966	0.7610	0.6407	0.1371	0.2421	0.3700	0.6648	0.6551	0.6304	0.5711	0.7331
	Dirichlet LM	0.7661	0.7203	0.6305	0.1225	0.2198	0.3518	0.6286	0.6192	0.6076	0.5613	0.7174
	Hiemstra LM	0.7898	0.7492	0.6576	0.1315	0.2365	0.3760	0.6456	0.6410	0.6326	0.5907	0.7344
Moderate stemming	BM25	0.7797	0.7610	0.6466	0.1336	0.2423	0.3714	0.6594	0.6553	0.6319	0.5790	0.7346
	DFR BM25	0.7797	0.7593	0.6458	0.1331	0.2415	0.3704	0.6581	0.6535	0.6309	0.5789	0.7341
	TF-IDF	0.7763	0.7542	0.6373	0.1326	0.2380	0.3656	0.6550	0.6499	0.6268	0.5676	0.7313
	Dirichlet LM	0.7593	0.7186	0.6254	0.1197	0.2175	0.3484	0.6225	0.6149	0.6025	0.5577	0.7138
	Hiemstra LM	0.7729	0.7390	0.6525	0.1275	0.2319	0.3710	0.6365	0.6337	0.6277	0.5868	0.7322
Heavy stemming	BM25	0.7797	0.7610	0.6466	0.1336	0.2423	0.3714	0.6594	0.6553	0.6319	0.5788	0.7346
	DFR BM25	0.7797	0.7593	0.6458	0.1333	0.2416	0.3705	0.6583	0.6537	0.6311	0.5788	0.7341
	TF-IDF	0.7763	0.7542	0.6373	0.1326	0.2380	0.3656	0.6550	0.6499	0.6268	0.5674	0.7312
	Dirichlet LM	0.7559	0.7186	0.6254	0.1189	0.2170	0.3479	0.6200	0.6145	0.6022	0.5571	0.7135
	Hiemstra LM	0.7729	0.7390	0.6534	0.1275	0.2319	0.3715	0.6365	0.6337	0.6282	0.5867	0.7322

Classical Baselines

Evaluation and results

Short-text retrieval: combined approach

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

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

Using DFR BM25, performance improved by up to **30.35%** over the baseline at NDCG@10.

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 apostrophes 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 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	0.4171	0.6933	0.6925	0.6918	0.6863	0.8065
	Hiemstra LM	0.8746	0.8305	0.7305	0.1524	0.2720	0.4372	0.7294	0.7211	0.7152	0.7040	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	0.2735	0.4353	0.7418	0.7218	0.7131	0.6756	0.8228
	TF-IDF	0.8847	0.8220	0.7203	0.1589	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
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	0.6800	0.6622	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 relative 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 relative 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 relative 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%**

Classical Baselines

Evaluation and results

Long-text retrieval (document content)

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	0.4847	0.4525	0.3839	0.0769	0.1281	0.1931	0.3883	0.3800	0.3765	0.3429	0.5764
	DFR BM25	0.4746	0.4475	0.3839	0.0758	0.1259	0.1925	0.3826	0.3763	0.3753	0.3416	0.5754
	TF-IDF	0.5288	0.4746	0.4110	0.0855	0.1382	0.2068	0.4275	0.4086	0.4045	0.3564	0.5927
	Dirichlet LM	0.4576	0.4186	0.3669	0.0696	0.1111	0.1701	0.3577	0.3544	0.3545	0.3110	0.5558
	Hiemstra LM	0.5390	0.4915	0.4314	0.0856	0.1402	0.2138	0.4289	0.4158	0.4189	0.3655	0.5990
Remove hyphens	BM25	0.5322	0.4966	0.4407	0.0869	0.1437	0.2265	0.4315	0.4270	0.4324	0.4042	0.6570
	DFR BM25	0.5288	0.4949	0.4407	0.0860	0.1425	0.2250	0.4259	0.4247	0.4302	0.4026	0.6547
	TF-IDF	0.5966	0.5390	0.4686	0.0942	0.1580	0.2425	0.4812	0.4677	0.4649	0.4224	0.6783
	Dirichlet LM	0.5051	0.4729	0.4153	0.0813	0.1338	0.2034	0.3961	0.3967	0.4014	0.3729	0.6364
	Hiemstra LM	0.5797	0.5542	0.4907	0.0946	0.1620	0.2503	0.4688	0.4726	0.4771	0.4338	0.6827
Remove apostrophes and hyphens	BM25	0.5322	0.4949	0.4407	0.0869	0.1434	0.2271	0.4300	0.4254	0.4328	0.4047	0.6567
	DFR BM25	0.5254	0.4932	0.4407	0.0857	0.1423	0.2256	0.4230	0.4232	0.4307	0.4031	0.6543
	TF-IDF	0.5966	0.5390	0.4678	0.0944	0.1581	0.2432	0.4788	0.4664	0.4650	0.4232	0.6790
	Dirichlet LM	0.5017	0.4729	0.4136	0.0809	0.1348	0.2043	0.3953	0.3983	0.4011	0.3737	0.6369
	Hiemstra LM	0.5831	0.5576	0.4907	0.0955	0.1636	0.2523	0.4721	0.4759	0.4790	0.4355	0.6840
Remove hyphens and accents	BM25	0.5322	0.4932	0.4280	0.0875	0.1442	0.2222	0.4349	0.4251	0.4273	0.4020	0.6585
	DFR BM25	0.5254	0.4932	0.4280	0.0861	0.1439	0.2211	0.4316	0.4272	0.4278	0.4007	0.6578
	TF-IDF	0.5831	0.5288	0.4525	0.0948	0.1541	0.2339	0.4744	0.4588	0.4555	0.4175	0.6769
	Dirichlet LM	0.5017	0.4576	0.4025	0.0779	0.1261	0.1923	0.3882	0.3847	0.3875	0.3645	0.6288
	Hiemstra LM	0.5831	0.5508	0.4763	0.0955	0.1645	0.2474	0.4812	0.4748	0.4737	0.4330	0.6855
Remove hyphens and stopwords	BM25	0.5797	0.5220	0.4619	0.0926	0.1547	0.2361	0.4614	0.4492	0.4521	0.4195	0.6722
	DFR BM25	0.5763	0.5203	0.4568	0.0918	0.1549	0.2344	0.4585	0.4483	0.4491	0.4189	0.6718
	TF-IDF	0.6102	0.5339	0.4661	0.0961	0.1574	0.2403	0.4849	0.4622	0.4612	0.4232	0.6777
	Dirichlet LM	0.4983	0.4695	0.4068	0.0742	0.1276	0.1946	0.3841	0.3860	0.3886	0.3604	0.6235
	Hiemstra LM	0.5932	0.5458	0.4797	0.0947	0.1601	0.2451	0.4780	0.4722	0.4714	0.4339	0.6840
Remove hyphens, apostrophes, and stopwords	BM25	0.5729	0.5153	0.4610	0.0922	0.1540	0.2366	0.4575	0.4457	0.4514	0.4199	0.6726
	DFR BM25	0.5695	0.5136	0.4576	0.0913	0.1542	0.2352	0.4545	0.4447	0.4492	0.4193	0.6720
	TF-IDF	0.6102	0.5322	0.4661	0.0960	0.1570	0.2410	0.4838	0.4618	0.4608	0.4236	0.6779
	Dirichlet LM	0.4881	0.4678	0.4042	0.0731	0.1270	0.1940	0.3786	0.3848	0.3858	0.3611	0.6232
	Hiemstra LM	0.6000	0.5492	0.4797	0.0959	0.1628	0.2471	0.4826	0.4760	0.4727	0.4358	0.6853
Remove hyphens, apostrophes, and moderate stemmer	BM25	0.5322	0.5017	0.4322	0.0849	0.1459	0.2236	0.4156	0.4176	0.4202	0.3981	0.6450
	DFR BM25	0.5254	0.4966	0.4297	0.0844	0.1447	0.2214	0.4110	0.4136	0.4173	0.3968	0.6435
	TF-IDF	0.6000	0.5441	0.4636	0.0942	0.1583	0.2364	0.4681	0.4589	0.4561	0.4160	0.6700
	Dirichlet LM	0.4915	0.4712	0.4076	0.0788	0.1330	0.1989	0.3795	0.3872	0.3911	0.3663	0.6274
	Hiemstra LM	0.5864	0.5559	0.4805	0.0965	0.1631	0.2475	0.4695	0.4734	0.4702	0.4301	0.6780
Average relative performance gains of the best model compared to the baseline		*15.39%*	*13.45%*	*13.75%*	*12.73%*	*17.33%*	*18.01%*	*13.43%*	*14.48%*	*14.35%*	*19.23%*	*14.44%*
Average relative performance gains of the best model compared to others within the same retrieval strategy		**9.61%**	**11.77%**	**11.47%**	**13.23%**	**16.40%**	**12.54%**	**9.64%**	**10.15%**	**11.08%**	**7.82%**	**4.65%**

Hiemstra LM performed best in almost all metrics and cutoffs.

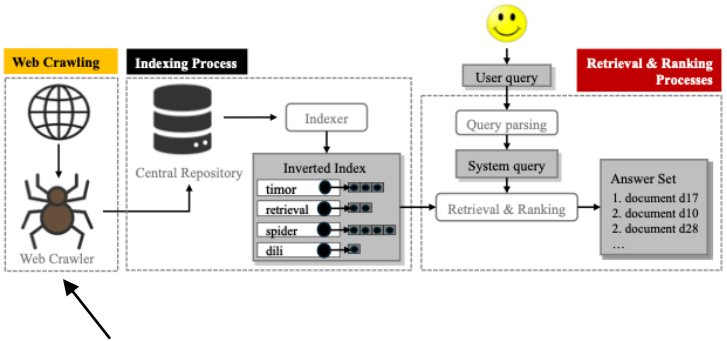
Using Hiemstra LM, performance improved by up to **19.23%** over the baseline at overall MAP.

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

Labadain **Search**

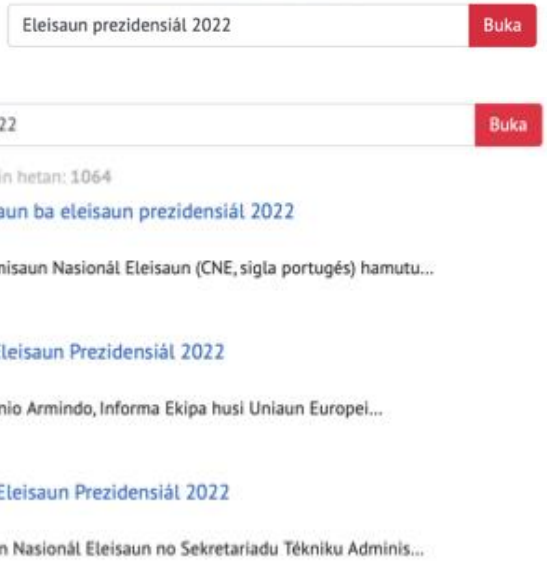
Showcasing the Best Retrieval Strategy for Tetun

General Architecture



Labadain Crawler

www.labadain.tl →



Zero-Shot and Hybrid Retrieval

Overview

Objective

Applying pre-trained models to Tetun without additional training or fine-tuning

Investigate the performance of dense retrieval models under zero-shot and hybrid strategies in out-of-distribution scenarios.

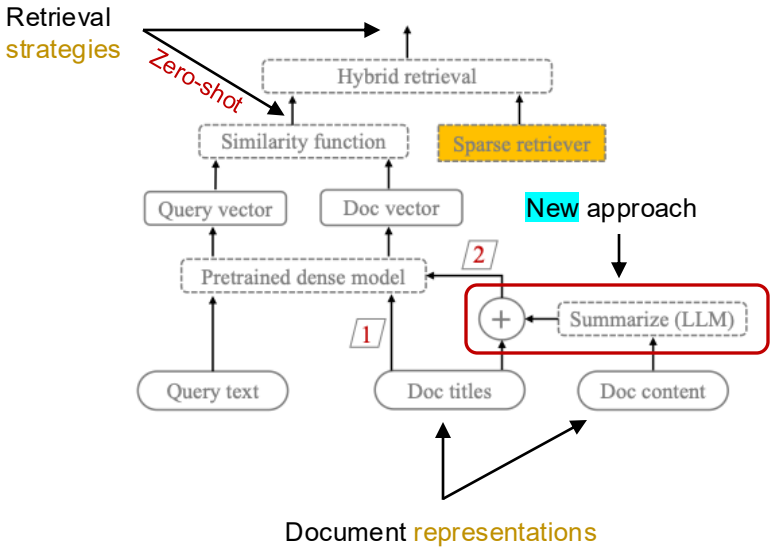
Combining traditional (sparse) and dense retrievers

Evaluating models on Tetun, a language they were not trained on.

Zero-Shot and Hybrid Retrieval

Methodology

Two factors studied



Four Retrieval **strategies** are defined:

- Basic zero-shot: adapted pretrained dense retrieval to queries and document titles for embeddings and evaluate their performance.
- Contextualized zero-shot: as above, but titles + document summaries for embeddings.
- Basic hybrid: basic zero-shot + sparse retriever.
- Contextualized hybrid: contextualized zero-shot + sparse retriever.

Zero-Shot and Hybrid Retrieval

Document summarization

Configurations

- Approach: one-shot prompting, with one example per LLM call.
- LLM: Claude 3 Haiku.
- Prompt engineering: qualitatively evaluated multiple prompt variants based on observations.

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

Hybrid framework

Score fusion through a linear combination

Literature

$$Score\ (q,d) = \alpha\ .Score_{lex}(\mathrm{q},\mathrm{d}) + Score_{sem}(\mathrm{q},\mathrm{d})$$
$$Score\ (q,d) = Score_{lex}(\mathrm{q},\mathrm{d}) + \alpha\ .Score_{sem}(\mathrm{q},\mathrm{d})$$
$$Score\ (q,d) = \alpha\ .Score_{lex}(\mathrm{q},\mathrm{d}) + (\alpha\ - 1)\ .Score_{sem}(\mathrm{q},\mathrm{d})$$

Our proposal

$$Score\ (q,d) = \alpha .Score_{lex}(\mathrm{q},\mathrm{d}) + \beta\ .Score_{sem}(\mathrm{q},\mathrm{d})$$

Dual-parameter tuning

A new parameter added

Intuition

Sparse and dense retrievers may assign different scores to the same document, and therefore their signals should be balanced to improve relevance.

Zero-Shot and Hybrid Retrieval

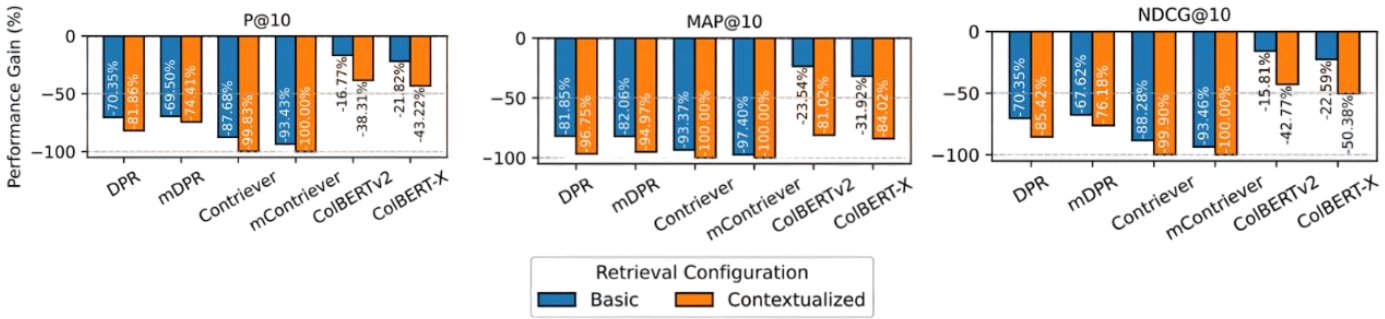
Experiment and results

Baselines

Model	P@10	MAP@10	NDCG@10
BM25	0.8373	0.2796	0.7347
DFR BM25	**0.8390**	**0.2804**	**0.7356**
Hiemstra LM	0.8305	0.2743	0.7245

Model performance was compared against the DFR BM25 baseline.

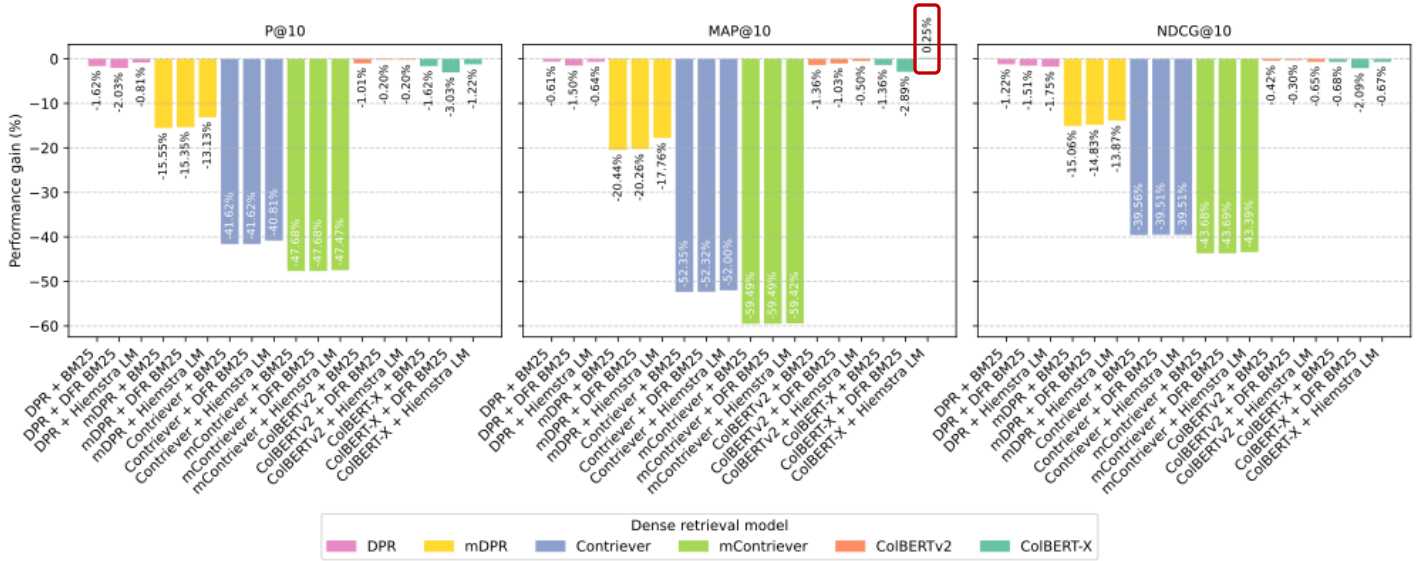
Zero-shot retrieval: relative performance gain over baseline



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, achieving up to **+4.24%** at MAP@10 ($p < 0.05$).

ColBERTv2 + sparse with a single parameter also outperformed the baseline, achieving up to **+4.03%** at MAP@10 when the parameter was assigned to the dense retriever.

	P@10	MAP@10	NDCG@10
Baseline			
DFR BM25	0.8390	0.2804	0.7356
Contextualized hybrid retrieval ($\alpha = 1.1$ and $\beta = 0.8$, top 2k results)			
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%)†	0.2226 (-20.61%)†	0.6236 (-15.23%)†
mDPR + DFR BM25	0.7102 (-15.35%)†	0.2234 (-20.33%)†	0.6262 (-14.87%)†
mDPR + Hiemstra LM	0.7288 (-13.13%)†	0.2305 (-17.80%)†	0.6334 (-13.89%)†
Contriever + BM25	0.3169 (-62.23%)†	0.0791 (-71.79%)†	0.2938 (-60.06%)†
Contriever + DFR BM25	0.3169 (-62.23%)†	0.0791 (-71.79%)†	0.2947 (-59.94%)†
Contriever + Hiemstra LM	0.3220 (-61.62%)†	0.0796 (-71.61%)†	0.2985 (-59.42%)†
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	**0.8441 (+0.61%)**†	**0.2862 (+2.07%)**†	**0.7453 (+1.32%)**†
ColBERTv2 + DFR BM25	**0.8458 (+0.81%)**†	**0.2858 (+1.93%)**†	**0.7480 (+1.69%)**†
ColBERTv2 + Hiemstra LM	**0.8559 (+2.01%)** †	**0.2923 (+4.24%)** †	**0.7536 (+2.45)** †
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%)	**0.2807 (+0.11%)**	0.7303 (-0.72%)
Contextualized hybrid retrieval (single parameter, $\alpha = 1.1$, top 2k results)			
ColBERTv2 + BM25	**0.8508 (+1.41%)**†	**0.2888 (+3.00%)**†	**0.7493 (+1.86%)**†
ColBERTv2 + DFR BM25	0.8373 (-0.20%)†	**0.2831 (+0.96%)**†	**0.7438 (+1.11%)**†
ColBERTv2 + Hiemstra LM	**0.8475 (+1.01%)**†	**0.2891 (+3.10%)**†	**0.7504 (+2.01%)**†
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%)
Contextualized hybrid retrieval (single parameter, $\beta = 0.8$, top 2k results)			
ColBERTv2 + BM25	**0.8458 (+0.81%)**†	**0.2866 (+2.21%)**†	**0.7467 (+1.51%)**†
ColBERTv2 + DFR BM25	**0.8441 (+0.61%)**†	**0.2856 (+1.85%)**†	**0.7476 (+1.63%)**†
ColBERTv2 + Hiemstra LM	**0.8542 (+1.81%)**†	**0.2917 (+4.03%)**†	**0.7530 (+2.37%)**†
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%)
Reciprocal Rank Fusion			
ColBERTv2 + BM25	0.7949 (-5.26%)†	0.2641 (-5.81%)†	0.7164 (-2.61%)†
ColBERTv2 + DFR BM25	0.7966 (-5.05%)†	0.2642 (-5.78%)†	0.7166 (-2.58%)†
ColBERTv2 + Hiemstra LM	0.8220 (-2.03%)†	0.2738 (-2.35%)†	0.7316 (-0.54%)†
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

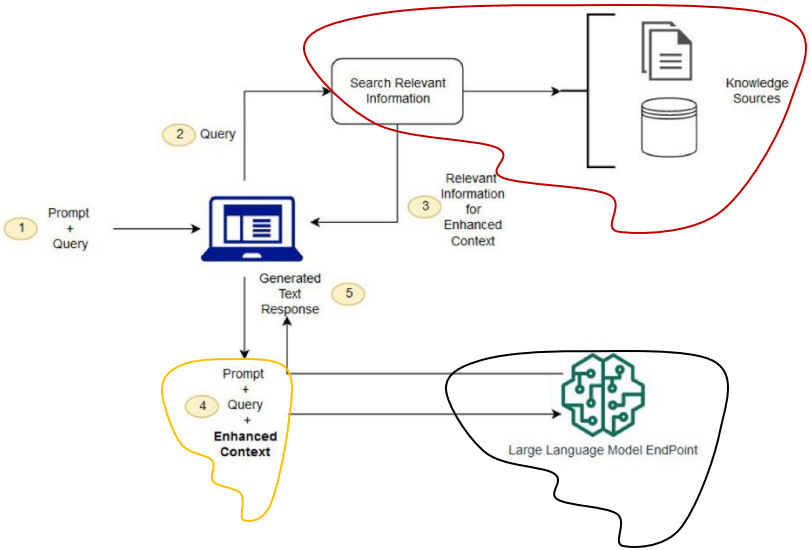
AI for Tetun

AI for Tetun (RAG Implementation)

What is RAG?

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

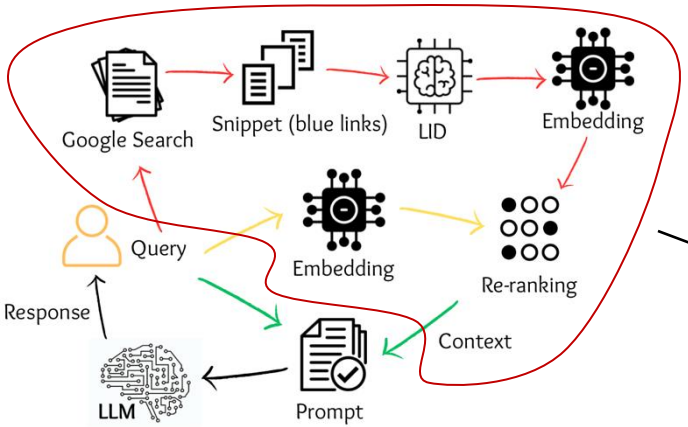


General RAG Architecture

Minaee et al. (2025). Large Language Models: A Survey. arXiv (2025). URL: https://arxiv.org/abs/2402.06196

Labadain **Chat** (AI for Tetun)

Architecture and Application

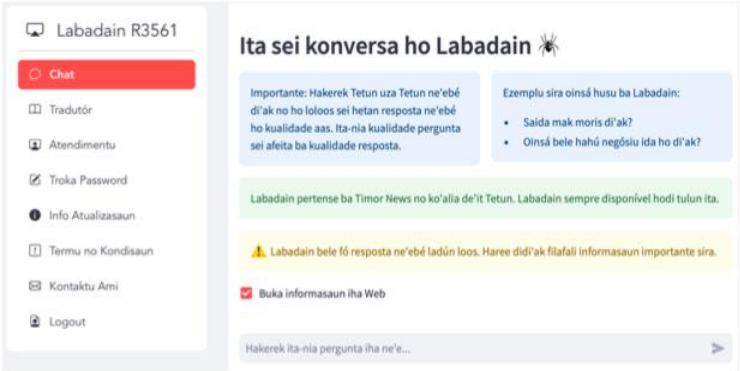


www.labadain.com

Learning from Labadain's Users

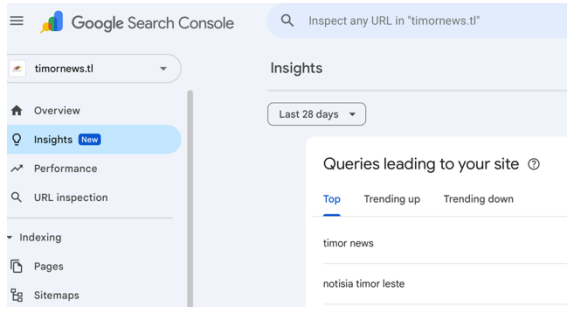
Search Log Analysis

Labadain Chat ← www.labadain.com

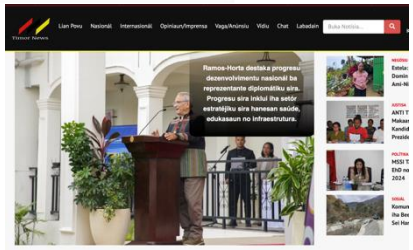


Labadain Search ← www.labadain.tl

Google Search Console from Timor News

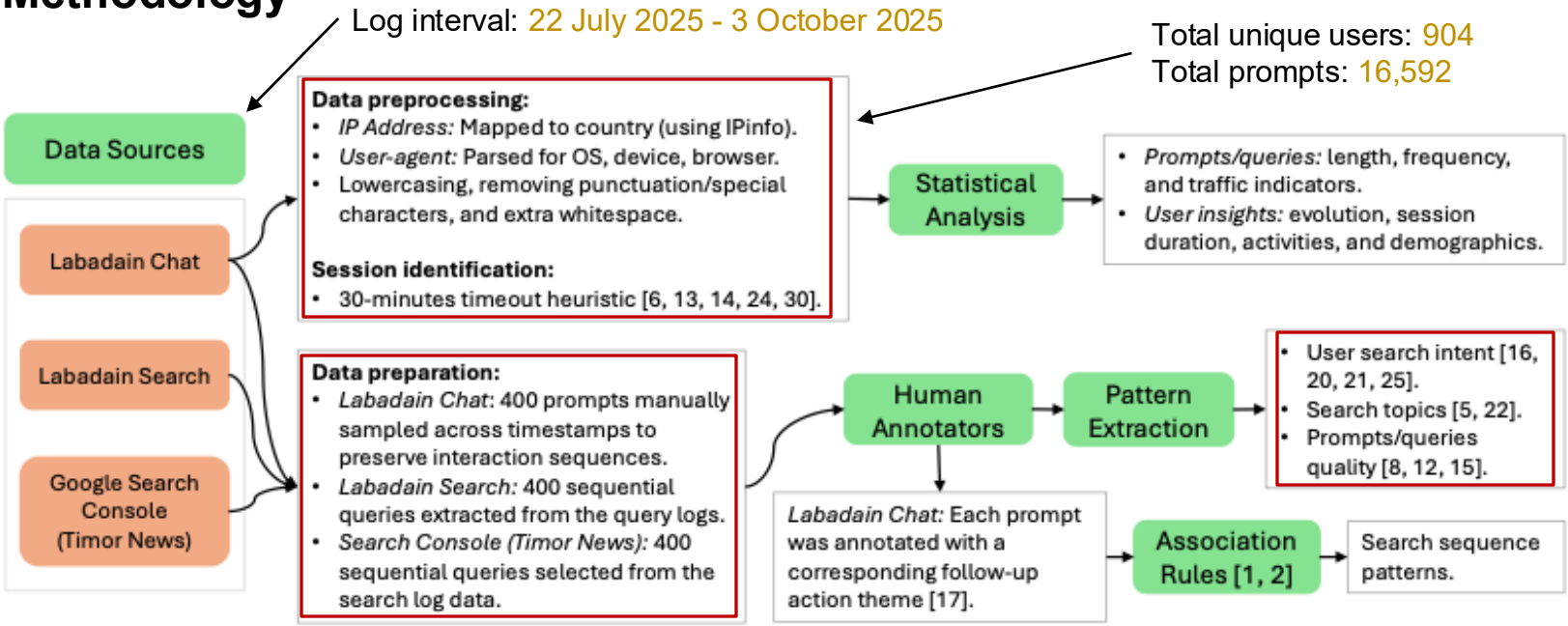


www.timornews.tl →



Learning from Labadain's Users

Methodology

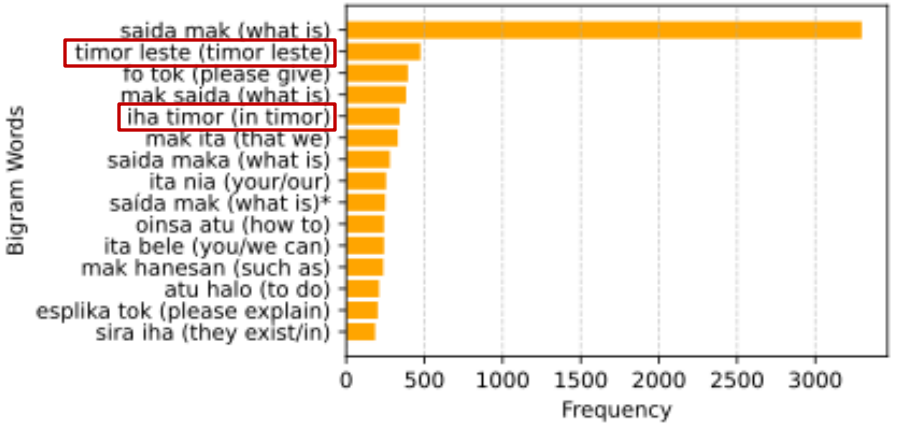


Learning from Labadain's Users

Results

Prompt log Analysis

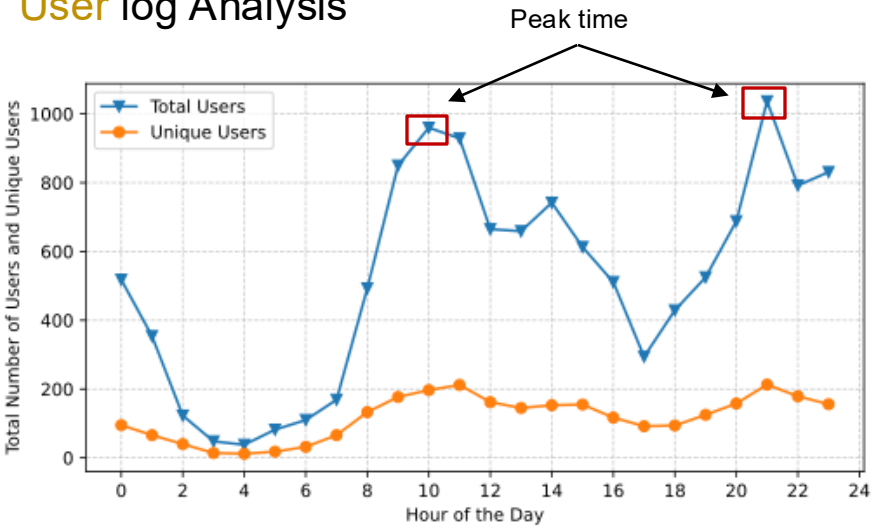
Query Type	#Prompt	Proportion	#User
Long prompts	107	0.63%	25
Unique long prompts	104	0.61%	25
Medium prompts	2,781	16.41%	436
Unique medium	2,647	15.62%	436
Short prompts	14,064	82.86%	884
Unique short prompts	12,512	73.81%	884

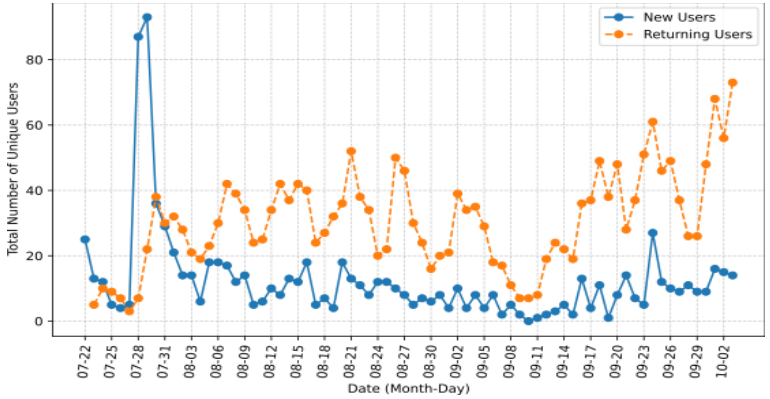


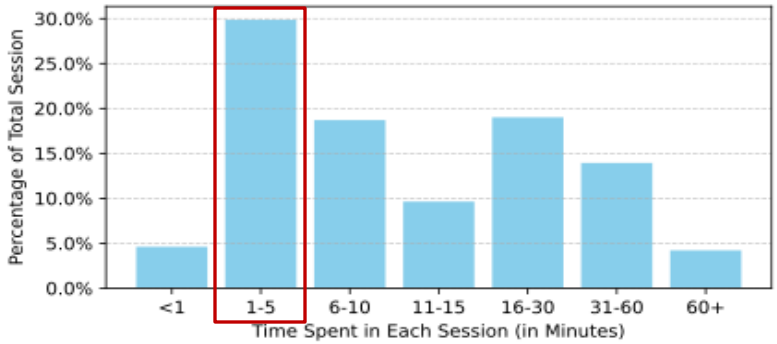
Learning from Labadain's Users

Results

User log Analysis







Learning from Labadain's Users

Results

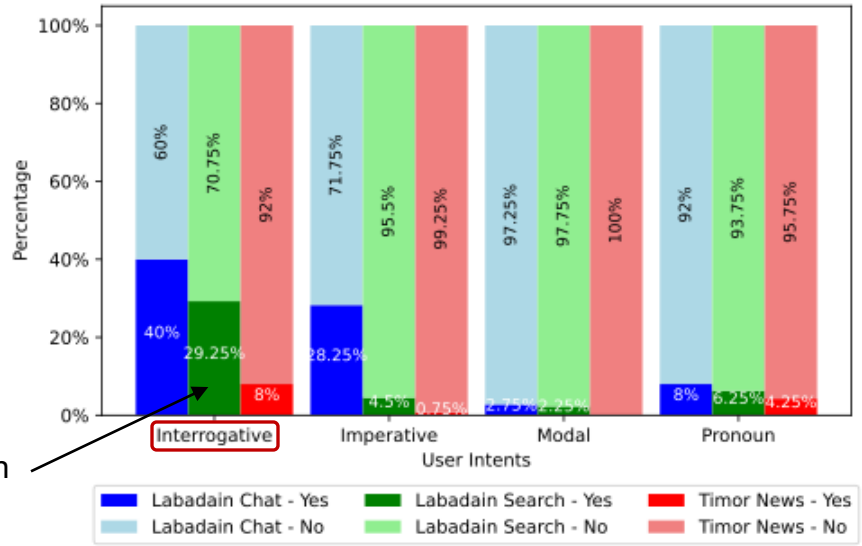
Traditional and LLM-based search

Distribution of user search intent

Search Intent	L. Chat	L. Search	Timor News
Information-seeking	87.75%	98.00%	99.50%
Task-oriented	10.50%	0.00%	0.00%
Conversational	1.75%	1.50%	0.00%
Other	0.00%	0.50%	0.50%

A global trend towards command-based instructions is evident in Labadain search (a traditional search engine developed in the era of LLMs).

Categories of user search intent



Learning from Labadain's Users

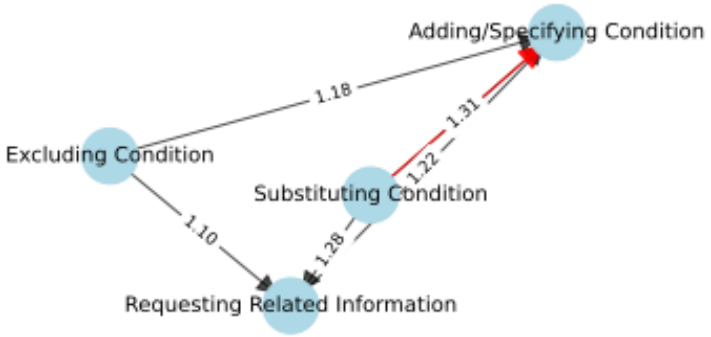
Results

Traditional and LLM-based search

Search topics

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%**

Follow-up search sequences



Paper: https://dl.acm.org/doi/abs/10.1145/3731120.3744596

Dataset: https://doi.org/10.25747/X2DK-5Y06

Why AI Matters for Tetun?

Rapid AI evolution

We are in an era where AI empowers education, government, and innovation.

Empowering Local Knowledge

AI trained in Tetun can preserve local expressions and ways of thinking.

Inclusive digital future

Ensure every citizen can access digital services, information, and education in Tetun.

Strengthening Governance

AI-driven e-government services make public communication more inclusive.

Bridging digital divide

Without Tetun-enabled AI tools, digital systems risk excluding the majority who use Tetun daily.

Building Tech. Sovereignty

Developing AI for Tetun fosters national digital independence.

Takeaways and Future Directions

Key Takeaways

Language is the key to inclusion

A digital future that speaks Tetun ensures that everyone can access, understand, and benefit from technology.

The foundation is set, but capacity remains limited

We now have initial resources, but continued research, training, and institutional support are crucial.

AI for Tetun is both cultural and national

Building AI in Tetun preserves identity, strengthens sovereignty, and embeds local knowledge in digital systems.

Collaboration is essential for progress

Academia, government, and int. partners must work together to develop tools, data, and infrastructure for sustainable growth.

The future is inclusive, intelligent, and local

Empowering Tetun through AI ensures Timor-Leste's digital transformation truly belongs to its people.

Challenges and Open Issues

Tetun text corpora remain limited.

Limited technical expertise in AI within Timor-Leste.

Few institutions focus on language technology research and development.

Insufficient funding for Tetun language innovation.

Advancing AI for Tetun remain a major challenge

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.

Thank You

AI for Tetun: Building Timor-Leste's Inclusive Digital Future

Empowering Language, Culture, and Innovation

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

TLNOG2
November 21, 2025
F-FDTL Auditorium Room
Dili, Timor-Leste