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

**Gabriel de Jesus**, Sérgio Nunes

INESC TEC / Faculty of Engineering, University of Porto (FEUP)

3rd Annual Meeting of the Special Interest Group on Under-resourced Languages

Lingotto Conference Centre, Turin, Italy
20 - 21 May, 2024

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

fct Fundação para a Ciência e a Tecnologia

INESCTEC TECHNOLOGY & SCIENCE | ASSOCIATE LABORATORY PORTUGAL

# Outline

- Introduction
- Our Contributions
- Related Work
- Dataset Construction
- Analysis and Discussions
- Conclusions & Future Work

# Introduction

- Text corpora are crucial for advancements in IR and NLP tools.
- Problems in constructing datasets for LRLs: lack of high-quality data, prevalence of informal text, limited Wikipedia resources, among others.
- We face similar problems in the construction of text corpus for Tetun.

# Introduction

Tetun

- The most widely spoken language in Timor-Leste.
- One of Timor-Leste's Official languages alongside Portuguese.

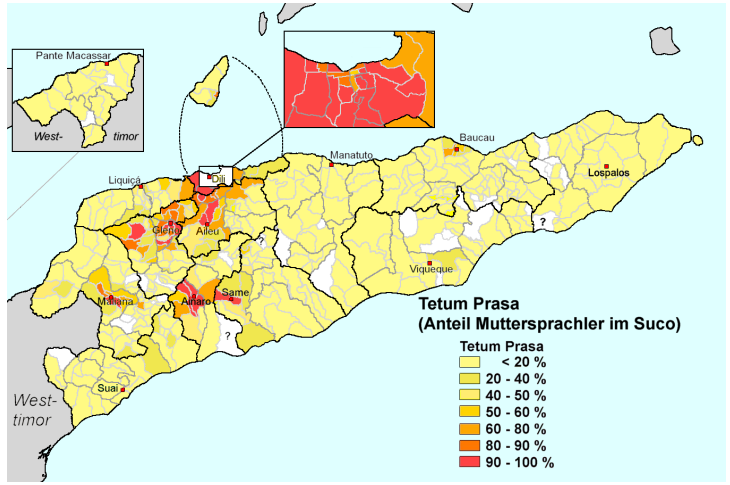
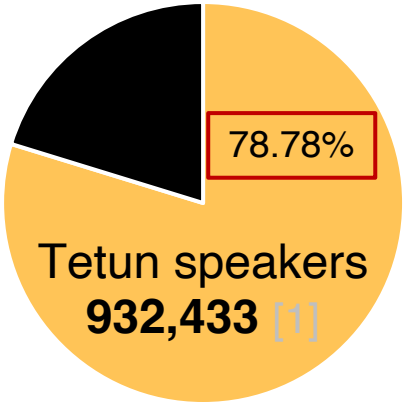



Image source: Wikipedia

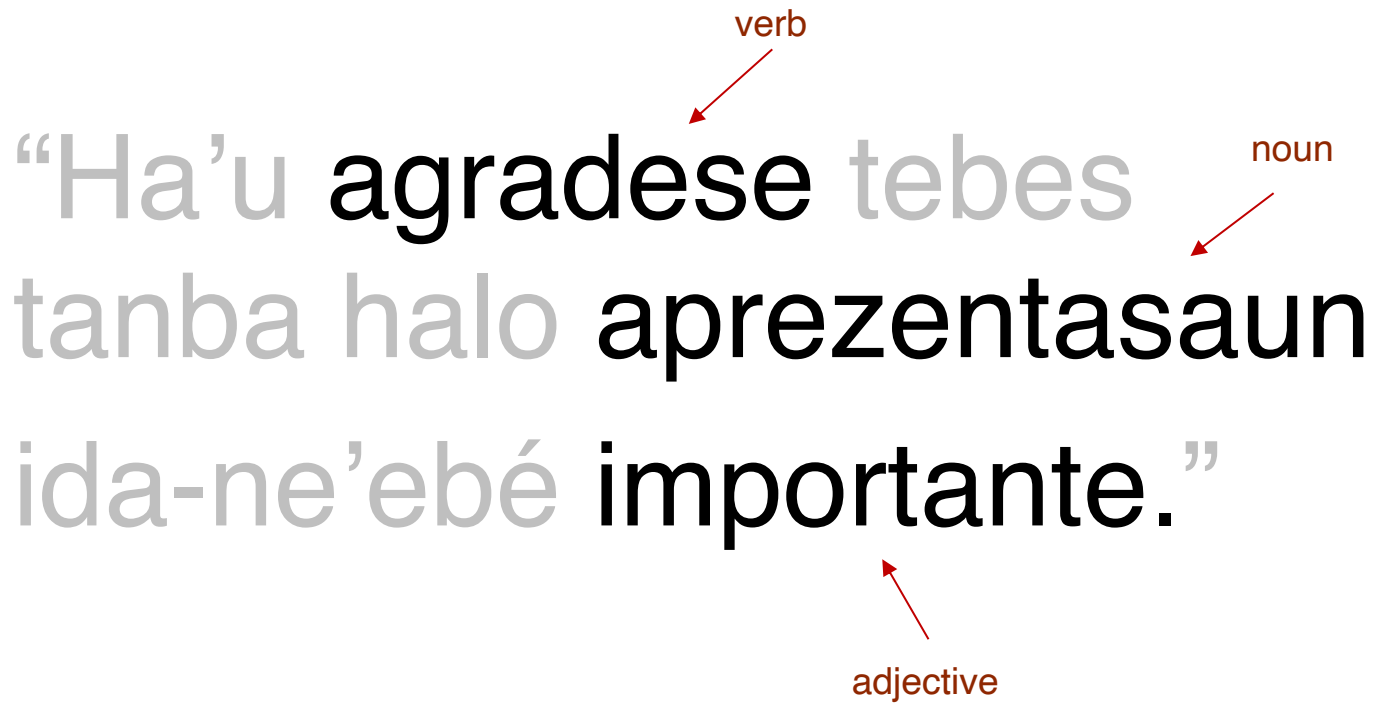



1.18 million population

[1] Gabriel de Jesus. Text Information Retrieval in Tetun. In Adv. in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III, volume 13982 of Lecture Notes in Computer Science, pages 429–435. Springer, 2023.

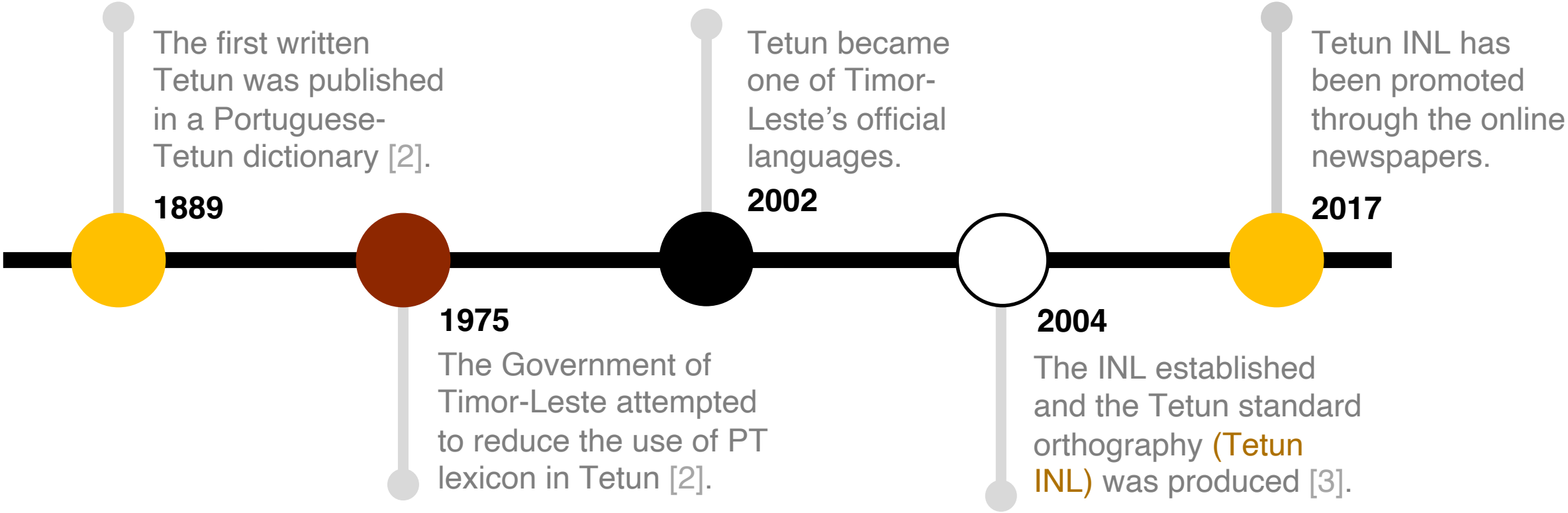
# Introduction

Portuguese loanwords



# Introduction

## Tetun evolution




[2] Zuzana Greksakova. Tetun in Timor-Leste: The Role of Language Contact in Its Development. PhD thesis, Universidade de Coimbra, Portugal, 2018.

[3] Timor-Leste's government Decree-Law no. 1/2004 of April 2004. URL: http://mj.gov.tl/jornal/lawsTL/RDTL-Law/RDTL-Gov-Decrees/Gov-Decree-2004-01.pdf

# Our Contributions

- A Tetun text corpus, audited at the document-level by native Tetun speakers.
- Insights of the Tetun text documents evolution on the web.
- Insights of the Tetun INL and Portuguese loanwords evolution.

# Related Work

## Kudugunta et al. (2023):

- Released MADLAD-400, a multilingual dataset obtained by processing CC snapshots (2008-2023).
- This dataset contains 40.4k Tetun text documents.
- Each document is separated by two consecutive newlines without any additional properties.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. CoRR, abs/2309.04662

# Related Work

## Wikimedia (2023):

- Released a multilingual Wikipedia dataset generated from Wiki dumps up to November 2023.
- This dataset contains 1.5k Tetun documents.
- Each document comprises ID, URL, title, and content.

The Foundation of Wikimedia. 2023. Wikimedia Downloads.

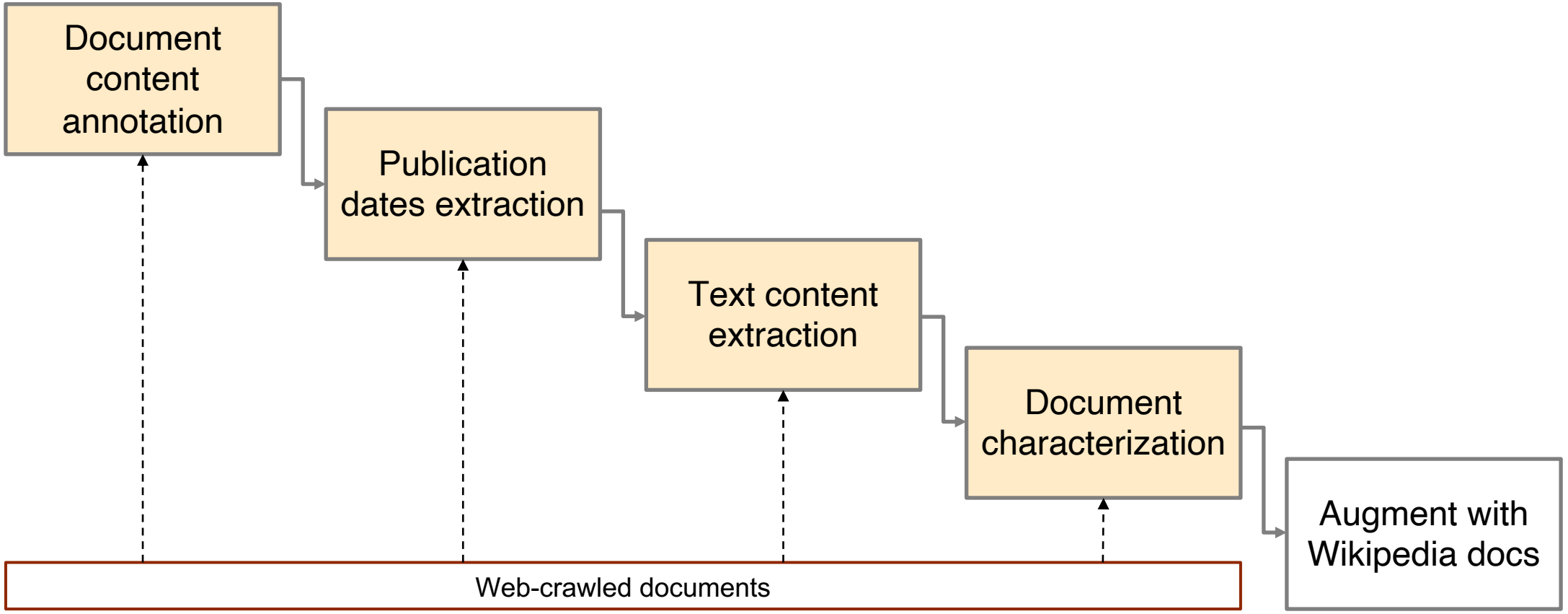
# Dataset Construction

## Document collections

- A collection of 32.1k Tetun text document we acquired from web crawling (title, URL and content).
- A collection of 1.5k Tetun Wikipedia articles from Wikimedia.

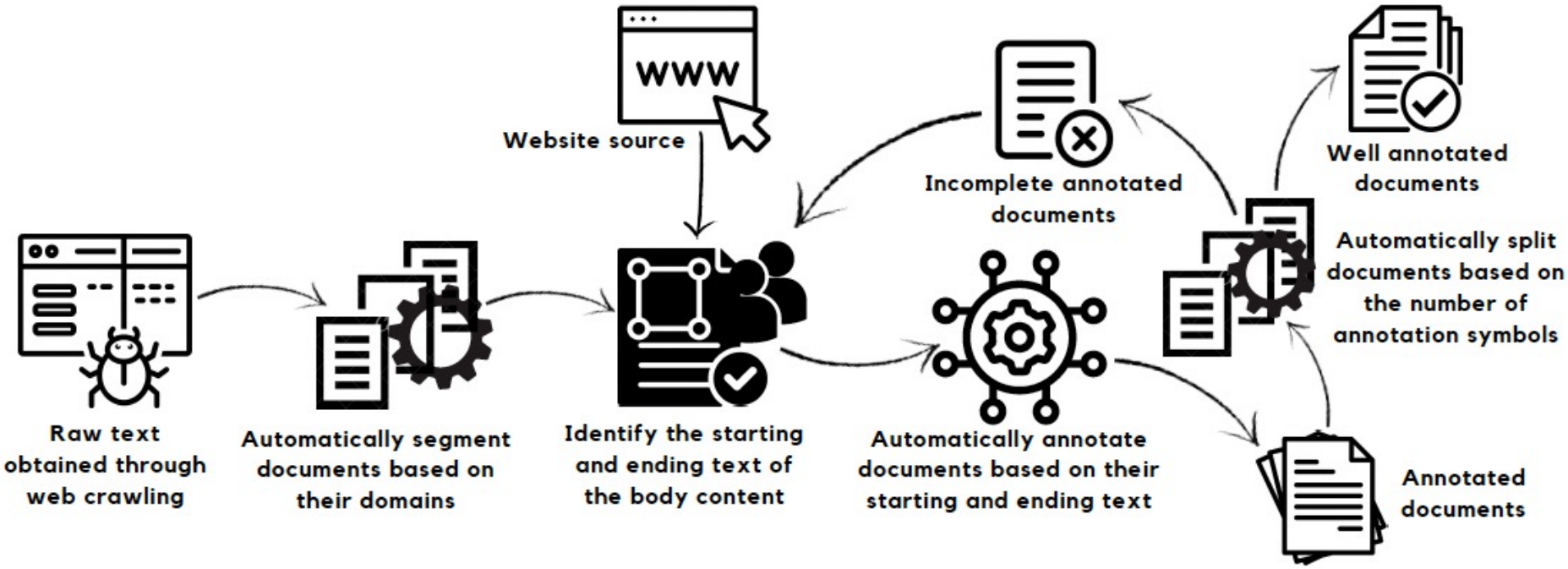
# Dataset Construction

## Web-crawled documents processing



# Dataset Construction

## Content annotation process



# Dataset Construction

## Content Annotation Algorithm

**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 $output\_file$ ▷ Refers to the "annotated documents" file in Figure 1.
  get $body\_content$ from $document$
  $annotation\_t\_counter \leftarrow 0$ ▷ 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 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**

# Dataset Construction

## Publication dates extraction

- Extract from URLs if they contain publication dates using regular expressions.

```
https://tatoli.tl/2024/05/14/mjdac-federasaun ....
```

- Extract from the web page source using the Beautiful Soup library.

```
Date in web page: December 18, 2023

CSS tag:
<time class="entry-date published" datetime="2023-
12-18T00:08:43+09:00">December 18, 2023</time>
```

# Dataset Construction

Content extraction and processing

- Extract the content text between the annotation symbols for each document.
- Deduplication: excluding documents that have the same title and URL.
- Generate document sources using their internet domain names.

# Dataset Construction

## Web-crawled dataset details

| Data source | #docs | Proportion |
|---|---|---|
| Online newspapers | 28,997 | 90.30% |
| Non-gov. portals | 1,889 | 5.88% |
| Government portals | 775 | 2.41% |
| Education portals | 184 | 0.57% |
| Blogs and Forums | 145 | 0.45% |
| Personal Pages | 74 | 0.23% |
| Banks and courts | 31 | 0.10% |
| Wikipedia | 18 | 0.06% |

The highlighted rows, corresponding a total of 2.9k documents, were chosen for characterization.

# Dataset Construction

Document characterization

- Three native Tetun speaker students carry out the characterization task.
- Each doc is categorized into one of seven predefined categories following guidelines.
- The characterize docs resulted in an inter-annotators agreement of Fleiss' *k* score of **0.4994**.

# Dataset Construction

Wikipedia documents processing

- Generate publication dates and document source using the previous approach.
- Perform HTML tags removal to ensure the cleanness of the documents.
- Each annotator manually audits the contents of approx. 500 documents.

# Dataset Construction

## Final dataset summary

| 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 | 0.10% |

The final dataset "Labadain-30k+ dataset," each document comprises a title, URL, source, publication date, and content.

# Comprehensive Content Analysis

## Labadain-30k+ corpus

| | |
|---|---|
| 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 |

## Document details

| | 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 |

## Documents by 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% |

## Documents 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% |

# Comprehensive Content Analysis

Documents 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 |

- Consistent increase has been observed since 2012.
- A notable surge occurred in 2020 (+11.75pp).
- There has decreased in 2023 as the crawled-data only cover up to Sep. 2023.

# Comprehensive Content Analysis

## Evolution of Tetun writing and PT loanwords

| | Before 2017 | | From 2017 to 2023 | | Difference |
|---|---|---|---|---|---|
| Words count in the corpus+ | 1,239,663 | | 10,689,158 | | ↑9.5M |
| Words count in the INL dictionary | 869,314 | 70.13% | 8,150,747 | **76.25%** | ↑6.12 pp |
| Words count in the loanword dictionary* | 286,493 | 23.11% | 3,014,218 | **28.20%** | ↑5.09 pp |
| Words count not found in the dictionaries | 331,090 | **26.71%** | 2,162,351 | 20.23% | ↓6.48 pp |

- Tetun INL standard usage in docs writing improved by +6.12pp from 2017 onwards.
- The use of Portuguese loanwords increased by +5.09pp along that trajectory.

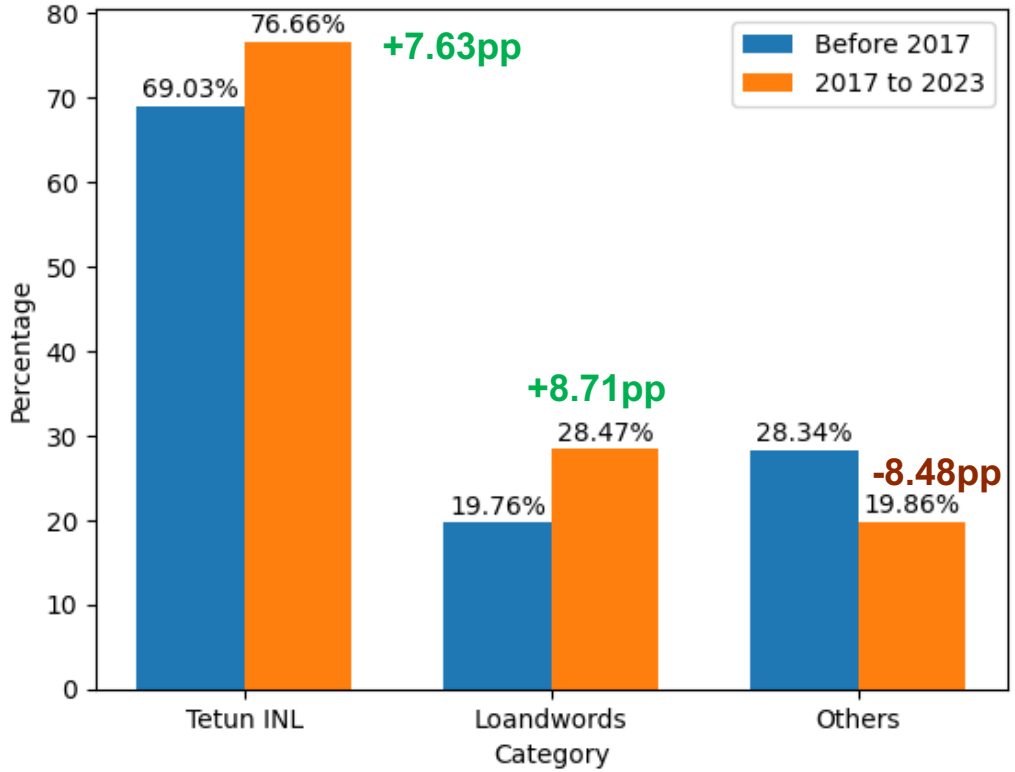
**Ground Truths**

Dictionary of INL

Dictionary of Portuguese loanwords

# Comprehensive Content Analysis

Evolution of Tetun writing and PT loanwords usage in news articles



# Discussions

## Comparison of Labadain-30k+ with other LRLs

| Language | #docs | #speakers |
|---|---|---|
| Tetun | 33.6k | 932k+ |
| Assamese | 33.8k[1] | 15M+[2] |
| Occitan | 36.4k[1] | 1.5M[3] |
| Mizo | 36.4k[1] | ~1M[4] |
| Swiss German | 42.7k[1] | 5M+[5] |

- Tetun, Occitan, and Mizo have comparable data size and number of speakers.
- Tetun has fewer speakers but is comparable data size to that of Assamese and Swiss German.

# Discussions

## Comparison of findings on PT loanwords

| | **Dataset** | **Prop. of PT Loanwords** | **Observations** |
|---|---|---|---|
| Van-Klinken and Hajek (2018) | Seven newspaper articles of 2009 | 32.00% | |
| Greksáková (2018) | 73,892 words from interview scripts | 35.00% | An increased of **+5pp** from 2018 to 2019. |
| Van-Klinken and Hajek (2019) | Newspapers and technical documents | 40.00% | |
| **Ours** | **818** documents in 2018 | 30.01% | A modest increase of **+3.5pp** and a lower overall percentage compared to the above findings. |
| | **1,164** documents in 2019 | 33.51% | |

# Conclusions and Future Work

- We introduce Labadain-30k+ and make publicly available for IR and NLP researchers.
- Labadain-30k+ is comparably sized to Tetun docs in MADLAD-400 (~6.8k fewer), yet it provides more contextual information.
- Labadain-30k+ is the first Tetun dataset audited by native Tetun speakers.
- In future work, we plan to use Labadain-30k+ for developing a list of Tetun stopwords, a stemmer, and an ad-hoc test collection for IR.

# Thank You

Scan here to access the Labadain-30k+ dataset!

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

**Gabriel de Jesus**, Sérgio Nunes

INESC TEC / Faculty of Engineering, University of Porto (FEUP)

3rd Annual Meeting of the Special Interest Group on Under-resourced Languages

Lingotto Conference Centre, Turin, Italy
20-21 May, 2024

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

fct Fundação para a Ciência e a Tecnologia

INESCTEC TECHNOLOGY & SCIENCE | ASSOCIATE LABORATORY PORTUGAL