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

Abstract

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

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

1. Introduction

Text corpora play a pivotal role in advancing the development of language technology tools, especially within the realms of natural language processing (NLP) and information retrieval (IR). However, the persistent problem of constructing datasets for low-resource languages (LRLs) remains unresolved. This problem includes issues such as the lack of usable text and the existence of low-quality dataset (Kreutzer et al., 2022; Koehn et al., 2019), the absence of official writing rules and the prevalence of informal context in which texts are typically written (Linder et al., 2020), the absence of standardized annotated tokens (Strassel and Tracey, 2016), data scarcity, and the limited availability of Wikipedia document (Yu et al., 2022; Suleman, 2018). Similar problems are also faced in the case of Tetun, one of the LRLs spoken in Timor-Leste by over 932,000 speakers (de Jesus, 2023).

Several studies have explored Tetun, primarily concentrating on the influence of Portuguese loanwords in Tetun (Greksáková, 2018; van Klinken and Hajek, 2018; Hajek and van Klinken, 2019). These investigations typically employed datasets collected through face-to-face interviews, extracted from print newspapers, and derived from translated text. To the best of our knowledge, no study has systematically analyzed Tetun documents acquired from the web so far.

Given that Timor-Leste is a multilingual country with two official languages (Tetun and Portuguese), two working languages (English and Indonesian) (Vasconcelos et al., 2011), and over 30 dialects (de Jesus, 2023), this multilingual environment emphasizes the prevalence of non-standardized Tetun, particularly in its written form. Consequently, this raises questions regarding the quality of documents available on the web.

As of 2023, two multilingual datasets incorporating Tetun documents have been released and made publicly accessible on Hugging Face[1], the Wikipedia dataset (Wikimedia, 2023) and MADLAD-400 (Kudugunta et al., 2023). Despite the Tetun documents included in both resources generally exhibiting good quality, as these datasets were not audited by native Tetun speakers, certain improvements are necessary for specific IR and NLP tasks. For instance, some Tetun documents in the Wikipedia dataset still include non-Tetun content, while in the MADLAD dataset, URLs are missing, posing challenges for NLP and IR tasks that depend on access to document sources and publication dates.

To address the aforementioned challenges, we introduce Labadain-30k+ (Labadain, a Tetun word meaning spider), a Tetun text dataset comprising

[1] `https://huggingface.co`

33,550 documents (de Jesus and Nunes, 2024b). Each document is constituted of a title, URL, document source, document category, publication date, and content. Out of these 33,550 documents, 32,113 were acquired through web crawling, and an additional 1,437 were collected from the Wikipedia documents (Wikimedia, 2023). The dataset obtained via web crawling underwent a two-stage audit process: initially, content auditing was performed at the document level to extract the body text from each web page text, followed by document characterization to classify the documents into categories. For Wikipedia documents, native Tetun speakers conducted a content audit to filter out empty content and non-Tetun documents to enhance document quality.

Furthermore, the resulting dataset was utilized to conduct a comprehensive content analysis with two main objectives: i) gaining insights into the evolution of Tetun text on the web and exploring the diversity of the documents, and ii) analyzing the lexical conformity to assess the evolution of texts that adhere to the established linguistic standards, particularly in terms of vocabulary usage, while evaluating the impact of Portuguese loanwords in Tetun. To assess the lexical adherence, the dictionaries from the *Instituto Nacional de Linguística* (INL) (Correia et al., 2005) and Greksáková (2018) were employed as ground truths. The former dictionary was used to determine whether the text conforms to the Tetun INL standard, while the latter was used to validate Portuguese loanwords.

The analysis revealed that the dataset encompasses diverse documents, with news articles representing the majority at 89.86% out of 33,550 documents. Additionally, the text written following the Tetun INL standard evolved in the post-2017 periods with a +6.12 percentage-point rise, indicating the evolution of document writing on the web over time.

2. Tetun Background

Tetun, alternatively written as Tetum or Tétum, is an Austronesian language spoken in Timor-Leste, a Southeast Asian island country. Tetun comprises two major varieties: Tetun Dili or Tetun *Prasa* (referred to as Tetun) and Tetun Terik (van Klinken et al., 2002). The first known Tetun materials appeared at the end of the 19th century in the Catholic catechism written by a Portuguese priest, Sebastião Aparício da Silva (van Klinken and Hajek, 2018; Greksáková, 2018), in the era of Portuguese colonialism in Timor-Leste, which lasted from 1702 to early October 1975 (Gunn, 1999). Throughout this period, Portuguese people conducted Tetun works, and consequently, Portuguese orthography rules were directly applied to Timorese Tetun (Greksáková, 2018).

In November 1975, Timor-Leste declared its independence, but in December 1975, Indonesia invaded Timor-Leste, subsequently declaring it as its 27th province. Tetun was primarily used as a church and trade language during the Indonesian invasion era until Timor-Leste regained its independence in early September 1999.

After Timor-Leste restored its independence on May 20, 2002, the government of Timor-Leste designated Tetun as one of the country's official languages alongside Portuguese (Vasconcelos et al., 2011). Since then, it has become a dominant language in public life. In 2004, the government established the INL and produced the standard orthography of Tetun, known as "Tetun INL" (DL 01/2004, 2004).

According to the 2015 census report, Timor-Leste's population was 1.18 million, with 78.78% of the population being Tetun speakers[2] (de Jesus, 2023). Among them, 30.50% considered Tetun as their home language, while 48.28% spoke it as a second or third language. The Census 2023 reported a population growth of 13.40%, from 1.18 million to 1.34 million (INETL, 2022). However, the report did not provide specific indicators for Tetun speakers.

Moreover, online newspapers in Timor-Leste primarily use Tetun, and the launch of Tatoli[3] by the government of Timor-Leste in March 2017 (GoTL, 2020) significantly contributed to the increased popularity of online news and promoted the use of the Tetun INL writing standard. By the end of 2021, over ten online newspapers were actively publishing daily news articles in Tetun (CITL, 2024).

3. Related Work

Constructing a highly suitable dataset for various NLP and IR tasks poses significant challenges, particularly in LRL scenarios where issues arise from both the number and quality of datasets (Kreutzer et al., 2022; Linder et al., 2020; Yu et al., 2022; Koehn et al., 2019; Suleman, 2018; Strassel and Tracey, 2016). The common technique for acquiring datasets involves crawling the World Wide Web, including those specific for LRLs (Körner et al., 2022; Linder et al., 2020; Tahir and Mehmood, 2021; Wenzek et al., 2020).

[2] The total population figure from the 2015 census report referenced in de Jesus (2023) has been adjusted based on the total population data provided in both INETL (2022) and GDS (2015). However, as neither of these sources provides specific data on the total number of Tetun speakers, the reference cited in de Jesus (2023) remains the basis for estimating the proportion of Tetun speakers up to the year 2015.

[3] `https://tatoli.tl`

However, datasets for LRLs are typically derived from automatically filtered content from CommonCrawl[4] (Artetxe et al., 2022), making the task of ensuring the quality of resulting datasets challenging. As an alternative, Artetxe et al. (2022) proposed a technique involving manual identification and scraping documents from websites with high-quality content, followed by human auditing to ensure the dataset quality. The auditing process employs the "quality at a glance" technique recommended by Kreutzer et al. (2022), suggesting that a quick scan of 100 sentences can be sufficient to detect major issues in data quality.

The MADLAD-400 dataset (Kudugunta et al., 2023), a multilingual dataset released by the Google Research and Google DeepMind teams in October 2023, also includes Tetun documents. This dataset was constructed from CommonCrawl snapshots ranging from 2008 to August 2022 and underwent document-level auditing using the aforementioned "quality at a glance" approach. Since Tetun documents in the dataset were not audited by native Tetun speakers, some documents lack titles and still contain template and layout elements, such as menu names, navigation paths, links text, and more. Furthermore, Tetun documents within the MADLAD-400 dataset lack URLs, posing challenges for certain NLP and IR tasks, including issues of exclusion and bias in language technology (Bender and Friedman, 2018). Emphasizing the significance of text source information, Yu et al. (2022) incorporated this aspect into their dataset construction framework.

Other Tetun documents are incorporated in the multilingual Wikipedia dataset, introduced by the Wikimedia Foundation as of November 2023 (Wikimedia, 2023). This dataset comprises identifiers, URLs, titles, and contents. Although Tetun documents in the dataset generally exhibit good quality, there are some content issues, such as non-Tetun and incomplete text. Despite these challenges, we extracted Tetun documents from this dataset and utilized them to augment our web-crawled data as both share similar structures.

Moreover, existing literature highlights the significant influence of Portuguese on Tetun, particularly in news media, such as newspapers (Hajek and van Klinken, 2019; Greksáková, 2018; van Klinken and Hajek, 2018). van Klinken and Hajek (2018) studied a selection of seven articles from different newspapers in 2009 and stated that an average of 32% of words are Portuguese loanwords, while Greksáková (2018) reported 35% of Portuguese loanwords in the analysis of 73,892 words from interview transcripts. In a recent study, Hajek and van Klinken (2019) described Tetun's influence from Portuguese in newspaper and technical writing rising to over 40%, with headlines often almost entirely in Portuguese. In light of this, we also conducted a comprehensive analysis to understand the document writing evolution and the impact of Portuguese loanwords in Tetun.

4. Document Annotation, Auditing and Characterization

To facilitate a better representation of the data, ensure its quality, and enable its broader usage, thorough data annotation, auditing, and characterization processes are crucial. The following subsections detail the processes of annotating, auditing, and characterizing Tetun text data in the construction of the Labadain-30k+ dataset.

4.1. Overview

The Labadain-30k+ dataset is derived from a collection of Tetun documents obtained from web crawling and Wikipedia documents extracted from the multilingual Wikipedia dataset released by Wikimedia. The web-crawled data includes titles, URLs, and plain texts, encompassing elements such as text headings, subheadings, links, body texts, comments, and more. The Wikipedia documents consist of IDs, titles, URLs, and contents. The web-crawled data was collected using the Labadain Crawler, a data collection pipeline we developed for LRLs (de Jesus and Nunes, 2024a).

4.2. Document-Level Annotation

Document-level annotation was carried out by two volunteer linguists, recent graduates specializing in Tetun native language. Their primary tasks included analyzing the crawled data to identify the body contents and publication dates for each internet domain name, utilizing the document URLs.

4.2.1. Annotation Processes

The annotation process is illustrated in Figure 1. Initially, the internet domain names were automatically extracted from the URLs of the raw text data. Subsequently, these domains were employed to split documents into files for each domain. The resulting partition comprises 79 domains, with 26 containing more than 100 documents each.

Before annotating documents, annotators were instructed to analyze page structures and publication date formats for each domain, referencing the website source browsed from the URL provided in each document. Following this, annotators identified a set of potential *start* and *end* texts for each domain based on its page layout. These lists were then employed in automating the content annotation, using the algorithm detailed in Algorithm 1.

The documents obtained from the automatically annotated documents were saved in the *annotated documents* file (Figure 1). Subsequently,

[4] https://commoncrawl.org/

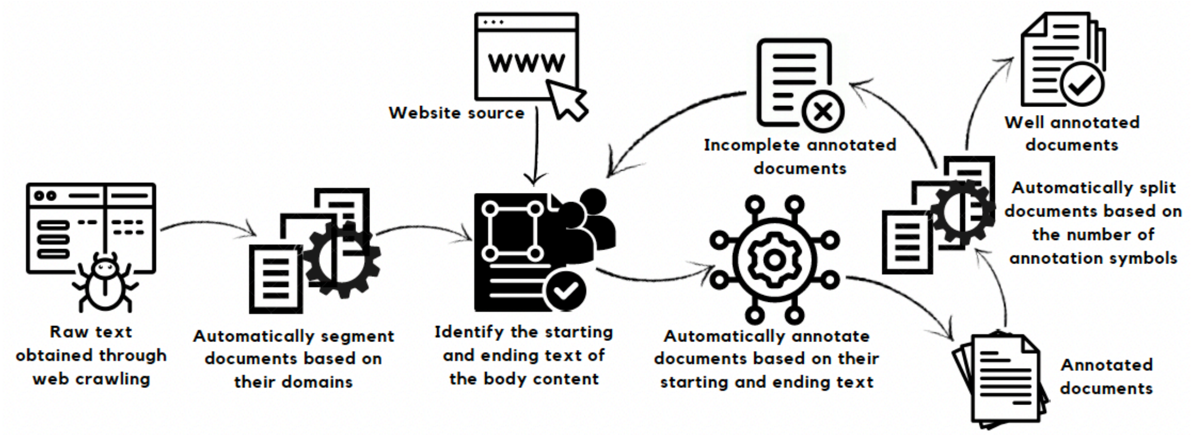
Website source
Incomplete annotated documents
Well annotated documents
Automatically split documents based on the number of annotation symbols
Raw text obtained through web crawling
Automatically segment documents based on their domains
Identify the starting and ending text of the body content
Automatically annotate documents based on their starting and ending text
Annotated documents

Figure 1: Document annotation process flow.

these documents underwent an automated verification process to verify if the documents were properly annotated. If a document contained a count of two *<t>*[5] annotations, it was considered well-annotated and then stored in the *well annotated documents* file; otherwise, it was saved in the *incomplete annotated documents* file.

The *incomplete annotated documents* were then returned to the annotators to analyze the annotation issues. The annotators updated the *start* and *end* texts, reapplied the content annotation algorithm and iterated through this process successively. Documents extracted from PDFs and presentation files underwent manual annotation, before being incorporated into the *well annotated documents* file.

4.2.2. Publication Dates Identification

The process of identifying publication dates employs two methods: first, analyzing the URLs of each internet domain name to verify whether they contain publication dates; second, browsing the website through the URLs to confirm if the page include publication dates. For documents extracted from PDF and presentation files, the dates within the files are utilized. When pages contain multiple documents with varying dates, the publication dates at the top of the page are selected and applied to all documents on that page.

For each domain, annotators provided instructions on how to access the publication dates. In cases where publication dates were not included in the URLs, additional details on date formats were also provided. This information was utilized in the configuration of publication date extraction from documents in each domain. The publication dates were formatted according to the ISO 8601 standard[6].

4.3. Content and Date Extractions

Using the *well annotated documents* as the input file, the content extraction process was automated by extracting content located between the *<t>* notations and excluding the remaining texts. For publication date extraction, if the document's URL contained the publication date, a regular expression was employed to automate the extraction process. If not, we browsed the corresponding website and inspected it to identify the CSS class tags associated with the publication date. Following the identification of these tags and the compilation of date formats for all domains, BeautifulSoup[7] was employed to automatically extract publication dates for all documents. The extracted publication dates along with the title, URL, document source, and content, were then saved in the output file.

Subsequently, an additional automated verification process was executed to ensure uniformity in date formats and structure across all documents. Incomplete information was recursively corrected and completed until all documents exhibited the same structure.

4.4. Deduplication and Post Processing

The deduplication process involved comparing document titles and corresponding URLs and excluding those with the same information. Moreover, any repeated occurrences of document titles within the content were also removed.

To improve the quality of document titles, in the case where the document source names were included in the titles, this information was manu-

[5] Note that *<t>* symbol was a preference notation chosen by the authors and can be replaced with any annotation symbol as preferred.

[6] https://www.iso.org/iso-8601-date-and-time-format.html

[7] https://www.crummy.com/software/BeautifulSoup/

ally removed using the find and replace function. For instance, “| Notísia Timor News” was eliminated from the document title [Povu mak sei Hili | **Notísia Timor News**].

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%

Table 1: Summary of the web-crawled dataset.

The resulting dataset consists of 32,113 documents, each comprising a title, URL, document source, publication date, and content. A summary of the web-crawled dataset is provided in Table 1.

4.5. Document Characterization

The document characterization task was carried out by three native Tetun speakers, who are students, and following the established guidelines. The subset of the dataset selected for the categorization task, refer to the highlighted rows in Table 1, comprises 2,879 documents sourced from non-governmental, governmental, education, and bank and court portals. These documents were chosen for their diverse content representations.

After conducting an overall preliminary analysis of the aforementioned documents, a total of seven categories were identified, which were then incorporated into the guidelines. These categories comprise news articles, legal and governmental documents, technical documents, correspondence letters, research papers, institutional information, and advertisements and announcements.

4.5.1. Annotation Processes

As the initial step of the document characterization process, annotators were instructed to read the guidelines to comprehend the task requirements. Following this, annotators were directed to familiarize themselves with the predefined categories by comparing examples of documents within each category in the guidelines.

Subsequently, a training session was provided to demonstrate practical annotation examples. After this session, annotators conducted three pilot testing sessions, each assessing ten documents. In each session, after completing the characterization, annotators compared their results and discussed the challenges encountered, suggesting improvements, and incorporating feedback to enhance the document characterization accuracy.

Finally, each annotator conducted a characterization of the 2,897 documents. The characterization task was carried out within two days, corresponding to approximately 16 hours, with an average characterization time of 20 seconds per document.

4.5.2. Inter-Annotators Agreement

To assess the reliability of inter-annotator agreement, we employed Fleiss’ Kappa measure (Fleiss, 1971), and the strength of the agreement was interpreted using the interpretation table provided by Landis and Koch (1977).

The evaluation resulted in a k value of 0.4994, indicating moderate agreement among the annotators. Subsequently, the annotators discussed their discrepancies and finally reached a consensus agreement for all documents. Documents based on this consensus encompass 1,223 legal and government documents, 1,153 news articles, 211 technical documents, 124 advertisements and announcements, 83 research papers, 53 institutional information documents, and 32 correspondence letters.

4.6. Wikipedia Documents Processing

To augment the existing crawled data, we leveraged the Tetun documents from the multilingual Wikipedia dataset available on Hugging Face. The process of extracting Tetun documents followed the documentation provided with the dataset. The extracted dataset contains 1,468 documents, consisting of ID, URL, title, and content.

To maintain uniformity with the structure of the aforementioned crawled data, we applied the same approaches outlined in subsection 4.3 to generate document sources and extract publication dates. Additionally, the document contents were organized in accordance with the crawled data format, where each document was separated by two consecutive newlines. We preprocessed documents by removing HTML tags that existed in some documents and excluding the document identification (ID) from the dataset. Afterward, we distributed these documents to the aforementioned three students for content audit, with each responsible for approximately 500 documents.

After thoroughly examining the document contents, a total of 13 documents were identified with empty content or content not written in Tetun. Some additional content issues, such as a mix of Tetun with Indonesian and English languages, were also reported. Nevertheless, as these texts were removed from the content during the auditing process, the final set of 1,455 documents is composed of clean documents.

4.7. Final Dataset

To compile the final dataset, we combined 29,234 documents that were not characterized, referring to non-highlighted rows in Table 1, with the 2,897 consensus documents described in subsubsection 4.5.2, and the 1,455 Wikipedia documents detailed in subsection 4.6.

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%

Table 2: Summary of the final dataset.

We identified 18 duplicate documents in the Wikipedia set, conducted deduplication, and ended up with a total of 1,437 unique documents from the Wikimedia dataset. These documents were merged with the 32,113 documents outlined in subsection 4.4, resulting in the final dataset comprising 33,550 documents (called **Labadain-30k+**). Each document includes metadata such as title, URL, document source, document category, publication date, and content. A summary of the final dataset is detailed in Table 2.

5. Comprehensive Content Analysis

This section provides a comprehensive content analysis to understand the composition and evolution of the dataset on the web, assess the evolution of Tetun documents written, and analyze the impact of Portuguese loanwords in Tetun.

The following terms are employed in this analysis: i) Document: A dataset unit consisting of a title, URL, source, publication date, and content. ii) Title: The document title. iii) Content: The body text of the document. iv) Corpus: A combination of document titles and contents. v) Paragraph: Each segment of text separated by a single newline in the document's content. vi) Sentence: Each line of text ending with a period (.), exclamation mark (!), or question mark (?). Periods within titles, such as Dr., Ph.D., etc., are not sentence endings. vii) Token: A text unit comprising a word or number, excluding punctuation and special characters. viii) Vocabulary: A set of unique tokens.

5.1. Dataset Description and Distribution

Table 3 summarizes a quantitative overview of the composition and characteristics of the dataset and Table 4 provides details information on the number of documents, paragraphs, sentences, individual text units, and unique tokens.

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

Table 3: Labadain-30k+ dataset description.

	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

Table 4: Summary of documents.

To identify the main contributors to the dataset and their origins, we grouped the documents by their sources. The results show that the top 5 contributors, in terms of quantity, predominantly originate from online newspapers (Table 5). Notably, Tatoli, the public online news agency in the country, emerges as the leading contributor, accounting for 27.19% of documents in the dataset.

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%

Table 5: Top five sources by document count.

To provide an overview of the dataset's composition, we grouped the distribution of documents based on their top-level domains (TLDs), as shown in Table 6. The ".com" domain notably predominates, while ".tl," representing Timor-Leste, holds the second position.

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%

Table 6: Summary of dataset per TLDs.

In the analysis of word frequency distribution within the corpus, we generated a plot to assess its adherence with Zipf's law (Zipf, 1949). Figure 2 illustrates the relationship between a word's rank and its frequency, confirming the characteristic pattern associated with Zipf's law. This pattern is characterized by an inverse proportionality between a word's frequency and its rank, a key feature indicative of Zipfian distribution. The most common words in the corpus, excluding stopwords, highlight prevalent terms such as "governu" (government), "timor-leste," "dili" (capital of Timor-Leste), among others.

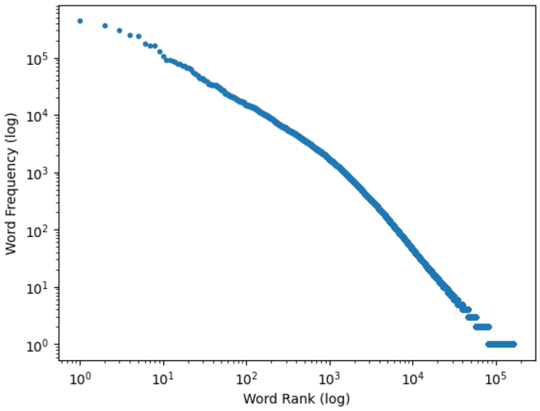

Figure 2: Word frequency vs. Word rank.

Furthermore, we analyzed co-occurring word sequences, explaining bigrams and trigrams as representative samples. The analysis of bigrams highlighted the prominence of pairs such as "covid 19" and "prezidente repúblika" (president of the republic). Shift to 3-grams, observed patterns such as "taur matan ruak" (name of the former prime minister of Timor-Leste) and "guterres lú olo" (name of the former president of Timor-Leste), emerged as the most frequent trigrams. Collectively, these n-gram words provide insights into the prevalence of specific terms within the dataset.

5.2. Document Evolution on the Web

The Labadain-30k+ dataset comprises documents spanning from 2001 to 2023, excluding the years 2004 and 2005 for which no documents are available. The absence of documents from 2004 and 2005 in the dataset may be attributed to various factors, including language barriers and limited digital archiving endeavors due to constraints in internet infrastructure. Furthermore, the dataset contains fewer than 100 documents for years preceding 2010, indicating similar challenges.

Starting in 2017, there was a substantial increase in document quantity (Table 7), corresponding to the increasing popularity of online news. This surge can be attributed to the launch of Tatoli in March 2017. Nevertheless, it was only from 2020 and onwards trajectory that a notable increase in document quantity on the web occurred, and the trend persisted, with document numbers continuing to rise until 2023.

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

Table 7: Evolution of document quantity over the years. $^{+}$Percentage point.

In the assessment of document writing evolution, we focused on evaluating the lexical adherence of Tetun text with the Tetun INL standard and the impact of Portuguese loanwords in Tetun. The evaluation grounded on the INL's dictionary to assess the evolution of Tetun text and Greksakova's dictionary to verify the presence of Portuguese loanwords. With the significant increase in web document quantity since 2017, we chose this year as the threshold for comparing Tetun text evolution and loanwords influence before and after 2017.

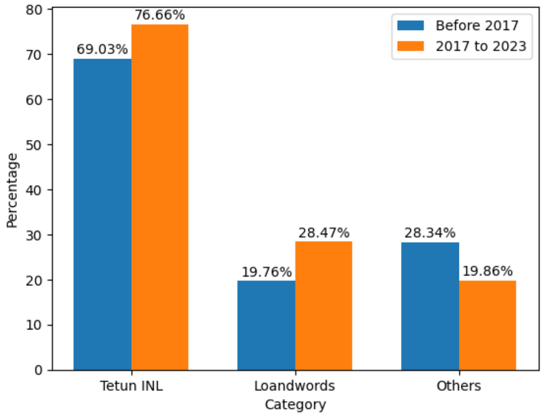

Figure 3: Evolution of document writing and loanword impact in news content pre- and post-2017.

The analysis revealed a substantial improvement in the use of Tetun INL in document writing, alongside the persistent use of Portuguese loanwords (see details in Table 8). There was also a notable decrease in words not found in the dictionary, encompassing terms such as misspelled and out-of-vocabulary (OOV).

Moreover, considering the predominance of news articles, we conducted a focused analysis on document writing evolution and the impact of Portuguese loanwords within this category. The findings indicated a higher use of Tetun INL and loanwords compared to the overall results (Figure 3).

6. Analysis on the Results

The Labadain-30k+ dataset comprises documents from a variety of sources (Table 1) and across multiple categories (Table 2). Although news articles are predominant (Table 5), substantial contributions also come from the Wikipedia and legal/government categories, along with lower contributions from seven other categories, each containing less than 300 documents. Analyzing the documents' origin based on TLDs, the majority originate from ".com," closely followed by ".tl," with a margin of 2.56 percentage points (Table 6).

From a linguistic perspective, the distribution of word frequencies in the Labadain-30k+ dataset adheres to Zipf's law, emphasizing the concept that a small number of words occur frequently, while the majority exhibit lower frequencies (Figure 2). Furthermore, the analysis of up to 5-gram words, excluding stopwords, suggests a substantial portion of the documents focus on the Covid-19 pandemic, events taking place in Dili, and topics related to the country and its government.

Regarding the evolution of document quantity on the web, a consistent increase has been observed since 2014. However, a notable surge occurred in 2020, marking an 11.75 percentage point rise compared to 2019 (Table 7). This upward trend persisted, with document numbers continuing to rise until 2023. However, since the data crawled only covers up to September 30, 2023, there has been a decrease of 3.79 percentage points in 2023 compared to the data from 2022. The evolution of document writing, assessed against the Tetun INL standard with a focus on vocabulary use, demonstrated a 6.12 percentage point improvement in Tetun INL standard usage from 2017 onwards compared to previous years. Additionally, the persistence of Portuguese loanwords remained evident, indicating an increase of 5.09 percentage points from 2017 onwards (Table 8).

7. Discussions

The Labadain-30k+ dataset showcases a diverse document composition collected from various sources and categories, emphasizing its richness in document variety. This diversity underscores the dataset's versatility, making it highly suitable for various NLP and IR tasks. Table 9 compares the Labadain-30k+ dataset size and the number of speakers with other LRLs. Tetun, Occitan, and Mizo have similar dataset sizes available on the web and indicate a comparable number of speakers. Despite Tetun having fewer speakers, its dataset size is comparable to that of Assamese and Swiss German.

Considering a substantial increase in the document quantity from 2020 onwards and the emergence of "covid 19" as the most frequent word pair, there is a noticeable correlation between the Covid-19 pandemic and the increase of Tetun documents on the web. With Approximately 90% of the documents being news articles, showcasing a substantial improvement in the use of Tetun INL standard in document writing within this category since 2017 (Figure 3), surpassing the overall improvement by 1.51 percentage points. Also, the occurrence of Portuguese loanwords in news articles exceeds the overall result by 3.62 percentage points. This evidence underscores the pivotal role of online news contributions in promoting the use of Tetun INL standard in document writing.

Since the existing literature reported a five percentage points increase in the prevalence of Portuguese loanwords in Tetun newspapers, rising from 35% to 40% between 2018 and 2019 (Greksáková, 2018; Hajek and van Klinken, 2019), where certain news titles were predominantly composed of Portuguese loanwords, we conducted a comparative analysis using news article titles from the same periods. Our findings revealed a similar trend but with a modest increase of 3.5 percentage points and a lower overall percentage of Portuguese loanwords: 30.01% in 2018 and 33.51% in 2019. While acknowledging that the variation may be attributed to differences in datasets, a comparable finding emerges regarding the upward trend of Portuguese loanwords in newspapers.

Table 8 shows that a total of 20.23% of words not found in dictionaries, categorized as misspelled, out-of-vocabulary (OOV), or from other languages used to represent specific terms and named entities. We analyzed the top 10 most frequent words in this category and identified words such as "hanesan" (such as), "Timor-Leste," "hetan" (get), PNTL (National Police of Timor-leste), and Covid as OOV words. This indicates that those words are not included in the dictionary entries, highlighting a limitation in the Tetun INL dictionary.

8. Conclusions and Future Work

This paper presents Labadain-30k+, the first Tetun dataset audited by native Tetun speakers, encompassing 33.6k documents enriched with metadata, including URLs, document sources, publication dates, categories, and contents. Comparable in size to Tetun documents in MADLAD-400,

	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

Table 8: Evolution in the use of Tetun INL in document writing before and after 2017. $^{+}$Numbers are excluded from the count. *Certain loanwords are also present in the Tetun INL dictionary.

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]

Table 9: Comparison of the Labadain-30k+'s dataset size and total number of speakers with other LRLs. [1]Kudugunta et al. (2023). [2]Britannica (2024). [3]Posner and Sala (2024). [4]UNESCO (2024). [5]Switzerland (2024).

Labadain-30k+ contains approximately 6.8k documents fewer, yet offers more contextual information for each document, enhancing its utility for various NLP and IR tasks.

Moreover, this paper outlines methodologies for document annotations and characterizations, and assessments of the evolution of Tetun text documents and Portuguese loanwords in Tetun. These approaches can be leveraged in constructing and analyzing textual data for other LRLs facing similar challenges.

In future work, we plan to utilize Labadain-30k+ to create a test collection for evaluating information retrieval tasks and explore its potential application in Tetun text classification.

9. Acknowledgement

This work is financed by national funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia (de Jesus, 2021).

10. Bibliographical References

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de-Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7383–7390. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Trans. Assoc. Comput. Linguistics*, 6:587–604.

The Editors of Encyclopaedia Britannica. 2024. Assamese language. Accessed on February 19, 2024.

Press Council of Timor-Leste CITL. 2024. The registered and licensed social communication agencies in timor-leste. Accessed on January 5, 2024.

Adérito José Guterres Correia, Geoffrey Stephen Hull, Geoge William Saunders, and Domingos dos Santos Rosa da Costa Tilman, Mário Adriano Soares. 2005. *Disionáriu Nasionál ba Tetun Ofisiál*. Instituto Nacional de Linguística, Universidade Nacional Timor Lorosa'e, Avenida Cidade de Lisboa, Dili, Timor-Leste.

W. Bruce Croft, Donald Metzler, and Trevor Strohman. 2009. *Search Engines - Information Retrieval in Practice*. Pearson Education.

Gabriel de Jesus. 2021. Pesquisa e recomendação computacional de couteúdo noticioso. Bolsa de Investigação na área de Engenharia Informática. Fundação para a Ciência e a Tecnologia (FCT), Portugal.

Gabriel de Jesus. 2023. Text information retrieval in tetun. In *Advances 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.

Gabriel de Jesus and Sérgio Nunes. 2024a. Data collection pipeline for low-resource languages: A case study on constructing a tetun text corpus. In *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Lingotto Conference Centre - Torino (Italia). Zenodo.

Gabriel de Jesus and Sérgio Nunes. 2024b. Labadain-30+: A monolingual tetun document-level audited dataset [data set]. INESC TEC. https://doi.org/10.25747/YDWR-N696.

Democratic Republic of Timor-Leste DL 01/2004, Government Decree-Law No. 1/2004 of 14 April. 2004. The standard orthography of the tetun language. Accessed on September 21, 2023.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382. Measures of inter-rater reliability for two or more rates (annotators).

The General Directorate of Statistics of the Ministry of Finance GDS. 2015. Timor-leste population and housing census 2015: Analytical report on agriculture and fisheries (volume 2). Accessed on February 19, 2024.

Government of Timor-Leste GoTL. 2020. Tatoli completes four years of existence. Accessed on January 5, 2024.

Zuzana Greksáková. 2018. *Tetun in Timor-Leste: The role of language contact in its development*. Ph.D. thesis, Universidade de Coimbra, Portugal.

Geoffrey C. Gunn. 1999. *Timor Loro Sae: 500 years*. Livros do Oriente.

John Hajek and Catharina Williams van Klinken. 2019. Language contact and gender in tetun dili: What happens when austronesian meets romance? *Oceanic Linguistics*, 58:59–91.

Instituto Nacional de Estatística Timor-Leste INETL. 2022. Timor-leste population and housing census. Accessed on February 19, 2024.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Miguel Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 3: Shared Task Papers, Day 2*, pages 54–72. Association for Computational Linguistics.

Erik Körner, Felix Helfer, Christopher Schröder, Thomas Eckart, and Dirk Goldhahn. 2022. Crawling under-resourced languages – a portal for community-contributed corpus collection. In *Proceedings of the 1st Workshop on Dataset Creation for Lower-Resourced Languages (DCLRL) @LREC2022, Marseille, 24 June 2022*. European Language Resources Association (ELRA).

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Balli, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Trans. Assoc. Comput. Linguistics*, 10:50–72.

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.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74. The reference contains interpretation of k-value of inter-annotators.The interpreptation is only for two annotators and two class. It is used in interpreting Fleiss' Kappa.

Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat, and Andreas Fischer. 2020. Automatic creation of text corpora for low-resource languages from the internet: The case of swiss german. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 2706–2711. European Language Resources Association.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England.

Rebecca Posner and Marius Sala. 2024. Occitan language. Accessed on February 19, 2024.

Stephanie M. Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Hussein Suleman. 2018. Information retrieval in african languages. *CoRR*, abs/1806.04735.

About Switzerland. 2024. Language – facts and figures. Accessed on February 19, 2024.

Bilal Tahir and Muhammad Amir Mehmood. 2021. Corpulyzer: A novel framework for building low resource language corpora. *IEEE Access*, 9:8546–8563.

UNESCO. 2024. World atlas of languages. Accessed on February 19, 2024.

Catharina Williams van Klinken and John Hajek. 2018. Language contact and functional expansion in tetun dili: The evolution of a new press register. *Multilingua*, 37:613 – 647.

Catharina Williams van Klinken, John Hajek, and Rachel Nordlinger. 2002. *Tetun Dili: a grammar of an East Timorese language*. Pacific Linguistics, Canberra, Australia.

Pedro Carlos Bacelar de Vasconcelos, Andreia Sofia Pinto Oliveira, Ricardo Sousa da Cunha, Andreia Rute da Silva Baptista, Alexandre Corte-Real de Araújo, Benedita McCrorie Graça Moura, Bernardo Almeida, Cláudio Ximenes, Fernando Conde Monteiro, Henrique Curado, et al. 2011. Constituição anotada da república democrática de timor-leste.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

The Foundation of Wikimedia. 2023. Wikimedia downloads.

Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. 2022. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3725–3743. Association for Computational Linguistics.

George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press.

Appendix A. Content Annotation Algorithm

The Content Annotation Algorithm is presented in Algorithm 1.

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
```