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

Gabriel de Jesus, Sérgio Nunes
INESC TEC and Faculty of Engineering of the University of Porto (FEUP)
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
gabriel.jesus@inesctec.pt, sergio.nunes@fe.up.pt

Abstract

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

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

1. Introduction

Constructing text corpora for low-resource languages (LRLs) to facilitate the development of natural language processing (NLP) and information retrieval (IR) tools has gained notable attention in recent year (Aji et al., 2022; Hedderich et al., 2021; Kusampudi et al., 2021; Magueresse et al., 2020). Despite the scarcity of resources, there has been a consistent rise of interest in this domain (Blaschke et al., 2023). Various solutions for constructing text corpora for LRLs have been proposed, ranging from employing the BootCat toolkit (Baroni and Bernardini, 2004) to more recent approaches such as manually scraping documents from websites containing high-quality data (Artetxe et al., 2022), employing crawling and collaborative techniques (Körner et al., 2022), extracting data from Common Crawl (Wenzek et al., 2020a), crawling by modifying the existing Nutch framework (Tahir and Mehmood, 2021), or developing a data collection pipeline from scratch (Linder et al., 2020). Moreover, challenges associated with the corpora construction for LRLs have also been studied, including addressing the quality of web crawler datasets (Kreutzer et al., 2022; Abadji et al., 2022), and language-specific issues (Aji et al., 2022).

However, the aforementioned techniques come with inherent constraints, such as the availability of data sources containing high-quality data, limited technical documentations, pipeline accessibility, and financial requirements. Consequently, implementing these approaches for a new LRL presents significant challenges. As an alternative solution, we propose Labadain Crawler (Labadain, a Tetun word meaning spider), a data collection pipeline designed to automate and optimize the process of constructing textual corpora from the web for LRLs.

Our proposed solutions consist of crawling, processing, and summarizing textual data from the web in a systematic and automated way, enabling the construction of a high-quality text corpus for LRLs. By leveraging Apache Nutch[1], an open-source web crawler and data extraction framework, our pipeline establishes a robust foundation for effective web crawling and data extraction. The automation of Labadain Crawler relies on three key components: an initial text corpus containing the target language, a tokenizer, and a language identification (LID) model. These components are crucial for generating the seed words and URLs required for Nutch to initiate the crawling process, and for extracting text in the target language from the crawled data to construct a text corpus.

The proposed solution was successfully tested with Tetun and effectively addressing the challenges associated with the construction of a Tetun text corpus. Since the language processing components for Tetun such as tokenizer and LID model did not exist, we initially developed these components, and then integrated them into the pipeline to enable the process of constructing a Tetun text corpus. All code is accessible at https://github.com/gabriel-de-jesus/labadain-crawler.

2. Related Works

In recent works, various techniques have been explored to address the challenges of constructing datasets for LRLs. Artetxe et al. (2022) proposed a tailored crawling, where this approach involved

[1] https://nutch.apache.org

LREC-COLING 2024, pages 4368–4380
20-25 May, 2024.

manually identifying data sources containing high-quality data and subsequently scraping their contents to construct a text corpus. This approach was implemented for Basque and resulted in the construction of a text corpus comprising 12.5 million documents containing 423 million tokens.

Körner et al. (2022) introduced Crawling Under-Resourced Languages (CURL), a data collection pipeline that used the Heritrix crawler (Mohr et al., 2004) and incorporated Brown's LID model (Brown, 2013). This pipeline was designed to facilitate community participation and corpus creation through a web portal. The implementation of CURL has resulted in the creation of corpora for 258 LRLs, and the Iranian Persian language has the largest corpus with over 3.9 million sentences.

Tahir and Mehmood (2021) proposed Corpulyzer, a customized crawling framework built by modifying the Apache Nutch framework and incorporating the LID model from Compact Language Detector 2[2]. This framework was applied to Urdu, resulting in the construction of UrduWeb20, a dataset consisting of 4.1 billion tokens extracted from eight million web pages crawled from 6,590 URLs.

Wenzek et al. (2020a) presented a pipeline for downloading and processing one snapshot of Common Crawl, employing sentence piece tokenizer (Kudo, 2018), the LID classifier from fastText (Grave et al., 2018), and computed the perplexity using a language model trained on a targeted domain as the documents quality score. This pipeline produced a monolingual text dataset by collecting 3.2 TB of compressed documents in 174 languages, including some LRLs such as Basque, comprising roughly 10.36 million sentences, and Malay, consisting of approximately 6.96 million sentences.

Linder et al. (2020) proposed SwissCrawl, a crawling tool developed from scratch to collect documents from the web and construct a Swiss German corpus. Moses' split sentences[3] was employed to split text documents into sentences and then filtered them using the LID model they developed. The resulting corpus comprised 562,524 sentences, gathered from a total of 62,000 URLs across 3,472 domains.

Tailored crawling is an effective technique for collecting a large amount of high-quality data, but it requires data sources that contain such content. It is particularly favorable for LRLs which have readily available data of adequate quality. The CURL project, which employs crawling and collaborative techniques, is advantageous for long-term data collection projects, but motivating the community to contribute content is challenging. von Holy et al. (2017) used a gamification technique with a rewarding strategy to increase community participation in contributing content to their dataset, but this approach requires a budget for rewards.

Constructing a dataset using publicly available snapshots from Common Crawl presents computational challenges since it requires adequate computing power to process the snapshots. The SwissCrawl presented limited availability of documentation and resources associated with the tool, and the Corpulyzer framework is not accessible.

As an alternative solution, we introduced the Labadain Crawler to address the challenges of constructing a text corpus for LRLs. To showcase its effectiveness, we tested it with Tetun and successfully constructed a Tetun text corpus well-suited for various NLP and IR tasks.

3. Tetun

Tetun is the language spoken in Timor-Leste, an island country in Southeast Asia. It was a dialect used as both church and trade language during the colonial era until Timor-Leste restored its independence and became a new sovereign state on May 20, 2002. In 2002, the Government of Timor-Leste designated Tetun as one of the country's official languages alongside Portuguese (Vasconcelos et al., 2011), leading to its widespread usage in public life. Tetun is a LRLs spoken by 78.78% of a 1.18 million populations (de Jesus, 2023)[4]. There are two major varieties of Tetun: Tetun Dili (commonly referred to as Tetun) and Tetun Terik (van Klinken et al., 2002). Tetun Dili encompasses Tetun *Instituto Nacional de Linguística* (Tetun INL) and Tetun Dili Institute of Technology (Tetun DIT), while Tetun Terik is one of Timor-Leste's dialects.

Tetun INL is a dialect for which the government of Timor-Leste has established a standard orthography through the INL, subsequently adopted as the official Tetun being used in the education system, official publications, and media (DL 01/2004, 2004). Tetun DIT was developed by linguists at DIT with a few standardized differences from Tetun INL in terms of writing conventions (van Klinken et al., 2002). For example, Tetun INL uses "ll" (e.g., millaun, meaning million), while Tetun DIT utilizes "lh" (e.g., milhaun).

Tetun INL is based on the Latin alphabet, distributed in 5 vowels: *a, e, i, o, u*, and 21 consonants: *b, d, f, g, h, j, k, l, ll, m, n, ñ, p, r, rr, s, t, u, v, x, z* (INL, 2004). The letters *C, Q, W*, and *Y* are

[2] https://github.com/CLD2Owners/cld2

[3] https://github.com/moses-smt/mosesdecoder

[4] 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.

not used in Tetun INL, except for proper names and international symbols. The accented vowels *á, é, í, ó, ú*, are also used, and the apostrophe (') denotes a glottal stop. Additionally, the hyphen is also introduced to indicate mono-semantic compound words.

Given that Tetun INL is the standardized Tetun established by the Government of Timor-Leste for use in the education system, official publications, and media, a tokenizer for Tetun was developed based on this standardization.

4. Tetun Tokenizer

Tokenization is an essential preprocessing technique in NLP and IR, segmenting texts into individual tokens to facilitate tasks related to text processing and analysis. We introduce a set of rule-based tokenization techniques for Tetun, where some techniques are devised specifically for Tetun based on the Tetun INL standard, and others are derived from commonly used tokenization rules.

Due to the unique language-specific characteristics of Tetun, directly applying existing tokenizers to Tetun words is not feasible. For instance, when applying tokenizers of English, Portuguese, Spanish, and French in NLTK[5] on the Tetun input sentence "ha'u-nia uma mak ne'e (this is my home)", these tokenizers incorrectly split the words "ha'u" and "ne'e" to ["ha", "'", "u"] and ["ne", "'", "e"], respectively. Therefore, a specialized tokenizer for Tetun was developed to tackle these tokenization challenges.

4.1. Characteristics

A Tetun tokenizer was developed using regular expressions, encompassing the following language-specific features: (1) A token may include accented vowels and apostrophes, such as "ne'ebé" and "ne'ebá". (2) An accented vowel can occur at the beginning, in the middle, or at the end of a token, for example "área", "líder", and "oinsá". (3) The accented consonant "ñ" typically appears within or in the middle of tokens, as in "kompañia", and "kampaíña". (4) Mono-semantic compound words can contain a combination of hyphens, accented vowels and apostrophes, such as "sanuluresin-ida", "ida-ne'ebá" and "ida-ne'e". (5) Person names may include Portuguese-based accentuation marks such as tilde (~), caret (^), and cedilla (ç), for instance "João" and "Conceição".

The Tetun Word Tokenizer and the Tetun Simple Tokenizer, as detailed in subsection 4.2, are tokenization techniques tailored for Tetun, aimed at accommodating its distinctive linguistic features.

4.2. Techniques

The Tetun tokenizer[6] incorporates several rule-based techniques, which are outlined below. The Tetun Word Tokenizer and the Tetun Simple Tokenizer techniques are designed specifically for Tetun, while the remaining techniques are adaptations of widely employed approaches.

1. **Word Tokenizer:** Extracting only word units from the input text and excluding numbers, punctuation, and special characters.
2. **Simple Tokenizer:** Extracting only words and numbers from the input text while discarding punctuation and special characters.
3. **Standard Tokenizer:** Segmenting the input text into tokens based on word boundaries, punctuation, and special characters.
4. **Sentence Tokenizer:** Splitting sentences by their ending delimiters such as period (.), question mark (?), and exclamation mark (!). The period used to represent titles, such as Dr., P.hD., etc., is preserved.
5. **Blank Line Tokenizer:** Segmenting the input text based on the presence of blank lines.

In this work, we exclusively utilized the Tetun Word Tokenizer, the Tetun Sentence Tokenizer, and the Tetun Simple Tokenizer, while the other tokenization techniques have been integrated into the Tetun tokenizer for potential future applications. The Tetun Word Tokenizer was employed in constructing the Labadain Crawler (referred to as tokenizer in Figure 3). Both the Tetun Word and the Tetun Sentence Tokenizers were applied in developing the language identification model described in section 5. The Tetun Simple Tokenizer was utilized in generating the summary of the corpus contents presented in subsection 6.3.

4.3. Evaluation and Result

The Tetun tokenizer was evaluated through a controlled experiment involving five Timorese volunteer students, all native Tetun speakers. The evaluation aimed at assessing the effectiveness of the tokenization techniques across different Tetun text samples sourced from the internet, such as Wikipedia, reports, news articles, and more. To ensure the reliability of the evaluation, the students followed a standard guideline.

Initially, each student was tasked with evaluating each tokenizer technique and collecting a minimum of three input texts in Tetun from various internet domains, ensuring a diverse range of linguistic characteristics and structural complexities.

[5] https://www.nltk.org

[6] https://pypi.org/project/tetun-tokenizer/

Subsequently, the student annotated the collected texts, creating tokens that served as the gold standard for evaluating the tokenizers' effectiveness. Finally, the evaluation module codes were distributed and configured on the student's laptops to assess the tokenizers' efficacy. The students provided the sample text and ground truth files as inputs and executed the evaluation module codes to observe the resulting outcomes.

The evaluators reported that all Tetun tokenizer techniques achieved 100% accuracy for each tested input text, comprising 200 to 500 tokens extracted from diverse internet domains. While the tokenizer was primarily developed based on the established Tetun INL standard, its broader applicability to Tetun as a whole was also assessed. Therefore, the Tetun tokenizer underwent testing with both Tetun INL and randomly collected Tetun texts from the web. These experiments ensured the tokenizer's general applicability to Tetun and further validated the effectiveness of the tokenization techniques outlined in subsection 4.2, thus providing additional validation for its usage.

5. Tetun Language Identification

Language identification (LID) is a pivotal component in constructing linguistic text corpora for LRLs (Jauhiainen et al., 2019). Timor-Leste, characterized by its multilingualism, comprises two official languages (Tetun and Portuguese), two working languages (Indonesian and English) (Vasconcelos et al., 2011), and over 30 dialects (de Jesus, 2023) spoken across the territory. This multilingual environment shapes diverse writing forms in Tetun, resulting in the prevalence of mixed digital text documents available on the web. Given the complexity of this linguistic landscape, developing a robust LID model for Tetun becomes a crucial need.

5.1. Data Preparation

The LID model was specifically trained to classify four languages spoken in Timor-Leste: Tetun, Portuguese, English, and Indonesian. The Tetun dataset was collected from Timor News[7], an online news agency based in Dili, Timor-Leste, while for Portuguese, English, and Indonesian, the CC-100 dataset[8] (Wenzek et al., 2020a) was utilized. Access to the Tetun dataset was unrestricted as it was obtained from a platform associated with the agency founded by the main author of this paper.

To efficiently manage the vast amount of data collected from the CC-100 dataset, a subset of equivalent size to the Tetun dataset was chosen through random sampling of texts in Portuguese, English, and Indonesian. These datasets were merged with the Tetun dataset and then tokenized into sentences, each associated with its respective language. The resultant dataset underwent preprocessing steps, including lower-casing and removal of punctuation, special symbols, and digits. A summary of the dataset is presented in Table 1.

5.2. Language Similarities Analysis

In the analysis of similarities among Tetun, Portuguese, English, and Indonesian, the dataset was first visualized using a Gaussian distribution plot. Next, we explored the interquartile range and standard deviation across various parameter values ranging from 1 to 3. The analysis revealed that a standard deviation of 1 provided the most suitable normalization solution for all input languages. Therefore, this setting was adopted to select the dataset that conformed to the normal distribution range, which was then utilized in the clustering.

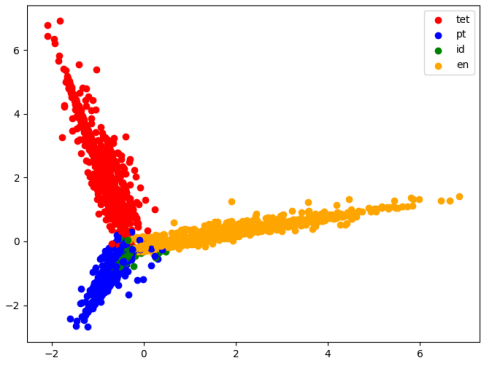

Figure 1: Gaussian-Mixture clustering algorithm applied to cluster Tetun (tet), Portuguese (pt), English (en) and Indonesian (id).

During the clustering process, the input sentences were tokenized into word tokens using Tetun Word Tokenizer and subsequently transformed the resulting tokens into numeric array vectors. These vectors were then processed using Principal Component Analysis (PCA) and followed by normalization of resulting features. Finally, Gaussian Mixture was employed to generate language clusters. This clustering technique was selected for its effectiveness in the application to LRLs scenario (Dovbnia et al., 2022). The resulting clusters were plotted into two-dimensional space (Figure 1), indicating significant dissimilarities among the four input languages.

5.3. LID Model Training

The LID model was trained using the dataset summarized in Table 1. Each sentence served as input features, while the target variable represented the

[7] https://www.timornews.tl
[8] https://data.statmt.org/cc-100/

	Tetun	Portuguese	English	Indonesian
Total sentences	18,108	29,056	34,509	31,888
Minimum words per sentence	2	1	1	1
Maximum words per sentence	209	1,122	220	1,746
Average words per sentence	29.12	20.89	18.99	16.47
Total words in document	527,258	606,867	655,328	525,298

Table 1: Summary of the dataset used for training the LID model.

respective language. The dataset was split into three subsets: 70% for training, 15% for development, and 15% for testing.

The experiments was conducted using three commonly employed machine learning models that have proven their effectiveness in the LID task (Jauhiainen et al., 2019; Espichán-Linares and Oncevay-Marcos, 2017): Support Vector Machine (SVM) with a linear kernel, Logistic Regression (LR), and Multinomial Naive Bayes (MNB). Throughout the training process, we assessed the performance of these models by evaluating accuracy at both the character n-gram and the word n-gram levels using the development set. At the character n-gram level, each character in the input text was treated as feature, whereas at the word n-gram level, each word served as feature. The training results are detailed in Table 2.

5.4. Model Evaluation and Result

The performance analysis presented in Table 2 demonstrates that the MNB model with 5-gram characters achieved the highest accuracy. Therefore, this particular model was selected for training and evaluating the LID model. Upon evaluation on the test set, it achieved an impressive overall accuracy of 99.77%. The corresponding F1 scores were 99.87% for Tetun, 99.84% for Portuguese, 99.79% for Indonesian, and 99.76% for English. The detailed evaluation results are illustrated in the confusion matrix in Figure 2.

The performance of the Tetun LID model aligns closely with findings highlighted in the survey conducted by Jauhiainen et al. (2019). Notably, employing character n-grams as the primary feature in training LID models has been effective, with 5-grams frequently yielding the highest accuracy among different n-gram sizes. Moreover, Zampieri (2013) demonstrated that MNB outperformed SVM in training LID models, particularly in classifying several closely related languages, including European and Brazilian Portuguese.

6. Labadain Crawler

The Labadain Crawler relies on three fundamental components: an initial text corpus containing the target language, a tokenizer, and a LID model. The following subsections outline its architecture, its implementation with Tetun, and the evaluation of the resulting text corpus.

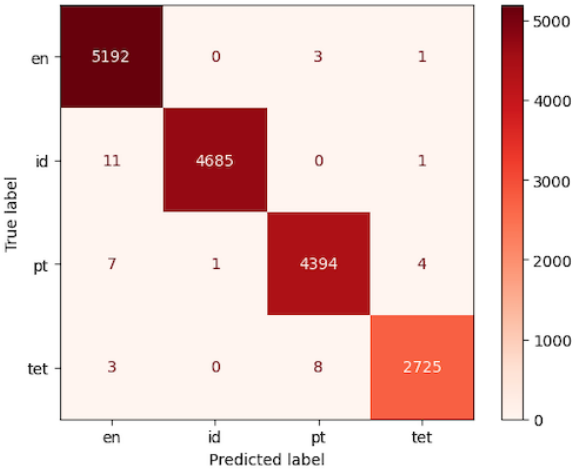

Figure 2: Confusion matrix of the evaluation on the test set: Tetun (tet), Portuguese (pt), English (en) and Indonesian (id).

6.1. Architecture

The Labadain Crawler's architecture is illustrated in Figure 3. It comprises several main stages: providing an initial text corpus as input, words seeding, crawling and indexing, text processing, and generating outputs (a final corpus and its summary). The modules used to process the initial corpus and performed seeding were adapted from techniques proposed by Baroni and Bernardini (2004) and Linder et al. (2020). Crawling was executed using Nutch, with the crawling outputs automatically transferred by Nutch and indexed in Solr. The modules for text processing and outputs generation represent our novel approaches integrated as part of the solutions proposed in this work. The following subsections detail the functional process of the Labadain Crawler.

6.1.1. An Initial Text Corpus

To enable the generation of seed words, an initial corpus containing text documents is provided. This corpus undergoes loading and preprocessing, which includes lower-casing and removal of

Model	Character n-gram						Word n-gram		
	1	2	3	4	5	6	1	2	3
SVM	0.9822	0.9954	0.9974	0.9975	0.9974	0.9968	0.9953	0.9689	0.8397
LR	0.9812	0.9953	0.9970	0.9975	0.9971	0.9960	0.9930	0.9522	0.8107
MNB	0.9452	0.9918	0.9967	0.9977	**0.9981**	0.9979	0.9973	0.9755	0.7806

Table 2: Accuracy of the models performance when evaluating using the development set.

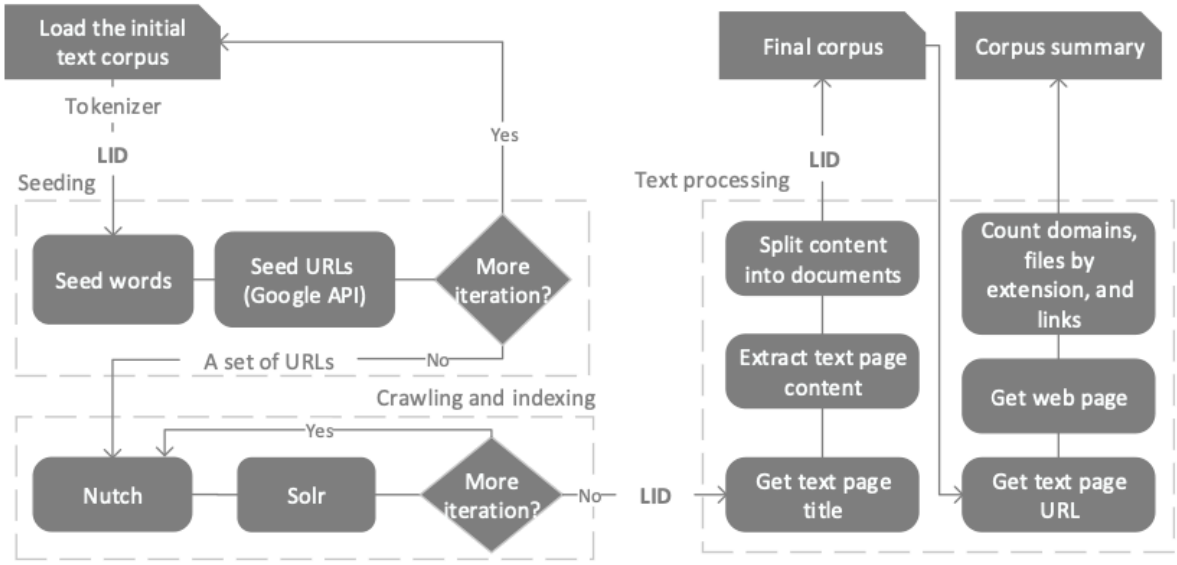

Figure 3: Architecture of the Labadain Crawler.

punctuation, special characters, and digits. Subsequently, the preprocessed corpus is tokenized into individual words, which are then passed through the LID model to validate whether their scores meet the predefined threshold on the target language. This process creates a vocabulary that is then used in the word seeding process.

6.1.2. Words Seeding

To generate a set of URLs, the frequency of the vocabulary is computed, and the query seeds are generated by randomly sampling three unique words separated by spaces, with probabilities following their frequency distribution. This query is then submitted to the Google search engine via the Google API[9] and the top ten URLs for the given query are retrieved and saved in the seed text file of Nutch. To ensure that the seed URLs are neither duplicated nor linked to PDF, Microsoft Office, images, audio, or video files, these types of URLs are filtered out during the URLs validation process. The generation of seed words and URLs is repeated based on the predefined configuration.

6.1.3. Crawling and Indexing

Using the provided set of URLs, Nutch initiates HTTP requests for the seed URLs, retrieves the associated web pages, and subsequently parses the fetched web pages to extract essential information, such as links, metadata, and content. Nutch's links extractor identifies and extracts hyperlinks from the parsed web pages, which are then used to discover new URLs for further crawling. Nutch maintains a queue that holds URLs to be crawled and performs repeated fetching and parsing steps for URLs in the frontier. Finally, a deduplication process is applied to remove duplicate URLs and contents. Upon completion of the crawling process, Nutch automatically transfers the crawled web pages to Solr[10], which are then indexed (we used Apache Solr version 8.9.0).

6.1.4. Text Processing

To retrieve the text content for the target language, the Solr API is utilized to access the web pages indexed in Solr. The following steps are executed: (1) Apply the LID model to each web page title. If the score obtained is greater than or equal to a predefined threshold, proceed to the subsequent processes. (2) Extract content of the given title, split it

[9] https://pypi.org/project/googlesearch-python/

[10] https://solr.apache.org

by a newline and store the resulting documents in a list. (3) Apply the LID model to each document in the list and filter out those do not meet the predefined threshold based on their scores. (4) Save the documents that meet the threshold along with their title and URL in the final corpus file. Each plain text extracted from the web page (referred to as text page) is separated by two subsequent newlines.

To generate the summary for the final corpus, the following steps are followed: (1) Load the contents from the final corpus file. (2) For each content, extract the respective URL and count the internet domain name and file type by extension from the URL. (3) Initiate the HTTP requests for the URL to retrieve the associated web page and utilize Beautiful Soup[11] to parse its content, enabling easy manipulation of the HTML elements. (4) Extract the inbound and outbound links associated with the page, count the number of links and store this information in the corpus summary file. The counting results of domain names and file types are saved in the corpus summary file at the conclusion of generating corpus summary.

6.2. An Experiment on Constructing a Tetun Text Corpus

The data collection pipeline was tested with Tetun under the following experimental settings: (1) The input text corpus was generated by randomly selecting 10% of the 3,500 Tetun news articles extracted from the Timor News database. (2) The Tetun Word Tokenizer was utilized to tokenize the input text into words (tokens) and the Tetun LID model with a score threshold of 0.95 was applied to filter the input tokens. (3) The seeding process was configured to repeat ten times and Nutch was set to adopt a depth-first crawling strategy with a depth limit of five (we used Apache Nutch version 1.19 released in September 2022). (4) The crawling process was set to repeat fifteen times with automated indexing in Solr.

After initiating the crawling process, all sequential steps were automatically executed until the construction of the text corpus and the generation of its summary were concluded. The experiment took place on a Linux Ubuntu virtual machine equipped with 16GB of RAM, 150GB of hard disk, and a single socket containing 4 cores. It concluded after approximately 46 hours of execution time.

6.3. Experimental Results Analysis

During the seeding stage, ten iterations were conducted to generate seed words and URLs, resulting in a total of 41 domains and 61 unique URLs. Subsequently, the crawling process started using these 61 URLs as the input seeds for Nutch. Upon completion of the crawling process, text processing was initiated, resulting in the generation of the final corpus and corpus summary. The top five internet domains from which Tetun text pages were retrieved are listed in Table 3 and the summary of the corpus content is presented in Table 4.

Domain	#text pages	Proportion
tatoli.tl	6,314	28.20%
timorpost.com	3,792	16.93%
naunil.com	3,343	14.93%
btl.tl	719	3.21%
laohamutuk.org	668	2.98%

Table 3: The top five domains from the dataset.

The corpus obtained from the crawled data was acquired from 149 different internet domains. Each text page represents the plain text content of a web page. The content of the crawled web page is parsed and structured within Nutch, generating a document appropriates for indexing in Solr. This document comprises essential fields such as title, URL, and content. Some of these text pages contain multiple articles, each associated with its corresponding title and URL. The corpus content includes text extracted from 589 PDF files and 13 PowerPoint files, with an average of 142.12 inbounds and 14.13 outbounds per URL.

Corpus content	Total
Text pages	22,392
Sentences	321,721
Tokens*	9,393,499

Table 4: Summary of the corpus content (excluding titles and URLs). *Tokens do not include punctuation and special characters.

Based on the internet domain names, we conducted a comprehensive analysis and then manually categorized the data sources into the following categories: (1) Online Newspapers: News portals affiliated with news agencies. (2) Governmental Institutions: Online portals maintained by government bodies. (3) Non-Governmental Institutions: Online platforms associated with national and international non-governmental organizations and agencies. (4) Educational Institutions: Online portals associated with universities and educational institutes. (5) Blogs and Forums: Online platforms originated from opinions and discussion forums. (6) Personal Pages: Online portals showcased individuals' personal information and interests. (7) Wikipedia: Documents sourced from Tetun Wikipedia. (8) Banks and Courts: Online

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

portals affiliated with banks and courts.
The corpus summary per data source category is presented in Table 5 and the distribution of text pages by domain name is summarized in Table 6.

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

Table 5: Total of the text pages per data sources.

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

Table 6: Total of the text pages per domain.

6.4. Evaluation of the Corpus Quality

To ensure that the text pages collected by the Labadain Crawler were Tetun documents and had clean titles and contents, we assessed these elements using data from the corpus summary and final corpus files. Initially, we utilized the corpus summary file, which provides a list of internet domain names alongside the total number of text pages for each domain, to analyze each domain name and determine the potential language of its contents. This process enabled us to ascertain whether the corpus contained texts in languages other than Tetun and gain a better understanding of its composition.
The analysis result showed that among the 149 domains in the corpus summary file, 22 domains were identified as potentially containing non-Tetun text pages. To confirm the language of these pages, we manually cross-referenced between the corpus content with the domain names. Our investigation found that out of the 22 domains, 20 did not contain Tetun text pages. These 20 domains were associated with 22 text pages, consisting of approximately 0.01% of the corpus content. As the LID model was applied in the process of generating the final corpus, those pages resulted in empty contents, meaning they did not contain body text, which were then excluded from the final corpus.
Furthermore, the text pages in the final corpus file were utilized to evaluate both the titles and contents of the text pages. This assessment employed the "quality at a glance" approach, as recommended by Kreutzer et al. (2022), which was also utilized by Artetxe et al. (2022) in their work. We randomly selected 50 text pages from the corpus and distributed them to six native Tetun speaker students, resulting in a total of 300 text pages being evaluated. The evaluators were given access to the sample corpus and Google Forms, facilitating them to assess various aspects of the corpus quality, including the quality of text page titles, noise level, recency and relevancy, and overall corpus quality. Details of the assessment criteria are outlined in Table 8.
The evaluation results reported that out of 300 text pages, five did not contain body text (Table 7). Upon conducting an in-depth analysis, we discovered that these five pages only consisted of text elements such as layout, menu names, and/or links, without any body text. This issue occurred due to these pages being linked to scanned PDF documents, web home pages, or pages lacking body texts in Tetun, despite having titles in Tetun. Overall, the qualitative evaluation highlighted that the generated corpus comprised high-quality textual data from diverse sources with clean titles and contents and written in Tetun. This underscores the effectiveness of the Labadain Crawler in collecting Tetun documents from the web.

7. Discussions

The Tetun Word Tokenizer, while primarily developed to address tokenization challenges in Tetun, showcases its adaptability by also being effective for other languages. For example, it has been successfully employed to tokenize English, Portuguese, and Indonesian texts into word tokens in the analysis of language similarities, as discussed in subsection 5.2. This versatility significantly enhances its utility and value, extending its applicability across a broader range of linguistic contexts beyond its initial target language.
Despite reports in existing Tetun literature highlighting the influence of Portuguese loanwords on Tetun, potentially reaching up to 40% (Hajek and van Klinken, 2019; van Klinken and Hajek, 2018; Greksáková, 2018), Figure 1 demonstrates that there is no significant indication of similarity between these two languages. This observation was further validated during the analysis and assessment of the corpus quality. Upon randomly inspecting text page titles in the corpus, we discovered that the LID model accurately classified some titles containing predominantly loanwords as Tetun. Examples include "CCLN Selebra Akordu Apoiu Grupu Veteranu" and "MTK Relata Progresu Investimentu Infrastrutura", where only "CCLN" and "MTK" are not Portuguese loanwords.

Quality metric	Description	#text pages	Proportion
Text page title quality	The text page title is in Tetun	300	100.00%
Text page content quality	The text page contains one or more articles	295	98.33%
Noise	The text page contains clean text	300	100.00%
Recency and Relevancy	Relevant content for present-day usage	278	92.67%
Overall Assessment	Diverse sources with high-quality content	295	98.33%

Table 7: The result of the corpus quality's assessment.

However, the LID model exhibited biases towards terms such as "Covid" and "Timor-Leste". These biases can be attributed to the higher frequency of these terms in the Tetun dataset used for training the LID model, while being less frequent or not appearing in the other three languages. Consequently, when these terms appear alongside other languages in short texts or as individual instances, the model tends to classify them as Tetun text. These biases were observed in the 20 domains associated with 22 text pages identified in the corpus analysis by internet domain names. These pages were collected due to their short titles, typically consisting of one to four words, which included either the word "Covid", "Timor-Leste", or both.

On the other hand, the Labadain Crawler has showcased its effectiveness in crawling the World Wide Web to collect Tetun document, even with limited computational resources, indicating its capability to operate on a personal computer with a minimum of 16GB of RAM and 100GB of hard disk space. Furthermore, the pipeline generates a summary of the crawled data, providing insights into the characteristics of the corpus, the quantity of documents per internet domain name, and the diversity of information collected.

In a multilingual environment like that of Timor-Leste, challenges arise from a mixture of texts obtained from the web, often within a single document. In certain instances, document titles are in one language while their contents are in another, leading to empty document body contents after applying the LID model. Therefore, auditing of document contents is essential to ensure the quality of the crawled data.

Furthermore, when comparing the Tetun text corpus to corpora constructing using different data collection techniques as outlined in section 2, it exhibits impressive statistics, comprising 321,721 sentences, with an average of 29.19 tokens per sentence, inclusive of words and digits. These metrics place it in a comparable size to the Swiss German corpus, which encompasses 562,524 sentences and was compiled from 3,473 domains. The comparison relies on the total number of domains utilized in constructing the Tetun text corpus, which amounts to merely 149 domains, significantly fewer than those employed in building the Swiss German corpus.

8. Conclusions and Future Work

This paper introduces Labadain Crawler, a data collection pipeline built on top of the Apache Nutch framework and incorporated with a tokenizer and a LID model. Since these language processing components were not previously available for Tetun, we initially developed them as the primary contributions of this work. Subsequently, we integrated these components into the Labadain Crawler pipeline and tested them with Tetun, resulting in the construction of the first-ever Tetun text corpus, encompassing 321,721 sentences extracted from 22,392 text pages (each text page representing plain text extracted from a web page). The dataset will be accessible to researchers upon request.

A manual assessment of the corpus content revealed its richness in diverse sources and high-quality content, rendering its suitability for various NLP and IR related tasks. The quantity and quality of the collected Tetun documents underscore the effectiveness of the Labadain Crawler pipeline in collecting textual data from the web. Moreover, the modular design of the pipeline enables easy customization and extension, streamlining its adaptation and application to other LRLs facing similar challenges in textual data construction.

Furthermore, we observed the accuracy of the Tetun LID model when integrated with the pipeline to discriminate Tetun text from a mixture of textual data acquired from the web. The model proved successful in distinguishing Tetun, particularly when dealing with loanwords. Nevertheless, biases in the LID model towards terms such as "Covid" and "Timor-Leste" were identified. These terms suggested to be more prevalent in the Tetun dataset used for training, leading to biases in the model, especially noticeable when processing short texts containing those terms.

In future work, we aim to improve the performance of the Tetun LID model by leveraging the constructed corpus to mitigate the model bias issues towards specific terms identified in this work. Additionally, we plan to expand the existing text corpus by crawling more data and make it available for the IR and NLP researchers.

9. Acknowledgement

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia under the PhD scholarship grant number SFRH/BD/151437/2021.

10. References

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 4344–4355. European Language Resources Association.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7226–7249. Association for Computational Linguistics.

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.

Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. A survey of corpora for germanic low-resource languages and dialects. In *Proceedings of the 24th Nordic Conference on Computational Linguistics, NoDaLiDa 2023, Tórshavn, Faroe Islands, May 22-24, 2023*, pages 392–414. University of Tartu Library.

Ralf D. Brown. 2013. Selecting and weighting n-grams to identify 1100 languages. In *Text, Speech, and Dialogue - 16th International Conference, TSD 2013, Pilsen, Czech Republic, September 1-5, 2013. Proceedings*, volume 8082 of *Lecture Notes in Computer Science*, pages 475–483. Springer.

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.

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. http://mj.gov.tl/jornal/lawsTL/RDTL-Law/RDTL-Gov-Decrees/Gov-Decree-2004-01.pdf, last accessed on February 21, 2024.

Olha Dovbnia, Witold Sosnowski, and Anna Wróblewska. 2022. Automatic language identification for celtic texts. In *Neural Information Processing - 29th International Conference, ICONIP 2022, Virtual Event, November 22-26, 2022, Proceedings, Part VI*, volume 1793 of *Communications in Computer and Information Science*, pages 264–275. Springer.

Alexandra Espichán-Linares and Arturo Oncevay-Marcos. 2017. Language identification with scarce data: A case study from peru. In *Information Management and Big Data - 4th Annual International Symposium, SIMBig 2017, Lima, Peru, September 4-6, 2017, Revised Selected Papers*, volume 795 of *Communications in Computer and Information Science*, pages 90–105. Springer.

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). https://www.laohamutuk.org/DVD/DGS/Cens15/2015-Census-Agriculture-and-Fisheries-report.pdf.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomás Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

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

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.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2545–2568. Association for Computational Linguistics.

Instituto Nacional de Estatística Timor-Leste INETL. 2022. Timor-leste population and housing census. https://inetl-ip.gov.tl/2023/10/04/2022-census-wall-chart/.

National Institute of Linguistics INL. 2004. The standard orthography of the tetum language: 115 years in the making. https://archive.org/details/the-standard-orthography-of-the-tetum-language/mode/2up, last accessed on February 21, 2024.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *J. Artif. Intell. Res.*, 65:675–782.

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.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *CoRR*, abs/1804.10959.

Siva Subrahamanyam Varma Kusampudi, Anudeep Chaluvadi, and Radhika Mamidi. 2021. Corpus creation and language identification in low-resource code-mixed telugu-english text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3 Sep., 2021*, pages 744–752. INCOMA Ltd.

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.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *CoRR*, abs/2006.07264.

Gordon Mohr, Michael Stack, Igor Ranitovic, Dan Avery, and Michele Kimpton. 2004. An introduction to heritrix - an open source archival quality web crawler. In *In IWAW'04, 4th International Web Archiving Workshop*. Springer Press.

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

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. http://hdl.handle.net/10400.22/4008.

Andreas von Holy, Alon Bresler, Osher Shuman, Catherine Chavula, and Hussein Suleman. 2017. Bantuweb: a digital library for resource scarce south african languages. In *Proceedings of the South African Institute of Computer Scientists and Information Technologists, SAICSIT 2017, Thaba Nchu, South Africa, September 26-28, 2017*, pages 36:1–36:10. ACM.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020a. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *LREC*, pages 4003–4012. European Language Resources Association.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020b. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4003–4012. European Language Resources Association.

Marcos Zampieri. 2013. Using bag-of-words to distinguish similar languages: How efficient are they? In *2013 IEEE 14th International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 37–41.

Appendix A. Assessment Criteria

The assessment criteria used for the qualitative evaluation is presented in Table 8.

Text page title quality	The text page title is in Tetun.
	A significant portion of the text page title is not in Tetun.
Text page content quality	The text page does not contain any article.
	The text page contains one or more articles.
Noise	The text page is clean.
	The text page contains extraneous elements such as HTML codes, non-Latin alphabets, and is not entirely clean.
Recency and Relevancy	The text page contains information from the past five years up to the present day.
	The text page contains older information, but it remains relevant for present-day usage.
	The text page contains outdated information that is not relevant for current use.
Overall quality assessment	Diverse sources with high-quality content.
	High-quality content but lacking diversity across sources.
	Diverse sources with medium-quality content.
	Medium-quality data but lacking diversity across sources.
	Diverse sources with low-quality content.
	Low-quality content with lacking diversity across sources.

Table 8: Assessment criteria used for the qualitative evaluation.